Response Processes in Noncognitive Measures: Validity Evidence from Explanatory Item Response Modeling

Michael Rodriguez[1], Okan Bulut[2], Kory Vue[1], Julio Cabrera[1]

[1] University of Minnesota
[2] University of Alberta

Minnesota Youth Development Research Group
www.mnydrg.com

April 2018

Paper presented at the annual meeting of the
National Council on Measurement in Education, New York, NY.

Citation:

Rodriguez, M.C., Bulut, O., Vue, K., & Cabrera, J. (2018, April). *Response processes in noncognitive measures: Validity evidence from explanatory item response modeling.* Paper presented at the annual meeting of the National Council on Measurement in Education, New York, NY.

Response Processes in Noncognitive Measures: Validity Evidence from Explanatory Item Response Modeling

Abstract

Consistent with improving the positive impact of assessment on teaching and learning, we explore score interpretation validation for a noncognitive measure of social competence, using a partial-credit explanatory item response model. Item and person characteristics interact in significant ways, influencing test-taker response processes and potentially influencing score interpretation.

Introduction

With increased attention to measurement of social and emotional learning and other noncognitive traits, psychometricians are finding their skills stretched. These measures are typically developed without psychometric expertise. As measurement specialists enter this arena, we bring the expectations from the *Standards for Educational and Psychological Testing* (hereinafter referred to as *Testing Standards*; AERA, APA, NCME, 2014), not commonly applied to noncognitive measures.

Among the most important of the *Testing Standards* are expectations for establishing and documenting validity. Empirical research of test-taker thinking processes provides support for score interpretation (Embretson, 2016), equally so for noncognitive measurement. Consistent with the NCME theme, *making assessment a stronger force for positive impact on teaching and learning*, we must secure validity (and fairness) evidence for all measures used to impact educational processes. Sources of validity evidence are described in the *Testing Standards*, but those sources of evidence are not intended to serve as a menu or recipe. The appropriate sources depend on the intended inferences and claims. In most cases, multiple sources of evidence are needed, which ideally indicate consistency and support for other sources.

One common critique of noncognitive measures is the potential for cultural influences in test-taker response processes (this critique is of course also commonly leveled against cognitive measures). Evaluations of measurement invariance and differential item functioning (DIF) analyses are rarely applied to noncognitive measures, but because noncognitive measures

estimate traits that may be more inherently culturally embedded (Rodriguez & Morrobel, 2004), the need is high.

We explore a measure of social and emotional learning through the lens of response processes in the manner presented by Embretson (2016), via an explanatory item response model. The impact of a cognitive feature of the item and a person characteristic on item functioning are explored. As a secondary issue, it in unclear as to what form of evidence such exploration provides. We briefly introduce the arena of social and emotional learning measurement, review relevant sources of validity evidence, present the interpretive argument for the target SEL measure, and describe the model and results.

Social & Emotional Learning

Social and emotional learning (SEL) goes by many names, including noncognitive measures, 21st century skills, soft skills, and many others. These names are not necessarily interchangeable, since the authors that align with one name often are interested in a particular set of constructs for a particular purpose (e.g., researchers examining 21st century skills are often interested in investigating career readiness and associated knowledge, skills, and abilities; see Greiff & Kyllonen, 2016, for example). But the broad range of constructs under each label overlap substantially. Nevertheless, the research foundations for the role of SEL in educational processes are well grounded (see Durlak et al., 2011, for a meta-analysis of school-based SEL interventions). However, the measurement challenges are significant (Kyllonen, 2012).

In the K-12 education arena, where SEL is gaining presence, there are many frameworks from which to draw domains and different definitions of similarly sounding domains (see the Collaborative for Academic, Social, and Emotional Learning [CASEL, 2017, 2018] for comprehensive access to SEL measurement resources). A broad group of leading scientists and scholars, in the Council of Distinguished Scientists, recently endorsed a consensus report of the National Commission on Social, Emotional, and Academic Development: *The evidence base for how we learn: Supporting students' social, emotional, and academic development* (Jones & Kahn, 2017). This consensus statement reflects on the evidence basis supporting the position that "social, emotional and cognitive competencies develop throughout our lives and are essential to success in our schools, workplaces, homes, and communities and allow individuals to contribute meaningfully to society" (Jones & Kahn, 2017, p. 7). In addition, "all students, regardless of their background, benefit from positive social and emotional development. At the same time,

building nurturing, and integrating social, emotional, and academic development in pre-K-12 can be a part of achieving a more equitable society" (Jones & Kahn, 2017, p. 12).

Minnesota recently adopted the CASEL framework for its SEL competencies and benchmarks, arguing that a well-rounded education includes explicitly teaching SEL (MN Department of Education, 2018a, 2018b). The School Safety Technical Assistance Center provides school districts with guidance and resources to support the integration of SEL into schoolwide teaching and learning practices. The CASEL framework includes five broad domains referred to as competencies, including self-awareness, self-management, responsible decision-making, relationship skills, and social awareness. It also explicitly acknowledges the roles of three spheres of contexts, including classrooms, schools, and homes and communities. Although the state provides a framework for schools regarding SEL implementation and assessment (MN Department of Education, 2018a), no specific assessments are required or preferred (and there are many entering the marketplace).

Positive Youth Development

One framework for approaching the work of SEL is based on positive youth development (PYD), the basis for the developmental asset framework (Benson, Scales, Hamilton, & Sesma, 2006; Scales & Leffert, 2004). PYD originates in a reaction to the overwhelming ubiquity of deficit-oriented (medical) models of youth development. Benson et al. (2006) identified five core characteristics common among numerous definitions concerning PYD:

> (A) developmental contexts (i.e., places, settings, ecologies, and relationships with the potential to generate supports, opportunities, and resources); (B) the nature of the child with accents on inherent capacity to grow and thrive (and actively engage with supportive contexts); (C) developmental strengths (attributes of the person, including skills, competencies, values, and dispositions important for successful engagement in the world); and two complimentary conceptualizations of developmental success; (D) the reduction of high-risk behavior; and (E) the promotion of thriving. (p. 896)

They include a wide range of contexts in their explanation of communities, including family, school, neighborhoods, programs, congregations, peers, and workplace. These interact with person characteristics (nature of the child and developmental strengths) and developmental success (less risk behaviors and more thriving).

Consistent with the developing models of PYD, the concept of developmental assets can be traced to 1990 (Benson, 1990). The National Research Council and Institute of Medicine (Eccles & Gootman, 2002) explicitly recognized the concept of assets, and argued that personal

and social assets facilitate well-being and successful transition through adolescence into adulthood, also noting the importance of context, including experiences, settings, and people. The developmental asset framework links "features of ecologies (external assets) with personal skills and capacities (internal assets)" (Benson et al., 2006, p. 906). Benson et al. hypothesized that developmental skills and supports impact all youth.

Validity Evidence

The *Testing Standards* (AERA, APA, & NCME, 2014) describe several possible sources of validity evidence. The relevant sources depend on the interpretive argument, the claims or intended score inferences and interpretations. The sources of evidence include content, response processes, internal structure, and associations with other variables. Evidence based on test content is often represented in the test specifications, which details the content, item formats, cognitive tasks, and other relevant item features in support of score interpretation. For example, "of particular concern is the extent to which construct underrepresentation or construct-irrelevance may give unfair advantage or disadvantage to one or more subgroups of test takers" (p. 15). In this way, content-related validity evidence supports score interpretation and meaning across relevant groups of test takers. "Evidence based on response processes generally comes from analyses of individual responses" (p. 15). However, the examples provided regarding forms of information consist of interview responses, performance process information (e.g., drafts of writing task responses), eye movements, and response times, as well as observer or performance judge behaviors. DIF is mentioned in the *Testing Standards* as a source of evidence based on the internal structure of the measure, in that differential functioning results when test takers with similar levels of the trait systematically respond differently to one or more items. Further, it was suggested that sometimes DIF is construct relevant, as in the case where a group of test takers with a common characteristic respond to a set of items with a common feature in a way that is consistent with that person characteristic, but in a way that differs from other test takers.

Another source of evidence described in the *Testing Standards* includes associations with other variables, not just other test scores, but such variables as group membership, which "become relevant when the theory underlying a proposed test use suggests that group differences should be present or absent if a proposed test score interpretation is to be supported" (p. 16). This

is particularly relevant to a validity argument that suggests such associations are construct relevant and contribute to the intended score interpretation.

In terms laid out by the *Testing Standards*, the role of group membership and item features comprises validity evidence based on test content, response processes, internal structure, and associations with other variables. As relevant item features constitute cognitive processes (e.g., items referring to one's self or others) relevant to person characteristics (e.g., cultural traditions or orientations), empirical evidence based on response processes can play an important role.

Response Processes

A new emphasis on assessment design and development has led to deeper investigations in the cognitive processes examinees use when responding to test items (Ercikan & Pellegrino, 2017). In doing so, we can answer questions about whether items tap intended knowledge, skills, and abilities. In addition, such investigations can examine item features that may be associated with item parameters (e.g., difficulty and discrimination), as well as whether such associations are invariant in diverse communities. Although response process data are gaining presence in the assessment arena, "the use of such data in test validation is rare" (Ercikan & Pellegrino, 2017, p. 2). According to Ercikan and Pellegrino (2017), "response processes refer to the thought processes, strategies, approaches, and behaviors of examinees when they read, interpret, and formulate solutions to assessment tasks" (p. 2). Response process data include test taker verbalizations from think-alouds and cognitive interviews, eye-movements recorded through eye-tracking, response process logs recorded during task interactions in computer delivered tests including the use of available test taking tools (e.g., onscreen calculator, dictionaries, or other resource tools), and response time information. These researchers and the many contributing authors to this edited volume consider achievement tests as their primary target, including response data that is obtained in addition to the item responses themselves.

Response process data have been used to support claims about test taker performance, particularly regarding engagement in complex tasks. But response process data can also be helpful to understand how test takers understand and interpret items, whether items assess intended constructs, and whether test takers from different backgrounds engage in similar response processes (Ercikan & Pellegrino, 2017; Kopriva & Wright, 2017). Without response process evidence, intended interpretations cannot be adequately supported. The strength of the

interpretive argument rests on the adequacy of the validity argument (consistent and convincing evidence), as articulated by Kane (2013). The *Testing Standards* position validation as preparing a sound scientific basis for score interpretation.

Embretson (2016) looks at response processes from a cognitive perspective that asks deeper questions of the responses to assessment items themselves, not requiring the collection of additional data or observations. In the IRT framework, an item response is the result of the item-person interaction. It is implicitly a function of the item and person characteristics. Embretson posed the question: "Can empirical research on the basis of examinees' responses to test items result in better measurement of the intended constructs?" (p. 6). This requires a deeper definition of the constructs and measure specifications, in terms of the relevant item features represented that allow for generalization to the construct domain. She defined construct representation as the representation of the processes, strategies, and knowledge test takers draw on when responding to the items in a measure. Moreover, she gives a central role to empirical research on the cognitive processes in which test takers engage as they respond to items. These cognitive processes are elicited by item features designed to be consistent with or require use of the target cognitive task.

Kane and Mislevy (2017) extend the types of data that encompass response process including patterns of responses across tasks. They recognize that cognitive theories can be utilized to examine process data in ways that connect it to score interpretation. They include the opportunity to examine trait interpretations. "A trait is a disposition to behave or perform in some way in some kinds of situations across some range of circumstances" (p. 11). Although they acknowledge that traits are important components of personality theory, they focus their attention to cognitive traits (e.g., reading ability or quantitative reasoning). They contrast trait interpretations with process-model interpretations, which involve specific cognitive processes with small grain sized analyses, and where score meaning is determined by the model. Process-model interpretations fit response data to specific models (e.g., a model for solving two-digit subtraction problems). "Trait interpretations tend to be relatively broad, focusing on performance domains associated with the trait, the grain size tends to be large, focusing on general competencies, and the meaning of the scores is, to a large extent, determined by the performance domain of interest, with cognitive models playing a supporting role rather than a defining role" (p. 12).

In describing approaches to validating trait interpretations, Kane and Mislevy (2017) reasoned that much of the evidence needed can be obtained during test development, but additional evidence can be gathered by "fitting more detailed process-motivated psychometric models" (p. 19), which may closely resemble internal structure sources of evidence. In particular, a trait interpretation can be challenged if different processes are employed by different test taker groups, such that scores may not support a common interpretation across groups (e.g., evaluated through DIF analyses or multigroup CFA, more structural approaches). But when the evidence is grounded in the cognitive processes employed by test takers, these become sources of evidence based on response processes reflected in the item response data. These authors remind us that Cronbach (1980) argued that the validation task is not to simply support a score interpretation, "but to find out what might be wrong with it" (p. 103).

Similarly, Embretson (2016) explored the use of explanatory item response models to contribute to the validity argument. She reminds us that whereas educational achievement tests tend to focus on evidence based on test content (supporting inferences about student knowledge, skills, and abilities vis-à-vis content standards), tests of aptitudes and other abilities rely more on evidence based on internal structure or associations with other variables. Similarly, she argues that test consequences as a basis for supporting a validity argument often include adverse impact and DIF. Moreover, she argued that evidence based on response processes is often not part of the validity argument, and rarely employed in item or test development. Embretson provided an integrated system of validity where content affects response processes and subsequently response processes affect internal structure and associations to other variables.

Through a series of empirical studies, largely based on explanatory item response models, Embretson (2016) investigated the cognitive features of items and how they play a role in item functioning, as it may relate to score interpretation. She suggested that the types of cognitive structures/features designed in items impacted what was measured and associations to other variables. She further pointed out that this has important test design implications, regarding the intended or required balance of such item features. Construct relevance can be controlled through item selection emphasizing certain item features. The target construct measured will depend on the balance of cognitive processes, characteristics, or features across the measure as a whole. What we need, in the context of SEL measurement, is a cognitive model of survey item responses. In the context of achievement tests, Embretson demonstrated how item difficulty

could be manipulated by modifying items regarding cognitive complexity based on a cognitive processing model. In addition, depending on the nature of the cognitive processes in items that present challenges to certain students, the most effective interventions may depend on which items (with which cognitive processes) posed challenges for those students.

Item responses can be linked to item features associated with specific cognitive processes (Embretson, 2016). Explanatory item response models can provide empirical evidence of the role of cognitive features. "Explanatory IRT models can be applied to traditional item response data to understand response processes" (pp. 20-21). The effects of cognitive features of items on item functioning can be empirically estimated. These effects are inherently important aspects of the internal structure of the measure. Moreover, various combinations of these item features may differentially affect associations with other variables and result in DIF, and ultimately "could impact the *consequential* aspect of validity" (p. 21), especially when considering score interpretation and use.

Trait Interpretations

Contemporary achievement tests are employing "next generation" standards that explicitly tap complex cognitive skills and abilities. For example, the next generation science standards employ big ideas and tasks grounded in the practices of scientists. Such assessments are relying more heavily on response process data as part of their validity arguments (see for example Nichols & Huff, 2017; Chapelle, Enright, & Jamieson, 2010). Measures of social and emotional learning are inherently complex, as they typically are grounded in complex contexts and describe complex social and emotional characteristics, including internal (personal) processes that interact with external (social) processes. Such processes also are rooted in cultural practices and traditions.

A particularly significant challenge in the measurement of SEL is the potential dependency of the trait definition on cultural characteristics. The expected performance over the domain of possible performances described earlier draw our attention to the target domain (Kane, 2013). The target domain can be

> defined in terms of performances that are thought to require the competencies associated with the trait. Although performance is expected to vary from task to task, traits are taken to be invariant over some sets of tasks, contexts, and occasions. (p. 17, Kane & Mislevy, 2017)

Messick (1970) defined a trait as:

> a relatively enduring characteristic of a person—an attribute, process, or disposition—which is consistently manifested to an appropriate degree when relevant, despite considerable variation in the range of settings and circumstances. (as cited by Kane & Mislevy, 2017, p. 17)

In this expanded context of response processes, DIF has been offered as a tool. Although DIF has traditionally been identified as a method to evaluate test fairness (lack of measurement invariance), Zumbo (2007a) identified one purpose of DIF as a method to try to understand item response processes. As such, it is a way to investigate the cognitive or psychosocial processes employed in responding to assessment items as a function of group membership. This directly considers the presence of limits to measurement inferences, where DIF is intimately tied to test validation, as DIF results might establish inferential limits or bounds to test score interpretations across diverse groups (Zumbo, 2007b; Zumbo & Rupp, 2004). Others have extended this work, including through the use of explanatory item response models, to investigate sources of DIF or to explain DIF results (Albano & Rodriguez, 2013; Li, Cohen, Ibarra, 2004).

In the context of social and emotional learning (SEL), traits are inherently multifaceted. However, the practitioner requires a simple index summarizing the layers implied by theory. SEL trait interpretations involve competencies which imply performance across contexts. For example, measures of empowerment might include items related to abilities to take on useful roles and responsibilities or to have a sense of safety to fully engage in such roles. In addition, these abilities may occur at home, at school, or in the community (see, for example, the Developmental Asset Profile, Search Institute, 2013).

An Interpretive Argument regarding Social Competence

One SEL skill is social competence. It requires a combination of self-awareness and social-awareness. It is an important vehicle for school success, and likely for college and career success (Benson, 2002; Scales & Leffert, 2004). In this study, we explore a measure of social competence administered though a statewide youth survey to evaluate the role of an item characteristic that potentially interacts with cultural tradition to influence item responses. This measure, *Social Competence* (italicized when referring to the specific measure rather than the construct) comes from the Developmental Asset Profile (Search Institute, 2013) and is characterized as an external asset comprising resistance skills, peaceful conflict resolution, and personal power.

A major inference regarding this measure (and SEL measures in general) is the invariance of its meaning across cultural groups. If we believe the assumptions described earlier, that greater attention to SEL can promote greater equity in schools and that these measures are relevant too all youth, evidence should be gathered regarding the response processes of diverse youth. One possible characteristic, particularly relevant to the concept of social competence is the individualistic versus collectivistic orientation of the cultural community. In some cultural traditions, there is a strong collectivistic orientation (in our case, within American Indian and Latino communities), whereas in other cultural traditions, there is a strong individualistic orientation (European-American communities; Brendtro, et al., 2002; Freeberg & Stein, 1996; Rhee, Uleman, & Lee, 1996).

Item content similarly plays an important role in measures such as *Social Competence*, since context is a theoretically core component of SEL in general, in the context of PYD and ecological models of development. The *Social Competence* measure includes four items referring to *self* and four items referring to *others* (see Table 1). This recognizes the roles of self-awareness and social-awareness. This also includes the notions of resistance skills and personal power (manifestations of self-awareness) and peaceful conflict resolution (a manifestation of social-awareness). These components could interact differentially in responses from individuals with more individualistic orientations (with a focus on self-awareness) versus individuals with more collectivistic orientations (with a focus on social-awareness). The hypothesis is that response processes to items with the self referent versus those with the others referent will function differentially for individuals with individualistic versus collectivistic orientations, and those orientations are found to be reflected in cultural communities. This is operationalized here as a function of race/ethnicity, where American Indian and Latino communities are known to embrace collectivism and European-American (White) communities are known to embrace individualism (relatively speaking).

Table 1
The Referent Basis for Each Social Competence Item

Item content	Referent
Resist dangerous/unhealthy things	Self
Build friendships	Others
Express feelings in proper ways	Self
Plan ahead and make good choices	Self
Avoid bad influences	Self
Resolve conflicts without violence	Others
Accept people who are different	Others
Sensitive to others' needs/feelings	Others

Scores on S*ocial Competence* should imply these characteristics of students, equally so across cultural communities. Organizations striving to improve educational equity in MN are using measures of SEL to support their efforts, particularly in communities facing persistent challenges, such as Generation Next (http://www.gennextmsp.org), Great Expectations (https://www.iocp.org), Ignite Afterschool (https://igniteafterschool.org), and Partners for Student Success (http://www.partnerforstudentsuccess.org). They look to *Social Competence* levels for groups of students (not individuals) to understand the extent to which students are equipped (ready) for learning (in addition to other measures such as *Positive Identity* and *Commitment to Learning*). When students perceive high levels of self-awareness and social-awareness to describe themselves (social competence), they are able to employ resistance skills, rely on their own personal power or agency, and peacefully resolve conflicts – all important skills for school success as well as college and career success. In the context of striving for educational equity, such skills become equally important across all cultural communities. Organizations can use score information to support decisions regarding the adoption of skill-building interventions and supports.

We return to this "equipped" level on *Social Competence* to estimate the potential impact of variation in scale performance (item performance) for American Indian students. How this level is estimated is described more fully in at the end of the results section when the impact on student scores is explored and described.

Our primary question could be framed in terms of measurement invariance, by examining the influence of person characteristics on the estimated item parameters, or in terms of content or

internal structure by examining the content of the items and the use of collectivistic or individualistic references. This investigation empirically explores possible differential response processes based on item and person characteristics, informing score interpretation and use (for better or worse).

Research Question

Does the item referent (self versus others) influence item parameter estimates for American Indian and Latino students differentially than for White students in the *Social Competence* measure?

Methods

Minnesota Student Survey

In this study, we used the 2016 Minnesota Student Survey (MSS; MN Department of Education, 2017) data set to examine the impact of an item characteristic (referent) and person characteristic (race and ethnicity) on item responses. The MSS is an anonymous survey administered every three years, including students in grades 5, 8, 9, and 11. The survey is designed by an interagency team, including the Departments of Education, Health, Human Services, and Public Safety.

Approximately 85% of MN public school districts voluntarily participated in the 2016 MSS. The sample closely matched the state population in terms of race and ethnicity (67% White only, 9% Latino, 5% American Indian, 5% Black non-Somali, 2% Somali, 4% Asian non-Hmong, 3% Hmong), as well as participation in special education (10%) and free and reduced-price lunch (28%).

In the final sample for this analysis, we identified White, American Indian, and Latino students who responded to all eight items in the *Social Competence* measure. This resulted in 30,962 White students (who are not also another race/ethnicity), 7,974 American Indian students, and 14,153 Latino students. In this analysis sample, 50.1% were female, with 23% in grade 5, 28% in grade 8, 28% in grade 9, and 21% in grade 11.

The MSS includes a number of measures that we consider to assess social and emotional learning characteristics. These were developed directly by Search Institute (2013), including three measures from their Developmental Asset Profile (DAP), and three additional measures associated with DAP domains, but based on similarly-worded items available on the MSS. The three DAP measures include (a) *Social Competence*, (b) *Positive Identity*, and (c) *Empowerment*. The three DAP-like measures include (d) *Commitment to Learning*, (e) *Family-Community Support*, and (f) *Teacher-*

School Support. These measures have been characterized as developmental skills (a, b, and d) and developmental supports (c, e, and f). Our focus here is on the developmental skill measure *Social Competence*.

Social Competence is measured with eight items (see Table 2), each including a 4-point rating scale (rarely, sometimes, often, almost always). The questions cover a range of skills and contexts related to social competence, defined as "the skills to interact effectively with others, to make difficult decisions, and to cope with new situations" (Search Institute, 2013). All 20 questions included in the three measures of developmental skills were submitted to a three-factor confirmatory factor analysis (CFA) with Mplus (v. 7; Muthen & Muthen, 2012), employing the WLSMV estimator, a robust weighted least squares probit-regression based method, which accounts for the categorical nature of the 4-point ordinal response scales.

The CFA resulted in fit indices with root mean-squared error of approximation (RMSEA) of .08, comparative fit index (CFI) of .92, and Tucker-Lewis Index (TLI) of .91, indicating adequate fit (Brown, 2015). The factor loadings for the *Social Competence* measure were uniformly strong and ranged from .60 to .80. The disattenuated correlations with the other two developmental skills and *Social Competence* were .68 (Commitment to Learning) and .85 (Positive Identity), whereas Commitment to Learning and Positive Identity were correlated .57. To support secondary analysis and research with these measures, they were then scaled using the Rasch model with Winsteps (v. 3.92; Linacre, 2016). In addition to scaling, Winsteps was used to conduct DIF analysis and no C-level DIF (Rasch DIF contrast > 0.64) was found as a function of sex, grade, or race and ethnicity.

Using the Rasch scaling values (scaled so that the average item location is 0), the scale was recentered so the midpoint of the rating scale was located at 10.0. Since the midpoint of the rating scale was associated with a Rasch measure of -0.04, the Rasch measures were transformed (Rasch Measure – -0.04 + 10), without changing the standard deviation. This resulted in a score scale ranging from 5.1 to 15.2 with the mean of 11.39 and a standard deviation of 1.6. The distribution is slightly positively skewed (0.4) and slightly leptokurtic (0.7), as the majority of students have strong perceptions of social competence, above the midpoint of the response scale. The response frequency of each item is reported in Table 2.

Table 2
Percent of Students Responding to the Frequency of Each Social Competence Item

Item content	Rarely	Sometimes	Often	Almost always
Resist dangerous/unhealthy things	7.1%	17.3%	25.2%	50.3%
Build friendships	5.7%	20.9%	38.2%	35.1%
Express feelings in proper ways	10.7%	29.8%	39.2%	20.3%
Plan ahead and make good choices	5.2%	25.9%	41.0%	27.9%
Avoid bad influences	6.4%	20.5%	29.6%	43.5%
Resolve conflicts without violence	5.5%	22.3%	38.9%	33.4%
Accept people who are different	1.9%	8.1%	33.4%	56.6%
Sensitive to others' needs/feelings	6.2%	19.3%	36.8%	37.6%

Variables

To support the modeling of the research question, two variables were derived from the survey items. The race/ethnicity characteristic was dummy coded with the focal groups including Latino (*n*=14,153) and American Indian (*n*=7,974) students, and White students (*n*=30,962) as reference group. The items were also coded regarding their referent basis, that is, whether the item referenced the students themselves or referenced other. With these two variables, race/ethnicity and referent, we can characterize the interaction of a person and item characteristic that theoretically play an important role in score interpretation of the *Social Competence* measure. As described earlier, we wondered about the role of cultural traditions in terms of being individualistic (represented by White students) versus collectivistic (represented by Latino and American Indian students) and how that may be reflected in responses to items that either focus on self or others.

Analytical Model

The partial credit model (PCM) represents the possible item scores from 0 to *J*, with adjacent item-response categories indexed by *j*, based on the log-odds of selecting response category *j* over *j*-1 on item *i* (*i* = 1, 2, 3, …, *I*) for person *n*. This is typically represented as

$$\log\left(\frac{P_{nij}}{P_{ni(j-1)}}\right) = \theta_n - \left(\delta_i + \tau_{ij}\right),$$

where θ_n represents the latent trait of person *n*. Although δ_i is often considered the overall item difficulty, this parameter represents the location of the threshold between the first (j = 0) and second (j = 1) response categories for item i. τ_{ij} represents the distance between the subsequent

thresholds. For example, if item *i* has three response categories, τ_{ij} would represent the distance between the 3×2 threshold and the 2×1 threshold. The threshold location can be interpreted as a "difficulty" parameter, or better, a trait level parameter, interpreted as the trait level required to have equal probability of choosing one of two adjacent category options for item *i*.

The cross-classified explanatory partial-credit model is a special characterization of the PCM, the log-odds of selecting response *j* over the adjacent category *j* −1 on item *i* for person *n* can be written as:

$$\log\left(\frac{P_{nij}}{P_{ni(j-1)}}\right) = \mathbf{Z_{nij}\theta_n} - \mathbf{X'_{nij}\delta_i} + \boldsymbol{\varepsilon_i}$$

where $\mathbf{Z}_{nij}$ is a matrix of fixed- and random-effects related to the latent trait θ_n distributed as $N(\mu_n, \sigma_n^2)$. $\mathbf{X}_{nij}$ is a matrix of fixed- and random-effects associated with individual items with the vector $\boldsymbol{\delta_i}$ of item locations, and $\boldsymbol{\varepsilon_i}$ represents an *I* x (*J*-1) matrix of the item threshold parameters estimated as random effects (e.g., Van Den Noortgate, De Boeck, & Meulders, 2003; Wang, Wilson, & Shih, 2006; Wang & Wu, 2011). It is important to note that in this model, both persons and items are considered random, as their random effects (variances) are estimated and reported. Also note that the parameterization is in item easiness form, such that the item locations and distances are in terms of -δ. However, in the summary tables below, the item parameter estimates have been converted to the traditional item location (difficulty) metric for ease of interpretation. Higher values indicate higher trait levels required to respond to higher categories. The models are estimated with *lme4* (Bates et al., 2015) in **R** (R Core Team, 2017).

In practical terms, each item, being a rating-scale response, has *J*+1 response categories and thus *J* thresholds (the point at which adjacent categories have equal probability of selection given theta). The remaining parameters estimate the distances from the first threshold for each item *i*.

In this study, we use the MSS data to examine the impact of item referent and person race/ethnicity variables on item responses in the *Social Competence* measure. The data set consists of 8 items and 53,089 students (with scores on the measure). Three models were evaluated here, including:

1. Model 1: Partial credit model with no predictors.
2. Model 2: Partial credit model + predictors (person race/ethnicity and item referent).
3. Model 3: Partial credit model + predictors + their interaction (Race/ethnicity x referent)

The three models allow us to estimate the fit of the model to the item response data without conditioning on person or item characteristics (Model 1), and compare that fit to the models where we

account for person and item characteristics (Model 2) and finally their interaction (Model 3). Model 3 is the model of highest interest, since the theoretical assumption in score interpretation for the *Social Competence* measures is that item responses, conditioned on trait level, are not influenced by person race/ethnicity (as an indicator of collectivistic/individualistic orientation) nor item referent (self versus other), nor their interaction (such that the effect of item referent does not depend on cultural orientation). Measures of fit are evaluated for all three models, as well as the model parameters (item thresholds and coefficients associated with the person and item characteristics and their interaction).

Results

Three models were examined to evaluate the functioning of ordinal responses from rating scale items in the measure *Social Competence*, given person characteristics (race/ethnicity as an indicator of individualistic versus collectivistic cultural orientations) and item characteristics (content reference to self versus others). The three models included the cross-classified unconditional PCM, the main effects PCM, and the full model including the interaction of person and item characteristics (does the effect of item referent depend on race/ethnicity). Here we briefly discuss results for each model. A more complete reporting of the *lme4* resulting output is provided in the Appendices A-C, one for each model.

As a check on the estimation of item thresholds in the cross-classified explanatory PCM, we compared the *lme4* estimates with those from the Winsteps PCM (Linacre, 2016). The distances from the first to second threshold and from the first to third threshold were correlated across the two estimation methods at .91 and .98 respectively. This indicates that *lme4* and Winsteps estimated the same relative distances between the three thresholds. We note that the explanatory item response model typically identifies the scale based on the average person location (person intercept random effects have $M = 0$), whereas Winsteps centers the scale at the average item location ($M = 0$).

Comparing AIC and BIC results, the full model substantially fits better than the unconditional model with no person/item characteristics. Model fit results are reported in Table 3. We focus the remaining presentation on the Model 3 results.

Table 3
Model Fit Indices for the Three Cross-Classified Explanatory Partial-Credit Model

Model	AIC	BIC	logLikelihood	Deviance	Residual *df*
1. Unconditional	816859	816951	-408421	816843	661798
2. Main Effects	815531	815667	-407754	815507	661794
3. Full with interaction	814695	814901	-407329	814659	661788

The person race/ethnicity results are very similar for the American Indian (0.55) and Latino (0.45) students (Model 3 in Table 4). These are main effects on person locations (person trait levels or thetas). This result, conditioned on item characteristic (referent), is consistent with the overall standardized mean difference observed in the scale scores for *Social Competence* (Figure 1); American Indian and Latino students report lower levels of *Social Competence* than do White students.

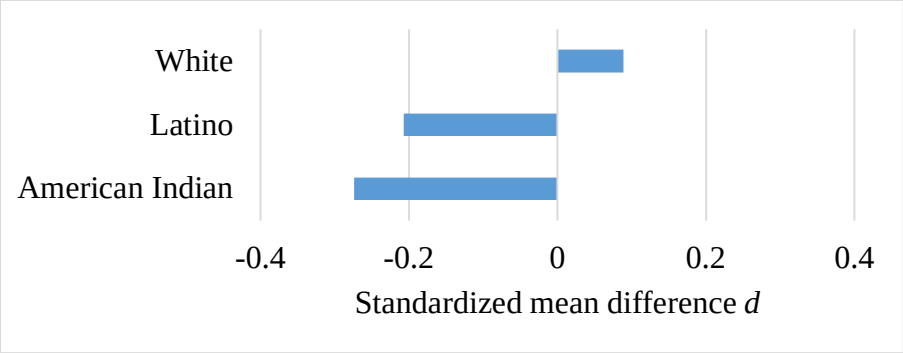

Figure 1. Standardized mean differences (from state average) in scores on *Social Competence* by race and ethnicity.

Of primary interest, each of the interaction effects (item × person characteristics; Model 3) are significant ($p < .005$, most at $p < .001$). Item referent (self versus other) and person race/ethnicity have significant effects on item thresholds. The interaction terms indicate that the effect of item referent depends on person race/ethnicity. We examined each partial effect relative to the item and person characteristic, acknowledging the primary role of the interaction.

Since most of the interaction terms are significant, the main effects on item parameters (for the item referent others) cannot be interpreted independently. In addition to the person effects being similar for both American Indian and Latino students, the interactions between item referent and person race/ethnicity were remarkably similar for both group

Table 4
Summary of the Explanatory Partial Credit Models

	Model 2		Model 3	
	Coefficient	SE	Coefficient	SE
Person Predictors				
American Indian	-0.46*	0.02	-0.54*	0.02
Latino	-0.37*	0.01	-0.45*	0.01
Item Predictors (Others referent = 1)				
Threshold 1 x Others	-1.50*	0.16	-2.73*	0.85
Threshold 2 x Others	-0.78*	0.08	-1.54**	0.39
Threshold 3 x Others	0.02	0.13	0.08	0.30
Item x Person Predictors				
Threshold 1 x Others x American Indian			0.19**	0.03
Threshold 2 x Others x American Indian			0.10**	0.02
Threshold 3 x Others x American Indian			-0.39**	0.02
Threshold 1 x Others x Latino			0.19**	0.03
Threshold 2 x Others x Latino			0.05*	0.02
Threshold 3 x Others x Latino			-0.33**	0.02

* $p < .005$, ** $p < .001$

For those items with the referent others, the distance between the first threshold and the third threshold is smaller for American Indian and Latino students. For students in these groups, with more collectivistic orientations, it takes slightly more social competence (0.19) to report that the characteristic in the time sometimes relative to rarely describes them compared to White students. For both American Indian and Latino students, 0.19 is added to the first threshold for items with an others referent. In addition, it takes less social competence (-0.39 to -0.33) to report that the characteristic almost always relative to often describes them compared to White students. For American Indian students, -0.39 is added and for Latino students -0.33 is added to the third threshold for items with an others referent. This results in a reduction of greater than 0.50 logits for the distance between the first and third thresholds: -0.58 (-0.39 – 0.19) smaller distance for American Indian students and -0.52 (-0.33 – 0.19) smaller distance for Latino students.

Since effects are additive in the PCM explanatory item response model, we estimated item thresholds, given the item referent and person race/ethnicity. Figure 2 contains estimates of these item thresholds (as reported in Table 5). For each item with an others referent (2, 6, 7, 8), there are three sets of estimates, including students who are White, American Indian, and Latino. For these items, the range of thresholds is smaller for American Indian and Latino students.

Table 5
Estimated Thresholds for the Unconditional Model and Final Model by Group

Items	Model 1 Thresholds			Model 3 Thresholds								
				White			American Indian			Latino		
	1	2	3	1	2	3	1	2	3	1	2	3
1	-1.8	-1.1	-0.7	-2.0	-1.1	0.2	-2.0	-1.1	0.2	-2.0	-1.1	0.2
2	-2.2	-0.8	0.5	-2.5	-1.2	0.4	-2.3	-1.1	0.0	-2.3	-1.2	0.0
3	-1.8	-0.7	1.0	-2.0	-0.7	2.3	-2.0	-0.7	2.3	-2.0	-0.7	2.3
4	-2.5	-0.2	1.3	-2.7	-1.0	2.4	-2.7	-1.0	2.4	-2.7	-1.0	2.4
5	-2.1	-0.7	0.0	-2.3	-1.0	1.0	-2.3	-1.0	1.0	-2.3	-1.0	1.0
6	-2.4	-0.5	0.9	-2.7	-1.2	0.8	-2.5	-1.1	0.4	-2.5	-1.1	0.4
7	-2.6	-1.4	0.1	-3.0	-2.2	-1.0	-2.8	-2.1	-1.4	-2.8	-2.2	-1.4
8	-2.0	-1.0	0.3	-2.3	-1.3	0.1	-2.2	-1.2	-0.3	-2.2	-1.2	-0.3

Note. Model 1 is unconditional; Model 3 adds person and item characteristic and their interaction. Model 1 also resulted in a main effect item intercept of -1.98, added here for comparability of scales.

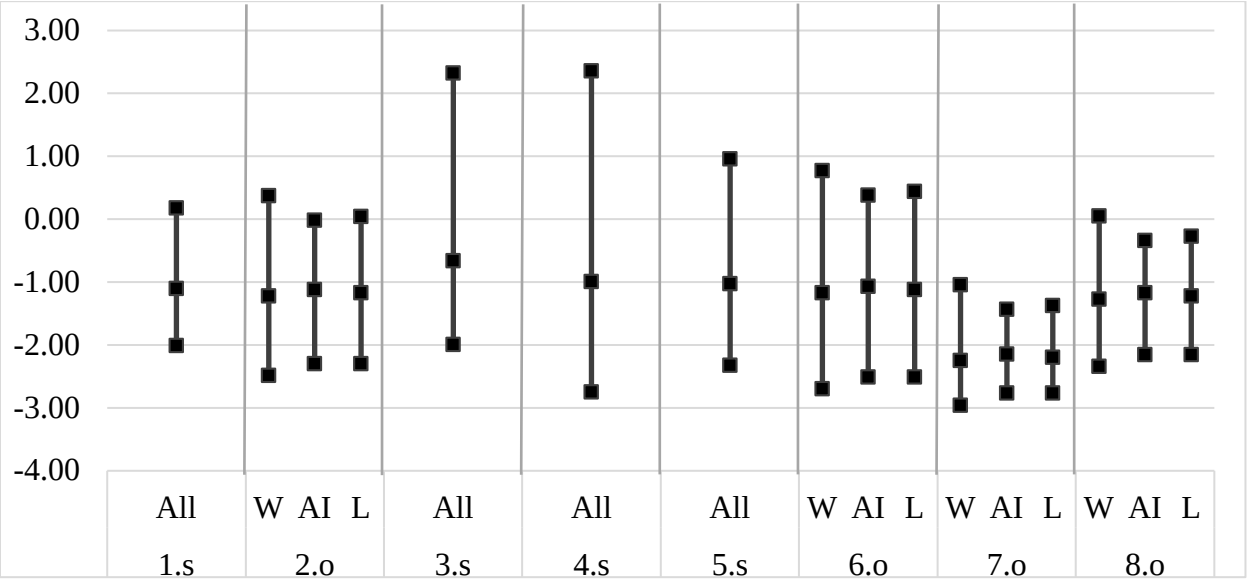

Figure 2. Location of item thresholds for students who are White (W), American Indian (AI), and Latino (L), for items referents "self" (denoted by .s) and "others" (denoted by .o).

Estimated Impact on American Indian Students

In an attempt to estimate the impact of this effect on item thresholds for students, we estimated the Rasch scores for American Indian students using the thresholds for White students and then again using the adjusted thresholds for American Indian students, based on the effects estimated above. We found that the estimates of *Social Competence* for American Indian students varied only slightly based on these variations in thresholds. First, the correlations with the original *Social Competence* scores as estimated by Winsteps and those based on the thresholds estimated here with lmer4 were essentially perfect (.999). The correlation between person scores using the two sets of thresholds was essentially perfect (same rank-ordering of persons). This can be seen in the scatterplot of the two sets of scores (Figure 3).

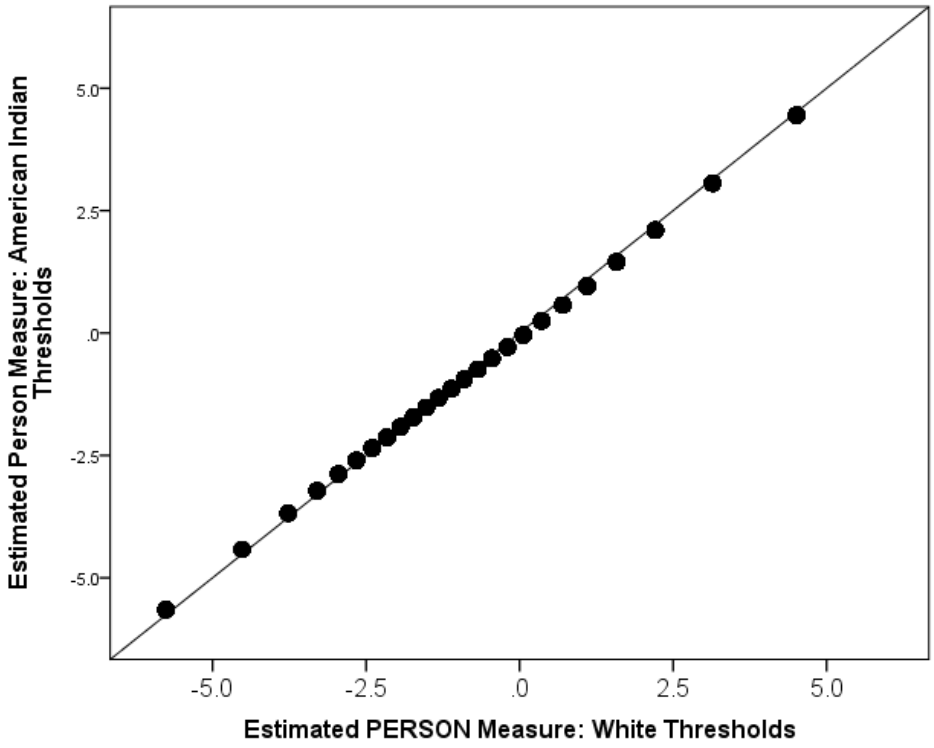

Figure 3. Scatterplot of *Social Competence* scores for American Indian students using the item thresholds estimated for White students compared to item thresholds estimated for American Indian students.

From Figure 3, we see the same rank ordering of scores across the scale (there are 24 observable score points in the 8 items with 4-point rating scale). Notice that American Indian students with higher scores (above 0) tend to be underestimated with White student thresholds

and those with lower scores (below 0) tend to be overestimated with White student thresholds. However, their rank order is the same.

As used in the community organizations monitoring student performance on this and other SEL measures, organizations use a criterion level associated with each measure as a goal for students in their communities. This is referred to as the "equipped" level, such that students who achieve that level "are equipped for learning". This equipped level is the point at which students report that the skills, behaviors, and beliefs represented in the measure are more like them than not. On the 4-point rating scale, this constitutes an average rating of 3 points. This average rating is then translated to the IRT metric through the test characteristic curve. On the measure of *Social Competence* using the thresholds for White students is -0.20 (an average rating of 3 out of 4). Using the scores based on the thresholds for White students, 49.6% of American Indian students would be identified as equipped for learning on *Social Competence*. Using the scores based on the thresholds for American Indian students, 41.9% of American Indian students would be identified as equipped for learning. Nearly 7.7% of American Indian students would be misclassified (610 of the 7974 students). We see this in the mean differences of scores using the two sets of thresholds, where when using the thresholds for White students, the average score for American Indian students is slightly lower than when using the adjusted thresholds for American Indian students (Table 6).

Table 6
Summary Statistics for American Indian Students Based on IRT Scores Estimated with Item Thresholds Estimated for White Students Compared to those for American Indian Students

Source of thresholds	Minimum	Maximum	*M*	*SD*
White students	-5.76	4.51	-0.09	1.69
American Indian students	-5.65	4.45	-0.15	1.66

Although these mean scores are very close, they are sufficiently different to result in different distributions of scores for American Indian students, resulting in different percentages of students identified as having the level of *Social Competence* to be equipped for learning by the groups using these measures to monitor progress toward closing achievement gaps.

It is also interesting to note that the differences in estimates based on the effects of ethnicity and item referent is not constant across the score scale. Overall we see a slight decrease

in mean scores when using the adjusted item thresholds, but in Figure 3, the effect is also dependent on the overall location of scores.

Discussion

To summarize the results, the primary results include:

1. The cross-classified (items and persons considered random) partial-credit explanatory item response model estimates closely match the PCM estimates from Winsteps.
2. Overall, American Indian and Latino students report a lower sense of social competence in the fully conditioned model, consistent with scale score differences.
3. For items with a referent to others (perhaps consistent with more collectivistic orientations), the first and second thresholds are higher for American Indian and Latino students, whereas the third threshold is lower for these students.
4. For items with a referent to others, the distance between the first and third thresholds is smaller for American Indian students, indicating a reduction in the additional amount of the trait (social competence) required to be likely to select a higher response options (suggesting that the item more often describes them).
5. As an example of impact, fewer American Indian students are identified as equipped for learning on the *Social Competence* measure when scores are estimated with the thresholds from White students than when scored with the thresholds corrected for American Indian students and the item referent of others.

Embretson (2016) posed the question: “Can empirical research on the basis of examinees’ responses to test items result in better measurement of the intended constructs?” (p. 6). She encouraged the use of empirical evidence of response processes in item design, item selection, and test design, as a way to better control construct representation and improve score interpretation and use.

Most notably, on items where the reference is to others (perhaps of concern to individuals from collectivistic community traditions) rather than to self, the lower response levels (rarely and sometimes) are associated with thresholds that are higher for American Indian and Latino students. This suggests that responses at these levels will be associated with higher levels of *Social Competence*. However, the highest threshold (between often and almost always) is lower

for American Indian and Latino students, indicating that reporting these others-referent items at the highest level (almost always) is associated with lower levels of *Social Competence*. This suggests that for American Indian and Latino students (potentially from more collectivistic communities), it doesn't take as much social competence to rate these items (with the reference to others) at the highest level (such that it is almost always true for them), as it does for White students.

The motivation for this investigation was inspired by the interpretive argument for *Social Competence* and the use of this measure in diverse communities facing persistent challenges. We find that the items function as expected, given the relevant item features across cultural communities in a way that is consistent – the others referent results in variation in item functioning for students from collectivistic communities. The balance of item referents between self (in four items) and others (in four items) further supports the individualistic and collectivistic orientations (respectively). As a concern regarding measurement development and construct representation, the balance of item referents seems appropriate and should remain an aspect of score interpretation. We worry, however, about the appropriateness of scaling based on the mix of students in different racial/ethnic communities, since scoring based on item thresholds estimate for White students appears to have a negative impact on American Indian students.

References

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

Albano, A.D., & Rodriguez, M.C. (2013). Examining differential math performance by gender and opportunity to learn. *Educational and Psychological Measurement, 73*(5), 836-856.

Bates, D., Maechler, M., Bokler, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48.

Benson, P.L. (1990). *The troubled journey: A portrait of 6th to 12th grade youth*. Minneapolis, MN: Search Institute. Retrieved from http://pub.search-institute.org/file/archive/1990-Benson-Troubled-Journey.pdf

Benson, P.L. (2002). Adolescent development in social and community context: A program of research. In R.M. Lerner, C.S. Taylor, & A. von Eye (Eds.), *New directions for youth development: Pathways to positive development among diverse youth, 95*, 123–147.

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). New York, NY: Wiley.

Brendtro, L.M., Brokenleg, M., & Van Bockern, S. (2002). *Reclaiming youth at risk* (rev. ed.). Bloomington, IN: National Educational Service.

Brown, T. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). New York, NY: Guilford Press.

Chapelle, C.A., Enright, M.K., & Jamieson, J. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice, 29*(1), 3-13.

Collaborative for Academic, Social, and Emotional Learning. (2017). *Continuous improvement: Establish systems for continuous improvement*. Chicago, IL: Author. Retrieved from https://drc.casel.org/continuous-improvement/#resources

Collaborative for Academic, Social, and Emotional Learning. (2018). *Measuring SEL: Using data to inspire practice*. Chicago, IL: Author. Retrieved from https://measuringsel.casel.org/

Cronbach, L.J. (1980). Validity on parole: How can we go straight? In W.B. Schrader (Ed.). *New directions for testing and measurement: Measuring achievement over a decade, No. 5* (pp. 99–108). San Francisco, CA: Jossey-Bass.

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

Eccles, J., & Gootman, J.A. (Eds.). (2002). *Community programs to promote youth development: Committee on community-level programs for youth*. Washington, DC: National Academy Press. Retrieved from https://www.nap.edu/download/10022

Embretson, S.E. (2016). Understanding examinees' responses to items: Implications for measurement. *Educational Measurement: Issues and Practice, 35*(3), 6-22.

Ercikan, K., & Pellegrino, J.W. (2017). Validation of score meaning for the next generation of assessments. In K. Ercikan & J.W. Pellegrino (Eds.), *Validation of score meaning for the next generation of assessments: The use of response processes* (pp. 1-8). New York, NY: Routledge.

Freeberg, A. L., & Stein, C. H. (1996). Felt obligations towards parents in Mexican-American and Anglo-American young adults. *Journal of Social and Personal Relationships, 13,* 457–471.

Greiff, S., & Kyllonen, P. (2016). Contemporary assessment challenges: The measurement of 21st century skills. *Applied Measurement in Education, 29*(4), 243-244.

Jones, S.M., & Kahn, J. (2017). *The evidence base for how we learn: Supporting students' social, emotional, and academic development.* Consensus statements of evidence from the Council of Distinguished Scientists. Washington DC: The Aspen Institute. Retrieved from https://eric.ed.gov/?id=ED577039

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

Kane, M., & Mislevy, R. (2017). Validating score interpretations based on response processes. In K. Ercikan & J.W. Pellegrino (Eds.), *Validation of score meaning for the next generation of assessments: The use of response processes* (pp. 10-24). New York, NY: Routledge.

Kopriva, R.J., & Wright, L. (2017). Score processes in assessing academic content of non-native speakers. In K. Ercikan & J.W. Pellegrino (Eds.), *Validation of score meaning for the next generation of assessments: The use of response processes* (pp. 100-112). New York, NY: Routledge.

Kyllonen, P.C. (2012). *Measurement of 21st century skills within the Common Core State Standards*. Princeton, NJ: Educational Testing Service, Center for K-12 Assessment & Performance. Retrieved from http://search.ets.org/researcher/

Li, Y., Cohen, A. S., & Ibarra, R. A. (2004). Characteristics of mathematics items associated with gender DIF. *International Journal of Testing, 4*, 115–136.

Linacre, J.M. (2016). Winsteps® (Version 3.92.0) [Computer Software]. Beaverton, Oregon: Winsteps.com. Retrieved from http://www.winsteps.com/

MN Department of Education. (2017). *Minnesota Student Survey*. Roseville, MN: Author. Retrieved from https://education.mn.gov/MDE/dse/health/mss/

MN Department of Education. (2018a). *SEL implementation guidance.* Roseville, MN: Author. Retrieved from https://education.mn.gov/MDE/dse/safe/clim/social/imp/

MN Department of Education. (2018b). *Social Emotional Learning*. Roseville, MN: Author. Retrieved from https://education.mn.gov/MDE/dse/safe/clim/social/

Messick, S. (1970). The criterion problem in the evaluation of instruction: Assessing possible, not just intended outcomes. In M. Wittrock and D. Wiley (Eds.), *The evaluation of instruction: Issues and problems* (pp. 183–202). New York, NY: Holt, Rinehart, and Winston.

Muthén, L.K., & Muthén, B.O. (2012). Mplus. (Version 7). [Software program]. Los Angeles, CA: Authors.

Nichols, P., & Huff, K. (2017). Assessments of complex thinking. In K. Ercikan & J.W. Pellegrino (Eds.), *Validation of score meaning for the next generation of assessments: The use of response processes* (pp. 62-74). New York, NY: Routledge.

R Core Team. (2017). *lme4*. Retrieved from https://CRAN.R-project.org/package=lme4

Rhee, E., Uleman, J. S., & Lee, H. K. (1996). Variations in collectivism and individualism by in-group and culture: Confirmatory factor analyses. *Journal of Personality and Social Psychology, 71*, 1037–1053.

Rodriguez, M.C., & Morrobel, D. (2004). A Review of Latino youth development research and a call for an asset orientation. *Hispanic Journal of Behavioral Sciences, 26*(2), 107-127.

Scales, P.C., & Leffert, N. (2004). *Developmental assets: A synthesis of the scientific research* (2nd ed.). Minneapolis, MN: Search Institute.

Search Institute. (2013). *Developmental Assets Profile: Technical summary*. Minneapolis, MN: Author. Retrieved from http://www.search-institute.org/surveys/dap

Van den Noortgate, W., De Boeck, P., & Meulders, M. (2003). Cross-classification multilevel logistic models in psychometrics. *Journal of Educational and Behavioral Statistics, 28*(4), 369–386.

Wang, W.-C., & Wu, S.-L. (2011). The random-effect generalized rating scale model. *Journal of Educational Measurement, 48*(4), 441–456.

Wang, W.-C., Wilson, M., & Shih, C.-L. (2006). Modeling randomness in judging rating scales with a random-effects rating scale model. *Journal of Educational Measurement, 43*(4), 335–353.

Zumbo, B.D. (2007a). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly, 4*(2), 223-233.

Zumbo, B.D. (2007b). Validity: Foundational issues and statistical methodology. In C.R. Rao & S. Sinharay (Eds.), *Handbook of statistics, Vol. 26: Psychometrics* (pp. 45–79). Amsterdam, The Netherlands: Elsevier Science B.V.

Zumbo, B.D., & Rupp, A.A. (2004). Responsible modeling of measurement data for appropriate inferences: Important advances in reliability and validity theory. In D. Kaplan (Ed.), *The SAGE handbook of quantitative methodology for the social sciences* (pp. 73-92). Thousand Oaks, CA: Sage.

Appendix A

Unconditional Partial-Credit Model of the 8-item Measure of Social Competence

```
Generalized linear mixed model fit by maximum likelihood
(Laplace Approximation) [glmerMod]

 Family: binomial  ( logit )
Formula: PCM ~ 1 + (1 | id) + (1 + PCMcategory | Item)


      AIC       BIC    logLik  deviance  df.resid
 816859.9  816951.1 -408421.9  816843.9    661798


Scaled residuals:
    Min      1Q  Median      3Q     Max
-7.1148 -0.8265  0.4455  0.7070  5.4443


Random effects:
 Groups Name             Variance Std.Dev. Corr
 id     (Intercept)      1.2183   1.1038
 Item   (Intercept)      0.1773   0.4211
        PCMcategorycat_3 1.6436   1.2820   -0.70
        PCMcategorycat_4 6.7442   2.5970   -0.76  0.98
Number of obs: 661806, groups:  id, 53089; Item, 8


Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.97989    0.09177   21.57   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


RESPONSE CATEGORY THRESHOLDS

       b1               b2              b3
1 -0.18741793       -0.9010517       -1.260251
2  0.18792600       -1.1825996       -2.521557
3 -0.20236571       -1.3169020       -2.962710
4  0.55059358       -1.7472495       -3.318658
5  0.11872778       -1.2820365       -1.963878
6  0.39386401       -1.4385312       -2.863300
7  0.65760021       -0.6224263       -2.128370
8  0.04386409       -0.9881244       -2.252236
```

Appendix B

Partial-Credit Explanatory Item Responses Model of the 8-item Measure of Social Competence with Item and Person Characteristics

```
Generalized linear mixed model fit by maximum likelihood (Laplace
Approximation) ['glmerMod']

 Family: binomial  ( logit )
Formula: PCM ~ -1 + AmInd + Latino + PCMcategory:others + (1 | id) + (1 +
    PCMcategory | Item)
   Data: mss4
Control: control

      AIC       BIC    logLik  deviance  df.resid
 815531.0  815667.8 -407753.5  815507.0    661794

Scaled residuals:
    Min      1Q  Median      3Q     Max
-7.3093 -0.8264  0.4434  0.7051  5.5674

Random effects:
 Groups Name             Variance Std.Dev. Corr
 id     (Intercept)      1.1771   1.0849
 Item   (Intercept)      0.4945   0.7032
        PCMcategorycat_3 0.3655   0.6045   -0.85
        PCMcategorycat_4 0.9897   0.9948   -0.80  0.93
Number of obs: 661806, groups:  id, 53089; Item, 8

Fixed effects:
                        Estimate Std. Error z value Pr(>|z|)
AmInd                   -0.46505    0.01584 -29.354   <2e-16 ***
Latino                  -0.37309    0.01283 -29.090   <2e-16 ***
PCMcategorycat_2:others  1.49979    0.15742   9.528   <2e-16 ***
PCMcategorycat_3:others  0.78104    0.08319   9.389   <2e-16 ***
PCMcategorycat_4:others -0.02062    0.13469  -0.153    0.878
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


RESPONSE CATEGORY THRESHOLDS
            b1                b2              b3
item1   0.4652550     -0.185473592        0.2547531
item2  -0.6597351      0.251043200        0.5148852
item3   0.4481826     -0.601169735       -1.4460554
item4   1.2047886     -1.034992651       -1.8041176
item5   0.7738789     -0.568876686       -0.4518336
item6  -0.4508910     -0.007038715        0.1695979
item7  -0.1880803      0.806445215        0.9077729
item8  -0.8041190      0.446452232        0.7839180
```

Appendix C

Partial-Credit Explanatory Item Responses Model of the 8-item Measure of Social Competence with Item and Person Characteristics and Item × Person Interaction

```
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation)
[glmerMod]

 Family: binomial  ( logit )
Formula: PCM ~ -1 + AmInd + Latino + PCMcategory:others + PCMcategory:others:AmInd +
    PCMcategory:others:Latino + (1 | id) + (1 + PCMcategory |      Item)
   Data: mss4
Control: control

      AIC       BIC    logLik  deviance  df.resid
 814695.8  814901.0 -407329.9  814659.8    661788

Scaled residuals:
    Min      1Q  Median      3Q     Max
-7.9382 -0.8262  0.4395  0.7003  5.5758

Random effects:
 Groups Name             Variance Std.Dev. Corr
 id     (Intercept)      1.187    1.089
 Item   (Intercept)      2.878    1.696
        PCMcategorycat_3 1.054    1.027    -0.95
        PCMcategorycat_4 3.548    1.884    -0.95  0.98
Number of obs: 661806, groups:  id, 53089; Item, 8

Fixed effects:
                              Estimate Std. Error z value Pr(>|z|)
AmInd                         -0.54417    0.01768 -30.786  < 2e-16 ***
Latino                        -0.44818    0.01432 -31.289  < 2e-16 ***
PCMcategorycat_2:others        2.72828    0.84842   3.216   0.0013 **
PCMcategorycat_3:others        1.54477    0.38969   3.964 7.37e-05 ***
PCMcategorycat_4:others       -0.07820    0.29985  -0.261   0.7942
AmInd:PCMcategorycat_2:others -0.18715    0.03396  -5.511 3.56e-08 ***
AmInd:PCMcategorycat_3:others -0.09641    0.02182  -4.419 9.94e-06 ***
AmInd:PCMcategorycat_4:others  0.39072    0.01962  19.910  < 2e-16 ***
Latino:PCMcategorycat_2:others -0.18673    0.02923  -6.389 1.67e-10 ***
Latino:PCMcategorycat_3:others -0.05216    0.01791  -2.912   0.0036 **
Latino:PCMcategorycat_4:others  0.33467    0.01572  21.288  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

RESPONSE CATEGORY THRESHOLDS
        b1               b2              b3
1  2.00917477      -0.91021767      -1.27764503
2 -0.24424349      -0.08000734       0.02791701
3  1.99121424      -1.32894591      -2.98655462
4  2.74867911      -1.76118618      -3.34207703
5  2.31856394      -1.29398797      -1.98726808
6 -0.03062577      -0.34139285      -0.31901055
7  0.22532019       0.47978281       0.41425111
8 -0.38748133       0.11486292       0.29683094
```