Development and Validation of the Intervention Skills Profile–Skills: A Brief Measure of Student Social-Emotional and Academic Enabling Skills

Stephen P. Kilgus[1], Wes E. Bonifay[2], Katie Eklund[1], Nathaniel P. von der Embse[3], Casie Peet[3], Jared Izumi[2], Hyejin Shim[2], & Lauren N. Meyer[1]

University of Wisconsin-Madison[1]

University of Missouri[2]

University of South Florida[3]

Journal of School Psychology

Note: The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A180515 to the University of Wisconsin-Madison. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education. Correspondence regarding this article should be addressed to Stephen Kilgus at the University of Wisconsin-Madison, Department of Educational Psychology, Madison, WI 53706; email: skilgus@wisc.edu

Abstract

The purpose of this study was to support the development and initial validation of the Intervention Selection Profile (ISP)–Skills, a brief 14-item teacher rating scale intended to inform the selection and delivery of instructional interventions at Tier 2. Teacher participants (*n* = 196) rated five students from their classroom across four measures (total student *n* = 877). These measures included the ISP-Skills and three criterion tools: Social Skills Improvement System (SSIS), Devereux Student Strengths Assessment (DESSA), and Academic Competence Evaluation Scales (ACES). Diagnostic classification modeling (DCM) suggested an expert-created Q-matrix, which specified relations between ISP-Skills items and hypothesized latent attributes, provided good fit to item data. DCM also indicated ISP-Skills items functioned as intended, with the magnitude of item ratings corresponding to the model-implied probability of attribute mastery. DCM was then used to generate skill profiles for each student, which included scores representing the probability of students mastering each of eight skills. Correlational analyses revealed large convergent relations between ISP-Skills probability scores and theoretically-aligned subscales from the criterion measures. Discriminant validity was not supported, as ISP-Skills scores were also highly related to all other criterion subscales. Receiver operating characteristic (ROC) curve analyses informed the selection of cut scores from each ISP-Skills scale. Review of classification accuracy statistics associated with these cut scores (e.g., sensitivity and specificity) suggested they reliably differentiated students with below average, average, and above average skills. Implications for practice and directions for future research are discussed, including those related to the examination of ISP-Skills treatment utility.

Development and Validation of the Intervention Skills Profile–Skills: A Brief Measure of Student Social-Emotional and Academic Enabling Skills

One of the most common ways by which schools support student social-emotional and behavioral (SEB) functioning is via multi-tiered systems of support, such as Positive Behavior Interventions and Supports (Sugai & Horner, 2009). Through these systems, schools provide a continuum of strategies to support students with varying levels of SEB need. At Tier 1, schools provide universal supports to all students in the interest of preventing SEB concerns. It is anticipated these supports will be effective for approximately 80% of students. At Tier 2, schools provide targeted interventions to students who are unresponsive to Tier 1 supports and at-risk for SEB concerns. Tier 2 interventions possess several common characteristics, including (a) efficiency, with supports requiring minimal time and effort to implement; (b) generality, as supports are designed to be broadly applicable and flexible to match specific student needs, and (c) integratable with school infrastructures, allowing for continuous availability to students (Hawken et al., 2009). Finally, at Tier 3, students who have more intense, diverse, and unique needs receive intensive and highly individualized support via a comprehensive intervention plan.

SEB Intervention

Though they vary in scope and intensity, the interventions delivered at Tier 2 and 3 can be organized into two general categories. The first category pertains to instructional interventions, defined as strategies that emphasize teaching positive SEB skills that students can use in lieu of other problem behaviors. Instructional interventions are typically based in a curriculum and are used to target a range of SEB-related skills, including social-emotional skills (e.g., self-awareness, relationship skills; Collaborative for Academic, Social, and Emotional Learning [CASEL], 2005) and academic enabling skills (e.g., academic engagement and study skills; DiPerna, 2006). The second category pertains to contingency management interventions, through which educators manipulate the antecedent and consequences of target behaviors to influence their future frequency. Such interventions include Check In/Check Out and differential reinforcement, which are commonly delivered at Tiers 2 and 3, respectively.

According to the well-established and widely recognized acquisition and performance deficit classification model, each of these two intervention categories is considered appropriate for a particular type of student (Gresham, 1981). Instructional interventions are appropriate for students who engage in problem behavior because they have not learned more appropriate skills by which to attain some desired outcome (Stichter, Malugen, & Davenport, 2018). For example, “Kevin” might push others to get their attention because he has not learned how to appropriately greet peers and initiate social interactions. Such limited skill display could represent either (a) an acquisition deficit, defined as a skill a student has yet to learn, or (b) a fluency deficit, defined as a skill a student has learned to some extent but continues to display inaccurately or in a manner not commensurate with their age or developmental level. In contrast, contingency management interventions are appropriate for students who have learned relevant skills but do not display them due to insufficient motivation. For example, though “Marlene” has demonstrated the ability to raise her hand and wait to be called upon, she still regularly calls out in class because it garners adult attention sometimes faster and more reliably than hand raising.

Research has supported the acquisition and performance deficit classification model as an effective means by which to identify appropriate intervention strategies within both the academic and SEB domains (Elliott, Gresham, Frank, & Beddow, 2008; Martens, Daly, Begeny, & VanDerHeyden, 2011). In addition, studies have supported the efficacy of individual contingency management and instructional interventions for a wide range of students with varying levels of need (e.g., Cook et al., 2008; Maggin, Zurheide, Pickett, & Baillie, 2015). However, studies have further indicated that these interventions are most effective when they are matched to specific student concerns. For instance, repeated studies have indicated contingency management interventions are most effective when they are aligned with the function of a student's problem behavior (McIntosh, Campbell, Carter, & Dickey, 2009; Newcomer & Lewis, 2004). Similarly, instructional interventions are more effective when they target a student's specific acquisition or fluency deficits (Barreras, 2009; Gresham, Cook, & Van, 2006). Below we narrow our discussion to the topic of instructional interventions and the skill assessments that can be used to support their delivery.

SEB Skills Assessment

To support matching of instructional interventions to student needs, it is necessary for educators to conduct SEB skill assessments to determine which acquisition or fluency deficits a student possesses. Relevant lesson plans can then be selected to address the documented deficits. A number of such assessments have been developed to date, such as the *Social Skills Improvement System* (SSIS; Gresham & Elliot, 2008), *Academic Competence Evaluation Scales* (ACES; DiPerna & Elliot, 2000), and the *Devereux Student Strengths Assessment (*DESSA; LeBuffe, Shapiro, & Naglieri, 2014). These and other similar measures are supported by evidence of their psychometric defensibility (DiPerna & Elliott, 1999; Gresham, Elliott, Cook, Vance, & Kettler, 2010). They are also notable for their length, as these scales include a rather large number of items that require a non-trivial amount of time to complete. For instance, the developers estimate it will take a rater approximately 10-15 minutes to complete the SSIS (Gresham & Elliott, 2008) and 10 minutes to complete the DESSA (LeBuffe et al., 2014). Though feasible and appropriate at Tier 3 where fewer students are served and SEB needs are greater, the use of such measures could prove challenging at Tier 2.

To illustrate this challenge, let us imagine an elementary school has elected to use one of these skill assessments with every student identified for Tier 2 support for the purpose of determining the best course of intervention. Assuming an enrollment of 500 students and that 15% of students are identified for Tier 2 intervention (Schanding & Nowell, 2013), it would be necessary for educators to assess 75 students and dedicate 750 to 1,125 minutes of their time. Though this may be feasible at a single point in time, dedication of such time and effort can become challenging over time when paired with other academic and SEB assessment activities.

This implementation challenge has been recognized in the broader area of SEB assessment (DiPerna, Anthony, & Elliott, 2019) and more specifically in relation to SEB universal screening and progress monitoring (Christ, Riley-Tillman, & Chafouleas, 2011; Glover & Albers, 2007). It has more recently been recognized in the area of SEB problem analysis, resulting in efforts to develop brief SEB skill assessments suitable for use at the Tier 2 level (Kilgus, von der Embse, Scott, & Paxton, 2015). One such measure is the *Intervention Selection Profile – Social Skills* (ISP-SS), a seven-item measure designed to inform small-group social skills instruction. To complete the ISP-SS, a teacher uses a four-point Likert scale (0 = *Never* to 3 = *Almost Always*) to indicate the frequency with which a student of interest has displayed seven social skills (e.g., Communication and Responsibility). Item scores are then evaluated in determining which skills corresponded to acquisition deficits, as evidenced by ratings equal to or less than one. Multiple studies have examined the measure to date (Kilgus, Eklund, & von der Embse, 2019; Kilgus et al., 2015). Though promising, support for the ISP-SS has been somewhat inconsistent, with results suggesting the measure predicts certain acquisition deficits better than others. For instance, Kilgus et al. (2019) found that while the ISP-SS accurately detected deficits in communication skills (sensitivity = .81 and specificity = .89), its capacity to accurately detect empathy skill deficits was limited (sensitivity = .44 and specificity = .91). Furthermore, the measure was limited to the social skills domain alone, omitting consideration of other skill types relevant to SEB functioning and academic success (DiPerna, 2006; Zins & Elias, 2007). Recognition of these limitations informed initial development of the novel *ISP-Skills* (Kilgus, von der Embse, & Eklund, 2018), a still brief but expanded measure intended to assess a wider range of SEB skills with greater accuracy and validity.

ISP-Skills

The ISP-Skills is a 14-item measure, intended for use with students who have been identified as requiring intervention (e.g., via teacher referral or universal screening). The measure is designed to detect specific skill acquisition deficits for the purpose of determining (a) whether the student requires an instructional intervention and (b) how to best match the intervention his or her needs.

The ISP-Skills is meant to improve upon the ISP-SS in multiple ways. First, whereas the ISP-SS pertains to student social skills alone, the ISP-Skills is meant to afford information regarding a wider variety of SEB skills. Scholars have developed a number of SEB frameworks (e.g., Collaborative for Academic and Social Emotional Learning [CASEL], Emotional Intelligence, Habits of Mind), each of which identifies distinct but interrelated "non-academic" skills that are key to youth success in school, work, and life (Jones, Bailey, Brush, & Nelson, 2019). Though social skills are typically a part of these frameworks, skills from other domains are also included given their documented relevance to youth success. Stephanie Jones and various colleagues recently endeavored to review these frameworks to identify areas of overlap and distinction (Jones, McGarrah, & Kahn, 2019). The ultimate goal of this "taxonomy project" was to derive broader categories under which a range of non-academic skills could be subsumed. The categories identified through this work included (a) *cognitive skills* through which students direct their behavior in pursuit of some goal (e.g., planning, organization, attention); (b) *social skills* though which students interpret social cues, navigate challenging situations, and establish positive relationships; and (c) *emotional skills* through which a student recognizes and manages their own affective states, while also appreciating and empathizing those of others (EASEL Lab, 2019a, 2019b). The ISP-Skills was designed to assess skills within each of these three major areas. More specific guiding frameworks were then selected to determine which specific skills should be targeted within each of these three categories. The CASEL Five Core Competencies framework was selected given its correspondence to both social and emotional skills, as well as its broad use to inform various aspects of both policy and practice (Eklund, Kilpatrick, Kilgus, & Haider, 2018). The academic enablers framework was then selected given its specification of cognitive-oriented skills through which one achieves academic goals, as well as its previous use to inform assessment tools (DiPerna & Elliott, 1999; DiPerna, Volpe, & Elliott, 2001).

Second, ISP-Skills scaling was designed to enhance the precision and objectivity with which each item is rated. Whereas ISP-SS items are rated using a Likert scale, ISP-Skills items are rated using a behaviorally anchored rating scale (BARS). Likert scale anchors frequently correspond to a single subjective term (e.g., *sometimes*, *almost always*). Accordingly, Likert scale raters are required to make high-inference judgments when selecting among anchor options (Christ & Boice, 2009). In contrast, BARS anchors are less ambiguous, listing and describing discrete behaviors that define and differentiate the anchors (Martin-Raugh, Tannenbaum, Tocci, & Reese, 2016). Within the ISP-Skills, each BARS anchor corresponds to a particular skill level, defined by level of *skill acquisition* (i.e., the degree to which the skill has been learned) and *skill* *utilization* (i.e., the degree to which the skill is used once learned). When taken together, the anchors represent the categories of skill development commonly conceptualized in research and practice, including *acquisition deficit, fluency deficit, performance deficit, typical,* and *strength* (Gresham, Elliott, & Kettler, 2010). See Figure 1 for the ISP-Skills BARS anchors.

Third, ISP-Skills precision and accuracy was designed to be enhanced through scoring founded upon diagnostic classification modeling (DCM; for an overview, see Rupp, Templin, & Henson, 2010). DCM represents a confirmatory multidimensional latent-variable model (Rupp & Templin, 2008). Importantly, DCM emphasizes *within-item* multidimensionality, such that each item is modeled as providing information about one or more discrete latent attributes, rather than the more familiar *between-item* multidimensionality that is common in factor analysis or related methods that aim to identify homogeneous clusters of items or indicators. Thus, DCM posits that an item set may be unidimensional in the traditional sense (i.e., all items may measure a single underlying construct), yet multidimensional due to the combinations of attributes that characterize each item. Though similar to item response theory (IRT), DCM differentiates itself in its conceptualization of latent variables. Through IRT, latent variables represent continuously scaled estimates of student ability, scaled in a manner consistent with *z* scores. Through DCM, latent variables represent estimates of probability of a certain attribute being present, scaled from 0-1. In accordance with this conceptualization, DCM can be used to estimate an attribute profile (i.e., latent class) for each student, specifying which attributes are present and which are not (Rupp & Templin, 2008).

DCM represents a modern and sophisticated method by which to gain actionable evidence about patients, students, and many other individuals (de la Torre, van der Ark, & Rossi, 2018). DCM-based profiles indicate which attributes an individual possesses or does not possess, thereby yielding information with potential implications for treatment planning (pending research supporting such applications). Due to this capability, DCM has been used extensively in relation to cognitive and academic assessments to generate estimates of the probability of skill mastery (Sessoms & Henson, 2018). Researchers have also recently begun to apply these models to mental health-oriented assessments, with scores representing probability of psychological disorder (e.g., de la Torre et al., 2018). Extension of DCM to school-based SEB skills assessment appears logical, as the estimation of skill mastery and the generation of skill profiles is of common concern to educators planning instructional interventions.

Summary and Purpose

To support more effective instructional interventions, it is necessary that educators have access to measures that can inform the matching of instructional content to student needs. Though a number of skill assessments exist, the time and effort required for their completion likely limits their use to Tier 3. Researchers have therefore recently begun to develop new brief skill assessments, such as the ISP-SS. Though previous ISP-SS studies have been promising, evidence has been somewhat inconsistent in relation to certain subscales. Accordingly, recent efforts have focused on revising the ISP-SS while expanding the measure to capture a wider range of SEB-related skills. These efforts have resulted in the novel ISP-Skills measure. The broader goal of this investigation is to inform the development and refinement of the ISP-Skills, as well as validate the measure relative to criterion SEB measures. In accordance with an approach that is common to other skill assessment tools (e.g., SSIS and DESSA), our development and refinement efforts were conducted within the context of a broad and normative sample of students presumed to exhibit the full spectrum of skill levels (e.g., from well below to well above average).

Specific purposes of this investigation were threefold. First, we applied DCM in evaluating the performance of the ISP-Skills in estimating student skill profiles. This analysis resulted in the generation of a variety of item characteristic statistics, student- and sample-level indicators of skill mastery, descriptive statistics indicative of broader test functioning, and statistics indicative of the fit of the chosen DCM model to the observed data. DCM was also used to generate a series of ISP-Skills scores for each student, each of which represented the probability the student had mastered a particular social-emotional or academic enabling skill. These scores were then examined as part of the second purpose of this study, which was to evaluate the criterion-related validity of ISP-Skills scores relative to multiple criterion measures (i.e., the SSIS, DESSA, and ACES). Scores considered within these analyses were continuously-scaled criterion measure subscale scores, as well as DCM-based ISP-Skills scores indicative of skill mastery probability. The third purpose of this study was to evaluate the classification accuracy of DCM-based ISP-Skills scores relative to the aforementioned criterion measures. Of particular interest was the examination of how well the ISP-Skills predicted (1) below average skills, which would likely represent acquisition or fluency deficits necessitating instructional interventions, and (2) above average skills, which would likely represent strengths within a student’s skill repertoire. Our focus on both of these skill levels is grounded in the recommendation that skill assessments should not only support the identification of targets for instructional interventions (i.e., skill deficits), but also strengths that can be leveraged and built upon when addressing these instructional targets (LeBuffe, Shapiro, & Robitaille, 2018).

This study represents an initial and important step in ISP-Skills validation, informing the development, refinement, and initial testing of a tool that can be subjected to further evaluation. To be clear, this subsequent testing will be necessary to support use of the ISP-Skills within applied settings. Given the nature of the ISP-Skills and its intended use within problem analysis, of particular importance will be studies examining ISP-Skills treatment utility, defined as the measure's capacity to promote positive intervention outcomes (Nelson-Gray, 2003).

Method

Participants and Setting

Participants included 196 teachers and 877 students from 11 elementary schools in the Midwest and Southeast (see Table 1). The teachers were predominantly White (82%) and female (92%). Over half of teachers held a Bachelor's degree (53%) and 44% held a Master's degree. Approximately one-third of teachers had 5 years or less teaching experience, 22% had 6-10 years of experience, 18% had 11-15 years of experience, and 29% had been teaching for 16 years or more. All participating teachers served as general education classroom teachers in kindergarten through sixth grade classrooms. Teachers completed behavior ratings for five students in each of their classrooms (see procedures section for more details). Fifty-three percent of the students in the current sample were male and 45% were White, 31% Black, 17% Hispanic/Latinx, 4% Other, 2% multiracial, and less than 1% Asian or Native American. Students ranged from kindergarten through sixth grade with a mean age of 8.39 years (*SD* = 1.9 years).

Across the three participating districts, an average of 14.33% of eligible elementary schools (range = 3–22%) participated in this study (although not all schools were approached for participation). According to data from the National Center for Education Statistics (2019), 55% of the 11 participating schools were designated suburban, 27% were urban, and 18% were rural. Total school enrollments ranged from 262 to 924 (*M* = 529.3, *SD* = 214.2). The percentage of students qualifying for free/reduced-price lunch ranged from 9 to 98% (*M* = 64.2%, *SD* = 36.3%). Five of the six schools from the Midwest site were majority White students, while the sixth was more diverse. The five schools from the Southeast site were all quite diverse, representing either majority Black or Hispanic/Latinx. On average across all 11 schools, 41.9% of students were White (*SD* = 35.2%, range = 3–86%), 29.8% were Black (*SD* = 27.4, range = 5–87%), 22.0% were Hispanic/Latinx (*SD* = 26.0%, range = 2–81%), and 5.4% were multi-racial (*SD* = 3.8%, range = 1–13%). Examination of these various demographic statistics suggest these schools were representative of their respective regions.

Measures

Each teacher completed behavioral ratings for five students within his or her classroom using the ISP-Skills and three criterion measures, each of which is considered a gold standard for skills assessment within its particular skill area. It should be noted that only a portion of the items from each of the three criterion measures were completed for the purposes of this study. Those completed were specific to subscales considered theoretically aligned with anticipated ISP-Skills scale scores. (To note, teachers completed all items represented within each chosen subscale.) The ISP-Skills and each criterion measure are described below.

ISP-Skills. The ISP-Skills (Kilgus et al., 2018) is intended to assess eight skills across two broad domains: Social-Emotional Skills and Academic Enabling Skills (see Figure 2 for an overview of conceptual definitions specific to each skill). Teacher raters use a five-point BARS to rate the frequency with which each student has exhibited a series of skills during the previous month. Note that for the DCM analysis, ISP-Skills ratings were dichotomized so that 0 = *Never/Sometimes-Insufficient Learning/Sometimes-Insufficient Motivation* and 1 = *Often/Almost Always*. Our decision to dichotomize these ratings was made in consideration of the state of the DCM literature. Though DCM-based methods are available for the evaluation of ordinal data, a much wider range of more rigorously evaluated techniques are available for dichotomous data.

Our decision to dichotomize was deemed appropriate given precedence established by alternative social-emotional assessments, which conceptualize *Never* and *Sometimes* ratings as indicative of skill deficits requiring some form of intervention (e.g., SSIS and ACES). We initially considered splitting the two *Sometimes* options, such that *Sometimes-Insufficient Learning* would represent skill non-mastery (0) and *Sometimes-Insufficient Motivation* would represent skill mastery (1). Justification for this approach would be founded in the theoretical expectation that insufficient motivation might not necessarily represent a skill deficit (i.e., non-mastery), but rather a performance deficit in an otherwise mastered skill. However, one could argue that the relative infrequency of skills receiving S*ometimes* ratings might still suggest challenges potentially indicative of skill non-mastery. Thus, in accordance with an over-inclusive scoring approach, both *Sometimes* ratings were coded as “0” within the dichotomization scheme.

The ISP-Skills was initially developed through a multi-step process. First, the ISP-Skills authors created an initial pool of 180 items to be considered for inclusion in the final ISP-Skills measure. Of these, 122 were hypothesized to correspond to Social-Emotional attributes, while the remaining 58 were hypothesized to correspond to Academic Enabler attributes. Within this pool, some items were designed to be specific to one of the eight attributes described above, with the item closely corresponding to the attribute’s conceptual definition. Other items were intended to represent either (1) a subcomponent of an attribute or (2) skills/behaviors relevant to multiple attributes. Justification for this latter approach was founded on the use of DCM-based scoring, which permits the use of multidimensional items that afford information regarding multiple skill attributes. In this way, DCM can support the development of abbreviated measures, as a smaller number of items can still afford information regarding a wide range of skill attributes.

Second, the authors drafted a BARS to be used in rating each ISP-Skills item. The initial BARS draft included four anchors corresponding to Never, Sometimes, Often, and Almost Always. Third, a panel of three experts in SEB assessment was convened to review and provide feedback on the ISP-Skills BARS and item pool. With regard to the BARS, multiple experts recommended the expansion of BARS anchors to support differentiation of fluency and performance deficits. To address this feedback, the authors revised the BARS to include two "Sometimes" scale anchors. The "*Sometimes – Insufficient Learning*" anchor describes a student exhibiting a skill sometimes but with limited fluency. In contrast, the "*Sometimes – Insufficient Motivation*" anchor describes a student exhibiting a behavior with adequate fluency, but only sometimes as a result of apparently limited motivation. It was hypothesized that BARS anchors should be ordered such that "*Sometimes – Insufficient Learning*" preceded "*Sometimes – Insufficient Motivation.*" Initial IRT analyses founded upon a nominal response model supported this ordering (Preston, Reise, Cai, & Hays, 2011). Specifically, a review of category boundary discrimination parameters specific to each BARS anchor ($a_{0\ldots}a_4$) suggested the anchors were ordered in accordance with expectations, such that $a_0 < a_1 < \ldots < a_4$. A review of the resulting category response functions specific to each item also supported the ordering, with graphs resembling typical graded response model findings. See the online supplemental materials to review these category boundary discrimination parameters (Table S1).

Next, in reviewing the ISP-Skills item pool, experts reviewed each of the eight attributes and selected the items they thought were particularly relevant to each. Experts were informed they could select the same item multiple times across the attributes if the item happened to be relevant to more than one attribute. No additional guidance was provided regarding the number of items they could select. From this group of items, the authors selected a subset of items for inclusion in the ISP-Skills form. Fourth, a second group of three SEB assessment experts reviewed ISP-Skills items to develop a Q-matrix, which is a foundational component of any DCM application. This particular process is described in greater detail below.

Social Skills Improvement System-Rating Scales (SSIS). The SSIS (Gresham & Elliot, 2008) was used to assess teachers' perspectives of their students' social skills in the classroom setting. Though the measure includes items specific to three broad areas (Social Skills, Competing Problem Behaviors, and Academic Competence), only the 46 Social Skills items were completed in this study. SSIS Social Skills items are divided across seven subscales, including Communication, Cooperation, Assertion, Engagement, Responsibility, Empathy, and Self-Control. Teachers rate each item using a 4-point Likert scale ranging from 0 = *Never* to 3 = *Almost Always*. The SSIS yields a number of scores, including seven subscale summed scores and a single broad Social Skills standard score (*M* = 100, *SD* = 15). When compared to age-based norms, SSIS scores can be classified into three behavior levels: Below Average (>1 SD below the normative mean), Average (within ±1 SD of the mean), and Above Average (>1 SD above the mean). (Note: combined gender norms were used for all normative scoring within this study.) The broad Social Skills scale was examined in this study. Research has supported the psychometric evidence of scores from this scale, revealing strong score internal consistency (α = .97), test-retest reliability (*r* = .82), and criterion-related validity relative to commonly used rating scales (e.g., Behavior Assessment System for Children, Second Edition [BASC-2] Social Skills scale; *r* = .80]; Gresham & Elliott, 2008; Gresham, Elliott, Cook, Vance, & Kettler, 2010; Gresham, Elliott, Vance, & Cook, 2011).

Academic Competence Evaluation Scales (ACES). The ACES (DiPerna & Elliott, 2000) was used to measure teacher perspectives regarding their students' approaches to learning.

The academic enablers subscale, comprised of 40 items, assesses four subscales: Motivation, Engagement, Study Skills, and Interpersonal Skills. The first three of these subscales were considered within this investigation. Each ACES item is rated using a 5-point Likert scale ranging from 1 = *Never* to 5 = *Almost Always*. Once ratings are completed, item scores are summed to yield subscale scores and an overall Academic Enablers score. Scores are then compared to grade-based norms and classified into three levels: Developing (>1 SD below the normative mean), Competent (within ±1 SD of the mean), and Advanced (>1 SD above the mean). The ACES subscales considered within this study possess strong evidence of various psychometric properties, including internal consistency (median α = 94), test-retest reliability (median *r* = .80), and criterion-related validity relative to the Social Skills Rating System (SSRS; median *r* = .77; DiPerna & Elliott, 1999).

Devereux Student Strengths Assessment (DESSA). The DESSA (LeBuffe et al., 2014) is designed to assess key social-emotional competencies in children kindergarten through grade 8. Teachers completed all 35 items from four DESSA subscales: Self-Awareness, Self-Management, Social Awareness, and Decision Making. Teachers are asked to assess student's behavior over the past four weeks on a 5-point Likert scale ranging from 0 = *Never* to 4 = *Frequently*. To score the DESSA, item scores are summed within each subscale and converted to *T* scores (*M* = 50, *SD* = 10), which are then classified into three levels: Need for Instruction (>1 SD below the normative mean), Typical (within ±1 SD of the mean), and Strength (>1 SD above the mean). The four DESSA subscales of interest possess strong psychometric evidence, including that related to internal consistency (median α = .92), test-retest reliability (median *r* = .93), and criterion-related validity relative to the Behavioral and Emotional Rating Scales–2 (BERS-2; median *r* = .66; LeBuffe et al., 2009; Nickerson & Fishman, 2009).

Procedures

A university-based Institutional Review Board at each study site approved the current study. Researchers contacted administrators in districts who had participated in studies our research team conducted in relation to universal screening for SEB risk. All districts had expressed an interest in building upon their existing screening initiatives through the use of problem analysis measures, which would support the determination of unique student needs to inform interventions. Administrators were asked to identify schools within their district that would be interested in and appropriate for participating in this study. Across the two sites, 11 schools agreed to participate. On average, 58% of teachers within each school chose to participate in this study (*SD* = 30.72%, range = 4-100%).

Study procedures were described to teachers during a staff meeting at each school where all teachers were invited to provide consent if they wished to participate. Parents of students within each classroom were sent an informational letter, which included details on how to opt their child(ren) out of the study if they did not wish for them to participate. Students for whom no form was returned were eligible for participation (99% of the current sample). Teachers completed ratings for five students in their classrooms. Each teacher selected two students who demonstrated behavioral or social-emotional concerns, based on teacher perceptions and no other quantitative information. Researchers then randomly selected the other three students from the class roster using a random number generator. Each teacher was sent a hyperlink to the survey and completed measures electronically via the online survey platform, *Qualtrics*. Measures were given in a counterbalanced fashion, thus removing the likelihood of ordering effects. When completing the criterion assessments for each student (i.e., DESSA, SSIS, and ACES), 20% of items were dropped via a planned missing data design. This approach was used to limit the amount of time and effort required of teachers when completing rating scales.

Q-matrix Development

Perhaps the most essential component of any DCM is an array known as a Q-matrix (Tatsuoka, 1983), which defines the attributes to which each item corresponds. Within a Q-matrix, rows represent items and columns represent the attributes associated with those items. Q-matrix entries of 1 signify attributes that must be mastered for item success, while entries of 0 signify attributes that are not involved in item success. A Q-matrix can be conceptualized in a manner consistent with a structural equation modeling (SEM) measurement model, wherein paths between an item and a latent variable represent a presumed relationship (i.e., a 1 in a Q-matrix) and omitted paths represent the absence of such a presumed relationship (i.e., a 0 in a Q-matrix). Unlike SEM, however, the multidimensionality expressed by the Q-matrix is within rather than between the items. Thus, a complex Q-matrix does not indicate that a test is multidimensional in the traditional sense (i.e., measuring multiple latent traits); instead, a test may be unidimensional (i.e., measuring a single latent trait) despite the presence of multiple interactions/cross-loadings in the Q-matrix (Rupp, Templin, & Hanson, 2010). As one might expect, misspecification of the Q-matrix can have adverse effects on validity (Rupp & Templin, 2008). Specifically, an incorrectly specified Q-matrix may result in biased parameter estimates, which can lead to misclassification of examinees' attribute mastery patterns and inaccurate model fit statistics (Henson, Templin & Wilse, 2009; Rupp & Templin, 2008; Liu, Xu, & Ying, 2012). Hence, correct Q-matrix specification is of great importance.

In the present study, three content experts within the area of SEB assessment developed the ISP-Skills Q-matrix. Each expert was a university professor with a doctorate in school psychology, experience developing and validating SEB assessment tools, and a number of peer-reviewed publications within that area. The development process was conducted via online video conference, through which the experts could view each other, the first author who led the development process, and PowerPoint slides the first author used to guide the meeting. The experts were first provided an overview of the ISP-Skills, including the (a) constituent items; (b) the attributes the ISP-Skills was designed to assess, along with conceptual definitions of each; and (c) the BARS scale. Experts were also given information related to DCM, with a particular emphasis on the role and importance of the Q-matrix.

Next, the experts worked to build the ISP-Skills Q-matrix. The process followed a multi-step expert consensus building procedure, which has been previously used within SEB assessment research (Jaffery et al., 2015). This process was completed twice, including once for the presumed social-emotional items and once for the academic enabler items. First, experts reviewed one ISP-Skills item at a time. Using Slido (www.sli.do), an online polling system, the experts selected the attribute to which they believed each item was related. Each expert's selections remained private until all experts were finished. Second, the experts reviewed the results of the initial attribute selection process. No discussion was necessary if experts were in agreement. In contrast, experts were given the opportunity to discuss their selections if any disagreement was observed. During discussions, experts typically provided a rationale for why they selected certain attributes while not selecting others. Experts also occasionally expressed uncertainty in their selections, indicating they considered other options. The first author made no comment throughout the majority of the discussion but would occasionally summarize or seek clarification from experts if necessary. Third, once the discussion appeared to reach a natural conclusion, the first author provided experts the opportunity to revise their selections in relation to the same item. Revisions were not required and were only carried out if at least one expert expressed interest in doing so. Once the revision process began it proceeded in accordance with the first step described above. Experts could choose to either make the same attribute selections as before or choose a new set of attributes. The experts then moved on to examine the next item once the re-selection process was complete.

Overall, the experts sorted five of the 14 ISP-Skills items with 100% agreement during the first round of selections. For the remaining 9 items, experts were in 100% agreement in their selection of one or two skills for each item, but disagreed with regard to one or more additional skills. In the second round of item sorting, five items were assigned with 100% agreement. The four remaining items were still not sorted with perfect agreement after the second round. For each of these items, unanimous agreement was reached for one or two attributes, though disagreement remained regarding one additional attribute. For three of these items, one expert had made an additional attribute selection with which the other two experts did agree. For the remaining item, two experts agreed in making an additional attribute selection with which the final expert did not agree. Our final Q-matrix specified item-attribute relations (i.e., codes of ‘1’) for which perfect agreement was noted.

As depicted in the Q-matrix, it was presumed ISP-Skills items would demonstrate *within-item* multidimensionality, such that certain items would load on multiple attributes within the ISP-Skills matrix. However, it was further presumed that items would also demonstrate *between-item* unidimensionality, with all nine social-emotional items loading on a single broad factor and all five academic enabler items loading on a separate broad factor. As a precursor to subsequent analyses, this latter presumption was evaluated through a confirmatory factor analysis (CFA), which used tetrachoric correlations and diagonally weighted least squares estimation, while also specifying the two covarying factors. The goodness-of-fit statistics provided further evidence of the between-item unidimensionality for both item sets, RMSEA = .05 (95% CI = [.04, .06]), CFI = .99, TLI = .99, SRMR = .05. Follow-up evaluations of the internal consistency of items within each broad factor revealed high coefficient alphas, with Social-Emotional α = .94 (95% CI = [.94-.95]) and Academic Enablers α = .93 (95% CI = [.92-.94]).

Data Analysis Plan

Research purpose 1. There are many types of DCMs, each with different assumptions about the underlying attributes, their (non-)compensatory relationships, and their interactive effects on item success. de la Torre (2011) developed the generalized deterministic input noisy “and” gate (G-DINA) model, a flexible model that encompasses many of the core DCMs, including the non-compensatory DINA model, the compensatory DINO model, and a number of additive DCMs that do not allow for any attribute interactions.

Mathematically, the G-DINA model partitions the item response function (IRF) into an interaction term, one or more main effects for each underlying attribute, 2-way interactions between each pair of attributes, and so on, up to the *K*-way interaction among all *K* attributes. The G-DINA model is formulated as:

$$P\left(x_{ij} = 1 \middle| \boldsymbol{\alpha}_{ji}^{*}\right) =$$

$$\delta_{j0} + \sum_{k=1}^{K_j^*} \delta_{jk}\alpha_{lk} + \sum_{k'=k+1}^{K_j^*} \sum_{k=1}^{K_j^*-1} \delta_{jkk'}\alpha_{lk}\alpha_{lk'} + \cdots + \delta_{j12\ldots k_j^*} \prod_{k=1}^{K_j^*} \alpha_{lk}.$$

The left side of this equation says that the probability of success (i.e., a response of *x* = 1) on item *j* for individual *i* is conditional on her or his attribute mastery pattern $\boldsymbol{\alpha}_{ji}^{*}$, where 0 = non-mastery and 1 = mastery of a given attribute.. The first term on the right side of this equation is the intercept δ_{j0}, which represents the baseline probability of item success when none of the required attributes have been mastered (i.e., for an individual with $\boldsymbol{\alpha}_{ji}^{*}$= {000}, $P(x_{ij} = 1) = \delta_{j0}$). Thus, δ_{j0} can be conceptualized as the G-DINA guessing parameter.

The δ_{jk} parameters denote the main effect of each attribute α_k, i.e., the change in the response probability (beyond the baseline) due to mastery of each of the attributes associated with the item. In other words, for an individual with $\boldsymbol{\alpha}_{ji}^{*}$= {100}, $P(x_{ij} = 1) = \delta_{j0} + \delta_{j1}$. The $\delta_{jkk'}$ parameter denotes the 2-way interactions between all pairs of attributes that characterize item *j*. More specifically, the $\delta_{jkk'}$ parameters reflect the change in the success probability due to mastering attributes α_k and $\alpha_{k'}$, over and above the additive change due to the main effects (i.e., for an individual with $\boldsymbol{\alpha}_{ji}^{*}$= {110}, $P(x_{ij} = 1) = \delta_{j0} + \delta_{j1} + \delta_{j2} + \delta_{j12}$). Finally, the $\delta_{j12\ldots k_j^*}$ parameter represents the increase in the success probability ascribed to mastery of all *K* attributes required by item *j*, over and above all main and lower-order interaction effects. That is, for an individual with $\boldsymbol{\alpha}_{ji}^{*}$= {111}, $P(x_{ij} = 1) = \delta_{j0} + \delta_{j1} + \delta_{j2} + \delta_{j3} + \delta_{j12} + \delta_{j13} + \delta_{j23} + \delta_{j123}$). The intercept must be positive and main effects are usually positive; however, the interaction effects can be either positive or negative (i.e., $P(x_{ij} = 1)$ cannot exceed 1.0, so if the sum of the main effects is greater than 1.0, then the interaction parameters necessarily will be negative).

Model evaluation. The G-DINA model was evaluated by first considering the overall goodness-of-fit to the observed ISP-Skills data. Following the advice of Maydeu-Olivares (2013), overall fit was assessed via the standardized root mean square residual:

$$SRMSR = \sqrt{\sum_{i<j} \frac{o_{ij} - e_{ij}}{n(n-1)/2}},$$

where o_{ij} is the observed correlation between items *i* and *j*, e_{ij} is the expected (i.e., model-implied) correlation between items *i* and *j*, and *n* is the number of items. Thus, the SRMSR fit statistic summarizes the differences between the observed and expected Pearson correlation of each item pair (i.e., the standardized residuals). Maydeu-Olivares recommends SRMSR ≤ 0.05 as the criterion for good fit when analyzing ordinal data.

The G-DINA model was also evaluated in terms of item-level fit. To assess the degree to which the specified model and Q-matrix represented the observed item responses, 10,000 attribute patterns were sampled from the posterior distribution of the attributes (described in more detail below). These resampled attribute patterns were then used in conjunction with the estimated model parameters to generate predicted item response patterns. The fit of each individual item was then estimated by comparing the observed proportion of *Often/Almost Always* ratings on that item in the sample to the expected proportion of *Often/Almost Always* ratings on that item across all resamples. Item fit was also evaluated by assessing the success with which the model was able to recover the observed correlations among all item pairs. In both the univariate and bivariate analyses, item-level goodness-of-fit would be supported by values close to zero (i.e., minimal differences between the actual response patterns and those reproduced by the model; Chen, de la Torre, & Zhang, 2013).

The expected mastery pattern (i.e., latent class membership) of each student was estimated using maximum a posteriori (MAP) scoring. MAP scoring is a Bayesian estimation method that combines the likelihood of the observed data with a prior distribution of the parameter(s) of interest. The resulting posterior distribution then characterizes the most precise estimate of the true parameter value or vector, as well as the amount of uncertainty about that estimate. In general, the MAP estimate is the mode of the posterior distribution. More specifically, in the G-DINA model, the MAP value of attribute mastery pattern $\boldsymbol{\alpha}_i$ of individual *i* is given by $\widehat{\boldsymbol{\alpha}}_i = \arg\max[P(\boldsymbol{\alpha}_c|\mathbf{X}_i)]$, where $P(\boldsymbol{\alpha}_c|\mathbf{X}_i)$ is proportional to the likelihood function

$L(\mathbf{X}_i|\alpha_c)$ times the prior probability $P(\boldsymbol{\alpha}_c|\mathbf{X}_i)$. The posterior probability can also be used to compute the probability $\hat{p}_{ik}$ that individual *i* has mastered attribute *k* (Wang, et al., 2015). In DCM, these posterior probabilities are aggregated to obtain the marginal skill probabilities, which are used to determine class membership (i.e., the likelihood of possessing each skill mastery pattern).

Once $\widehat{\boldsymbol{\alpha}}_i$ has been estimated, it is then possible to evaluate the classification accuracy of the G-DINA model. By comparing the observed (MAP-based) classification probabilities to those implied by the model, one can estimate classification accuracy in terms of attributes (i.e., classification as a non-master or master of a particular attribute), patterns (e.g., classification as a member of the [00], [01], [10], or [11] classes of 2-attribute item), and the overall test instrument. The percentage of agreement between the observed and expected classifications indexes the accuracy of each of these classifications (Wang, et al., 2015).

Q-matrix validation. Within the G-DINA framework, it is possible to empirically validate the Q-matrix (de la Torre & Chiu, 2016). Statistical validation of the Q-matrix (which is comprised of *K* q-vectors) is based on twofold reasoning. First, a q-vector is deemed appropriate if it results in latent classes with homogeneous within-group success probabilities. Second, among the appropriate q-vectors, the correct one is that which contains the fewest attribute specifications. To quantify these two definitions, de la Torre and Chiu developed a discrimination index (ς^2) that can be used (after the domain experts have developed the initial, theory-driven Q-matrix) to identify and replace any misspecified Q-matrix entries:

$$\varsigma_j^2 = \sum_{c=1}^{2^{K_j^*}} w\left(\boldsymbol{\alpha}_{cj}^*\right)[P\left(\boldsymbol{\alpha}_{cj}^*\right) - \bar{P}_j]^2,$$

where $w\left(\boldsymbol{\alpha}_{cj}^*\right)$ is the posterior probability of a given attribute pattern for item *j* (referred to as the *reduced* attribute pattern), $P\left(\boldsymbol{\alpha}_{cj}^{*}\right)$ is the item success probability associated with that pattern, and $\bar{P}_j$ is the mean success probability for item *j*. The discrimination index is therefore the weighted variance of correctly answering item *j*, given the distribution of a particular reduced attribute pattern. Thus, there would be a separate ς_j^2 for every possible reduced attribute pattern within each item *j*.

The Q-matrix validation process involves a search algorithm that determines $\hat{\varsigma}_j^2$ for all possible q-vectors, stopping after computing $\hat{\varsigma}_{j1:K}^2$, the discrimination index associated with mastery of all required attributes. $\hat{\varsigma}_{j1:K}^2$ will maximize discrimination, but it represents the most complex attribute specification and will therefore serve as a reference value by which to select a more parsimonious, but similarly appropriate q-vector. Identification of appropriate q-vectors is based on the proportion of variance accounted for (PVAF) by a particular ς_j^2 relative to $\hat{\varsigma}_{j1:K}^2$. Specifically, when $\hat{\varsigma}_j^2/\hat{\varsigma}_{j1:K}^2 \geq \epsilon$, where ϵ is a user-defined PVAF, then the q-vector associated with $\hat{\varsigma}_j^2$ is considered to be appropriate. In the present study, PVAF was set at .95 to identify any relatively simple q-vectors that accounted for at least 95% of the variance accounted for by the most complex q-vector. For example, suppose the original expert-constructed attribute specification for item *j* was q = {010} with $\hat{\varsigma}_j^2$ = 0.15, while the maximum discrimination (associated with pattern q = {111}) was $\varsigma_{j1:K}^2$ = 0.21. The ratio $\hat{\varsigma}_j^2/\hat{\varsigma}_{j1:K}^2$ = 0.15/0.21 = 0.71, which is below the prespecified PVAF, so the search algorithm would try to uncover a more appropriate q-vector. For example, q = {011} may yield $\hat{\varsigma}_j^2$ = 0.20, which, when compared to the reference q-vector, would exhibit a satisfactory PVAF of $\hat{\varsigma}_j^2/\hat{\varsigma}_{j1:K}^2$ = 0.20/0.21 = 0.95. The Q-matrix validation process proceeds in this manner until all appropriate and correct q-vectors have been identified.

Nesting. It should be acknowledged that the current data are nested in nature, with student scores clustered within classrooms. Only a few studies on multilevel DCMs that would account for this nested structure have been published to date (e.g., Huang, 2017; Wang & Qiu, 2019). Both works advanced a multilevel extension of the DINA DCM model, which only allows for non-compensatory skill interactions. Wang and Qiu note that a future direction of their work would be a consideration of multilevel modeling in the context of the G-DINA model, which was used in this paper. Unfortunately, that extension has not yet occurred.

However, to investigate the influence of nesting, we ran a multilevel IRT model with fixed items and random intercepts for the grouping variable of school. The intraclass correlation coefficient (ICC) for this model, which indicates the influence of school membership on the latent trait, was just .007. In other words, students' classroom membership only accounted for 0.7% of the variability in their levels of social-emotional and academic enabling skill. This low ICC suggests that the nested structure of the data can safely be ignored (Pornprasertmanit, Lee, & Preacher, 2014).

Research purpose 2. As described above, DCM was used to estimate an attribute profile for each student. Here, rather than rounding to a binary 0/1 mastery classification as is typical in DCM, the profiles used the original (unrounded) marginal probabilities were used to enhance the accuracy of the following analyses. Each profile included eight probability scores representing the likelihood the student had mastered the eight proposed ISP-Skills factors. The criterion-related validity of these scores was examined relative to three criterion measures, including the SSIS, DESSA, and ACES. Criterion scores of interest were continuously-scaled scale scores, expressed in the form of either summed item scores (ACES), T scores (DESSA; *M* = 50, *SD* = 10), or standard scores (SSIS; *M* = 100, *SD* = 15). The relation between ISP-Skills and criterion scores was evaluated via Spearman's rho (ρ) correlation coefficients. The non-parametric Spearman's ρ coefficients were preferred over the more common Pearson's product-moment (*r*) correlation coefficient given the presumed non-normal distribution of ISP-Skills probability scores. Based upon previous work (de la Torre et al., 2018), it was anticipated probability score distributions would be bimodal, with modes at the lower and upper ends of the distribution, representing near certainty regarding the absence or presence of skill mastery (respectively).

In evaluating correlational findings, emphasis was placed upon the comparison of expected convergent and discriminant relations. Convergent relations were defined as those between an ISP-Skills scale and its most closely theoretically-aligned criterion measure scale. Discriminant relations were then between an ISP-Skills scale and any other criterion measure scale. It was anticipated all correlations would be positive and at least small (>.10) or medium (>.30) in magnitude, mostly due to common method and informant variance, as well as the presumed interrelatedness of all social-emotional and academic enabling items (DiPerna, 2006; Ross & Tolan, 2018). However, it was further anticipated convergent correlations would be large (>.50) and exceed those of discriminant correlations given the enhanced theoretical alignment between scales represented within the former coefficients.

Research purpose 3. Receiver operating characteristic (ROC) curve analysis was used to evaluate the classification accuracy of ISP-Skills scores relative to criterion scales. All ROC curve analyses were conducted twice. The first set of analyses examined ISP-Skills capacity to predict below average skills. Within these analyses, the outcome variable corresponded to a dichotomous criterion measure score, where 1 = Below Average and 0 = Average or Above Average. The second set of analyses examined ISP-Skills capacity to predict above average skills. The outcome variable once again corresponded to a dichotomous criterion measure score, where 1 = Above Average and 0 = Below Average or Average. The ISP-Skills score and criterion measure examined within each ROC curve analysis were the same as those represented in the convergent relations described above relative to Research Purpose 2. See Table 2 for the percentage of students falling within each behavioral level across the three criterion measures.

The broader purpose of the ROC curve analyses was to identify suitable cut scores around which ISP-Skills scales could be dichotomized in support of decisions regarding the presence of above or below average skills. A series of statistics were calculated to evaluate each possible cut score's classification accuracy. Sensitivity (SE; true positive rate) is defined as the proportion of students who truly possessed the condition of interest (e.g., below average skills) who were correctly identified as such via the ISP-Skills. In this scenario, the criterion measure defined the "truth" regarding the condition of interest. Specificity (SP; true negative rate) is defined as the proportion of students who truly did *not* possess the condition of interest who were correctly identified as such. Positive predictive values (PPV) represent the proportion of students identified as possessing the condition who truly did possess it. Negative predictive values (NPV) represent the proportion of students identified as *not* possessing the condition who truly did not possess it. Finally, correct classification (CC) is defined as the proportion of overall students who were correctly identified as either possessing or not possession the condition.

There are no definitive guidelines by which to evaluate these statistics with respect to "acceptable" classification accuracy. This is particularly true of PPV and NPV due to their sample dependence, such that PPV tends to be higher when a condition of interest is more prevalent while NPV tends be higher when the condition is less prevalent. Multiple heuristics have been proposed for SE and SP. In accordance with previous skills assessment research (Kilgus et al., 2015, 2019), we applied acceptability thresholds of SE ≥ .80 and SP ≥ .70.

Next, the Youden *J* index was used to support identification of a suitable cut score for each ISP-Skills scale. Through this approach, the Youden index is calculated for each observed score within a scale, such that *J* = SE + SP – 1. The observed score then selected to serve as a cut score is that possessing the highest *J* value. Given its formulation, the Youden index equally weighs the costs associated with false positive decisions (=1 – SP) and false negative decisions (=1 – SE); accordingly, by using the statistic, a test developer is attempting to maximize both SE and SP (Smolkowski & Cummings, 2015).

A final classification accuracy statistic corresponded to the area under the curve (AUC, with corresponding 95% confidence intervals). AUC statistics are regarded as effect size-type indicators of a measure's overall classification accuracy. The statistic is interpreted as the likelihood that a randomly selected at-risk individual would have a more at-risk score than a randomly selected not at- risk individual. An AUC equal to .50 indicates a measure's accuracy is no better than chance, while an AUC equal to 1.00 indicates a measure possessing perfect accuracy. Common interpretive guidelines define AUCs of .50–.69 as low, .70–.89 as moderate, and .90–1.00 as high (Streiner & Cairney, 2007).

Missing data. As noted above, we employed a planned missing data design, such that teachers only completed 80% of the items within each criterion measure. Which items were dropped was randomly determined for each student, supporting a conclusion that these data were missing completely at random (MCAR; Enders, 2010). No other unplanned missing data were present within the dataset. Expectation maximization, a single imputation method, was used to handle missing criterion measure data when conducting analyses specific to Research Purposes 2 and 3. This particular approach was considered appropriate given the presumed MCAR nature of the planned missing data (Gold & Bentler, 2000).

Results

Research Purpose 1: Apply DCM to ISP-Skills Data

Q-matrix validation. Table 3 presents the original Q-matrix (as constructed by the team of content experts) and the estimated Q-matrix (as suggested by the Q-matrix validation algorithm) for the social-emotional items. Entries of 1 in Table 3 denote consensus among the experts regarding the particular attribute(s) that a student must master to achieve an *Often/Almost Always* rating on an item; entries of 0 denote attributes that are not associated with the item. The asterisks in Table 3 indicate two modifications that were suggested by the Q-matrix validation algorithm. More specifically, the entries of 1* indicate that the experts included two unnecessary attributes. To assess whether the modified Q-matrix actually improved upon the original Q-matrix, the G-DINA model was fit to the data using each matrix. The modified Q-matrix resulted in slightly better fit relative to the original, with the former G-DINA model yielding lower AIC and BIC values. However, a chi-square test revealed that this difference in fit was not statistically significant, $\chi^2(8) = 8.59, p = 0.38$. Accordingly, all findings described below are based on the original Q-matrix as assembled by the team of experts. Table 4 presents the expert-created Q-matrix for the academic enabler items. The absence of asterisks indicates that the Q-matrix algorithm perfectly validated the original Q-matrix that was arranged by the expert team.

Parameter estimates. The G-DINA parameter estimates are presented in Tables 5 (social-emotional items) and 6 (academic enabler items). Note that the items in Table 5 are listed according to the number of attributes rather than simple numerical ordering. The first column presents the probability of earning an *Often/Almost Always* rating when none of the requisite attributes have been mastered (i.e., δ_0, the baseline probability). Most items had baseline probabilities close to zero; an exception was Item 8 (“Initiates or joins activities with peers”), for which students had a $\hat{\delta}_0$ = .38 probability of receiving an *Often/Almost Always* rating even if they had not mastered Attribute A4 (relationship skills). The δ_k parameters in columns 2-4 of Table 5 represent the change in the baseline probability due to mastering a single attribute (i.e., the main effects), where applicable. As an example, the probability of receiving an *Often/Almost Always* rating on Item 1 increased by .90 if the student had mastered Attribute A1 (self-awareness). Notice that the first and third main effects on Item 9 did not affect the baseline probability at all; that is, Attributes A3 (self-management) and A5 (responsible decision-making), on their own, had zero impact on the probability of success.

The $\delta_{kk'}$ parameters in columns 5-7 present the pairwise interaction effects among the attributes (i.e., the change in the baseline probability due to mastery), over and above the additive impact of mastering each attribute in the pair. For instance, the probability of success on Item 3 increased beyond the baseline (.01) by $\hat{\delta}_1$ =.24 if Attribute A1 (self-awareness) is mastered, and by $\hat{\delta}_2$ =.58 if Attribute A3 (self-management) is mastered; if both were mastered, then there was a change in the baseline probability, beyond the additive effect of A1 and A3 (.24 + .58 = .82), of $\hat{\delta}_{12}$ = .12. Thus, the G-DINA model estimated the probability of earning an *Often/Almost Always* rating on Item 3, when Attributes A1 and A3 were both mastered, as $\hat{\delta}_0 + \hat{\delta}_1 + \hat{\delta}_2 + \hat{\delta}_{12}$ = .01 + .24 + .58 + .12 = .95. Note that the interaction parameter estimate will be negative whenever the main effects sum to a value greater than 1.0.

Finally, δ_{123} estimates in the rightmost column represent the three-way interaction, which is the change in the baseline probability due to mastery of all required attributes, over and above the additive impact of the main and two-way interaction effects. The $\hat{\delta}_{123}$ values are interpreted just as the lower-order interaction effects. For example, if Attributes A3 (self-management), A4 (relationship skills), and A5 (responsible decision-making) have all been mastered, then the probability of receiving an *Often/Almost Always* rating on Item 9 can be computed as $\hat{\delta}_0 + \hat{\delta}_1 + \hat{\delta}_2 + \hat{\delta}_3 + \hat{\delta}_{12} + \hat{\delta}_{13} + \hat{\delta}_{23} + \hat{\delta}_{123}$ = .01 + .00 + .37 +.00 + .45 + .90 + .09 -.83 = .99.

The G-DINA results can perhaps be clarified by plotting the conditional response probabilities. Figure 3 presents the probability of success (i.e., receiving an *Often*/*Almost Always* rating) on the nine social-emotional items, given the student's attribute profile. A simple example is provided by Item 1. The upper left panel in Figure 3 indicates that the probability of earning an *Often*/*Almost Always* rating on Item 1 is .98 for students who have mastered Attribute A1 (self-awareness) but only .08 for students who have not mastered this attribute. A similar interpretation can be made for Item 2. The results are more informative when item success is contingent on mastering multiple attributes. The upper right panel presents the success probabilities for Item 3 ("Monitors own emotions and controls his/her behavior"), conditional on the student's (non-)mastery of the underlying attributes: A1 (self-awareness) and A3 (self-management). The probability of earning an *Often/Almost Always* rating is .01 for students who have mastered neither skill and .95 for those who have mastered both. Further, the probability of success is .25 if students have only mastered self-awareness and .59 if they have only mastered self-management. This difference reflects a partially compensatory relationship between the attributes: although self-management mastery on its own is not associated with an especially high probability of item success, it will partially compensate for a lack of self-awareness. The inverse relationship, however, does not hold.

The most complex probability patterns in Figure 3 are related to the social-emotional items that were characterized by three underlying attributes. Items 6 is associated with attributes A3 (self-management), A4 (relationship skills), and A5 (responsible decision-making). For Item 6, the relationship among these attributes was disjunctive, meaning that failure to achieve an

Often/Almost Always rating was only likely among students in attribute class 000 (non-mastery of all three attributes). Mastery of any of the three attributes was associated with a moderately high probability (.62 or higher) of earning an affirmative rating on this item. Closer inspection of the conditional response probabilities reveals that the highest probability of helping others was related specifically to self-management and relationship skills: patterns 110 and 111 were both associated with a .99 probability of earning an *Often/Almost Always* rating on Item 6.

A particularly nuanced pattern was estimated for Item 9. This item involved the same attributes as Item 6, though they interacted in unique ways to influence the probability of an *Often/Almost Always* rating. One noticeable difference is that no single attribute was sufficient for earning a high rating on Item 9. However, while self-management mastery alone only had a .01 probability of success, it was found to interact with each of the other attributes in achieving a high rating. When self-management and relationship skills were both mastered (pattern 110), the probability of an *Often/Almost Always* rating was .82; when self-management and responsible decision-making were both mastered (pattern 101), the probability increased to .91; and when all three were mastered (pattern 111), the probability was .99. Conversely, students who had mastered the latter two attributes, but did not possess self-management skills (pattern 011) only had a .46 probability of *Often/Almost Always* responding appropriately to others.

Finally, Table 6 displays the G-DINA parameters that were estimated from the academic enabler item data. Here, all items were characterized by a single attribute, which greatly simplifies interpretation of the results. Each item includes two parameters: δ_0 (the baseline probability of earning an *Often/Almost Always* rating) and δ_1 (the change in that baseline if the single attribute was mastered). Summing these estimates provides the probability of item success given mastery of the attribute. For example, there was only a $\hat{\delta}_0$ = .03 probability of earning an

Often/Almost Always rating on Item 10 when Attribute A6 (study skills) was not mastered. If a student had mastered study skills, then there was a $\hat{\delta}_1$ = .89 increase in the probability of earning a high rating. As shown in the rightmost column of Table 7, the probability of receiving an *Often/Almost Always* rating was then $\hat{\delta}_0 + \hat{\delta}_1$ = .03 + .89 = .94.

Descriptive statistics. The parameter estimates from a DCM can be used to describe various aspects of the test as well as the objects of measurement (in this case, the ISP-Skills and the student sample, respectively). At the test-level, one can explore the prevalence of attribute mastery in the sample. Table 7 presents each of the eight attributes from the Q-matrix and their estimated proportions of (non-)mastery in the present sample. For example, based on the observed rating patterns to all of the social-emotional item that involved Attribute A1 (self-awareness), the G-DINA model estimated that 51.4% of the students in this sample had mastered self-awareness. Attributes A1, A2, A3, A4, and A7 each had higher proportions of mastery, suggesting that these attributes were "easier" to master. Attributes A5, A6, and A8, however, had higher proportions of non-mastery, indicating the "difficulty" of these attributes. Attribute A5 (responsible decision-making) in particular was especially difficult to mastery: only an estimated 21.3% of the students exhibited mastery of responsible decision-making.

Another way to summarize the results is by focusing on student-level descriptive statistics. Table 8 displays the *Often/Almost Always* rating probabilities for each of the 2^5 = 32 possible patterns of mastery across the five attributes underlying the nine social-emotional items. Students with mastery pattern $\boldsymbol{\alpha}$ = {11100}, for example, had a high success probability (P > .95) on Items 1, 2, 3, and 7, a moderate success probability on Items 4 (P = .48) and 6 (P = .68), a low success probability on Items 5 (P = .20) and 8 (P = .38), and essentially no probability of success on Item 9 (P = .01). In other words, students who had mastered self-awareness, social awareness, and self-management, but not relationship skills or responsible decision-making, were highly likely to earn *Often/Almost Always* ratings on the items pertaining to emotions and sharing objects, and less likely to earn such ratings on items related to interacting with others.

Table 9 presents the rating probabilities on the academic enabler items, conditional on the $2^3 = 8$ possible combinations of the three underlying attributes. As an example, students who had mastered Attributes A6 and A7, but not A8 (i.e., mastery pattern $\boldsymbol{\alpha} = \{110\}$) were likely to earn an *Often/Almost Always* rating on Items 10, 11, and 12, and unlikely to earn a high rating on Items 13 and 14. Inspecting the content of these items reveals that students who had mastered study skills and academic engagement, but not motivation, were likely to receive high ratings on items about academic preparation and participation, and unlikely to receive high ratings on items about showing interest and independence.

Model evaluation. The G-DINA coefficients and descriptive statistics demonstrate the unique perspective afforded by DCM analysis. However, in order to draw meaningful inferences from these results, it is essential to demonstrate that the model closely reflects the observed data. The overall model-data fit of a DCM can be established via the SRMSR described earlier. Results indicated that the G-DINA models, based on the expertly constructed Q-matrices, yielded good fit relative to the observed social-emotional and academic enabler item rating patterns, SRMSR = .031 and .032, respectively.

Table 10 presents the univariate item fit statistics, i.e., the expected and observed proportions of *Often/Almost Always* ratings, along with *Z*-test associated *p*-values. Differences between the observed item success proportions and the model-implied probabilities of success were less than .003 for all items; accordingly, all *p*-values were well above .05, indicating failure to reject the null hypothesis of no difference between the expected and observed proportions.

Thus, the fitted G-DINA models with the expert-specified Q-matrices were able to account for the actual response proportions of each item.

Although not shown here due to space considerations, the G-DINA models were also able to reproduce almost all of the pairwise correlations among the items. For the social-emotional items, only two of the 36 item pairs were not accounted for by the model; specifically, the *Z*-test revealed a significant difference between the observed and expected correlations between Items 3 and 4 ($p < .001$) and Items 7 and 8 ($p = .02$). For the academic items, the correlation between Item 2 and 5 ($p = .01$) was the only correlation that was not precisely reproduced by the G-DINA model. In sum, the near-perfect modeling of the univariate item success proportions and the satisfactory reproduction of almost all pairwise correlations offer strong support for the item-level fit of the G-DINA models that were specified in this study.

Classification accuracy. Finally, because DCMs are (constrained) latent class models, it is possible to estimate the classification accuracy of each attribute mastery pattern, of the attributes themselves, and of the overall test. Regarding the social-emotional attributes, the last two columns in Table 8 present the estimated counts of each mastery pattern, based on MAP estimation, and the pattern-level classification accuracy. The largest classes, by far, characterized students who expressed mastery or non-mastery of all five attributes. The first row, for instance, indicates that 393 students (44.7%) were estimated, with .99 accuracy, to possess mastery pattern $\boldsymbol{\alpha} = \{11111\}$. The second most likely pattern was $\boldsymbol{\alpha} = \{00000\}$, which was estimated for 279 students (30.6%) with classification accuracy of .97. Several smaller yet notable patterns also emerged. For example, 30 students (3.4%) were estimated to have pattern $\boldsymbol{\alpha} = \{00011\}$, which denotes a lack of self- and social awareness and self-management, but mastery of relationship skills and responsible decision-making. Notably, classification accuracy was only .69 for this mastery pattern. Such a finding could be related to the smaller frequency of this pattern with the sample. However, it is also worth noting that the classification accuracy was estimated as .00 for 13 of the 32 possible patterns because these patterns were not observed. The low prevalence of patterns defined by a mix of skill mastery and non-mastery (e.g., $\boldsymbol{\alpha}$ = {11100}) could suggest a problem with the capacity of the ISP-Skills to differentiate among various skill deficits.

The note at the bottom of Table 8 presents the attribute-level classification accuracy for the social-emotional items. Mastery vs. non-mastery classification was precise for all five attributes, with accuracies ranging from .94 to .97. Finally, the overall classification accuracy of the 9-item instrument was .885, indicating that the G-DINA model of the social-emotional items was successful in accurately classifying students into the correct mastery pattern.

The pattern frequencies and classification accuracies of the academic enabler items are displayed in the last two columns of Table 9. Three main mastery patterns emerged. As in the previous analysis, patterns of complete mastery and non-mastery were the most frequent: $\boldsymbol{\alpha}$ = {111} was estimated for 371 students (42.2%) with .99 accuracy and $\boldsymbol{\alpha}$ = {000} was estimated for 354 students (40.2%) with .93 accuracy. Surprisingly, 127 students were classified, with .94 accuracy, as having pattern $\boldsymbol{\alpha}$ = {010}, denoting mastery of academic engagement, but no study skills or motivation. The table note also indicates that the mastery and non-mastery classification of each of the three attributes was highly accurate, as was the overall test-level classification.

Research Purpose 2: Criterion-related Validity

See Table 11 for a summary of correlations between ISP-Skills probability scores and various criterion measure subscale scores. With one exception, all correlations exceeded the threshold for “large” correlations (>.50), indicating each ISP-Skills score was strongly related to all social-emotional and academic enabling skills. Further examination of correlational findings indicated that ISP-Skills social-emotional scores were more strongly related to criterion social-emotional subscales than academic enabling subscales. The converse was also true, such that ISP-Skills academic enabling scores were more strongly related to criterion academic enabling subscales. Finally, results indicated the expected pattern of convergent and discriminant relations did not emerge, as (1) hypothesized discriminant correlations tended to be just as large as hypothesized convergent relations, and (2) some hypothesized discriminant correlations exceeded those of hypothesized convergent correlations for certain ISP-Skills subscales.

Research Purpose 3: Classification Accuracy

Below average skills. The Youden index was used to detect suitable cut scores along each ISP-Skills probability score scale, which could be used to differentiate students with below average skills (=1) from students with average or above average skills (=0). See Table 12 for a list of selected cut scores across scales, as well as the classification accuracy statistics associated with each. Results indicated the majority of selected probability cut scores approximated zero (.01–.06). A sole exception was the Relationships Skills scale, which yielded a cut score of .41. Across cut scores, SE and SP values consistently exceeded the thresholds for acceptable performance (≥.80 and ≥.70, respectively), with SE values ranging from .82–.94 and SP values ranging from .70–.86. NPV values (.93–.97) consistently fell above PPV values (.50–.76), though such differentiation was to be expected given the low prevalence of the condition of interest (i.e., below average skills) within the sample. CC values ranged from .74–.86, suggesting that overall, the ISP-Skills correctly classified students with regard to their below average skill status the majority of the time. AUC values indicated that overall, three of the ISP-Skills scales were moderately accurate predictors of below average skills (.84–.89), while the remaining five were highly accurate (.90–.92). To note, none of the 95% confidence intervals associated with these AUC values fell out of the moderate range or even below .80.

Above average skills. The Youden index was once again used to identify cut scores for differentiating students with above average skills (=1) from students with below average or average skills (=0). See Table 13 for a list of selected cut scores and associated classification accuracy statistics. The selected cut score across all ISP-Skills scales was equal to 0.99. SE and SP values once again exceeded the thresholds for acceptable performance, with SE values ranging from .90–.98 and SP values ranging from .74–.81. NPV values (.96–1.00) again exceeded PPV values (.29–.62), reflecting the limited prevalence of above average skills within the sample. CC values ranged from .78–.83, suggesting four of every five students were classified correctly with regard to their above average skill status across scales. AUC values indicated that all eight ISP-Skills scales were moderately accurate predictors of above average skills (.84–.89). Again, none of the AUC 95% confidence intervals fell out of the moderate range or even below .80.

Discussion

The purpose of this study was to develop a brief tool that can be used to inform interventions for SEB concerns. The novelty of the ISP lies within its brevity, allowing it to be used with a wider range of students to match Tier 2 interventions to student concerns rather than waiting for Tier 3 to individualize intervention. The three specific purposes of this investigation were to (1) evaluate the performance of the ISP-Skills in estimating student skill profiles through DCM, (2) examine the criterion-related validity of ISP-Skills scores relative to multiple criterion measures, and (3) evaluate the classification accuracy of ISP-Skills scores.

Research Purpose 1

The first research purpose was to evaluate ISP-Skills item functioning via a DCM framework. An initial step in this process was to evaluate the fit of the expert-driven Q-matrix and chosen G-DINA model to the observed data. Results suggested the Q-matrix fit the data well, thereby supporting the proposed model of relations between ISP-Skills items and hypothesized attributes. More specifically, an examination of the estimated Q-matrices suggested only slight modifications were needed to the expert-driven social-emotional Q-matrix to better represent the relation between items and the five attributes. Furthermore, no modifications were necessary for the expert-driven academic enablers Q-matrix. Given findings indicating the estimated Q-matrix fit no better than the expert-driven counterpart, one might conclude the experts did a good job representing how the various attributes were related to "success" on each item (i.e., ratings of *Often/Almost Always*). Conversely, one might also conclude the Q-matrix was an appropriate representation of which items predicted the probability of social-emotional and academic enabler skill mastery.

DCM results also supported the fit of G-DINA model to ISP-Skills data at both the test and item levels. At the test level, SRMSR fit statistic fell in the acceptable range. At the individual item level, the G-DINA model was found to generate accurate response pattern estimates, defined as the proportion of individuals earning Often/Almost Always ratings on items. The G-DINA model also yielded accurate pairwise inter-item correlations, with the G-DINA model yielding expected correlations that were not statistically significantly different from observed correlations for all but two of the 36 possible coefficients. Taken together, results indicated the expert-driven Q-matrix appropriately represented the relationship between ISP-Skills items and related social-emotional and academic enabling attributes. Furthermore, the selected G-DINA model was a good representation of ISP-Skills performance, with the model reliably reproducing observed item response patterns and correlation estimates.

G-DINA parameter estimates and descriptive statistics suggested the majority of items functioned as intended. Results indicated that the more attributes a student mastered, the higher the probability teachers would rate the item as *Often*/*Almost Always*. More specifically, the likelihood of an *Often/Almost Always* for a given item was greater when the student had mastered the attributes to which that item corresponded. These findings again supported the appropriateness of the specified Q-matrix and the G-DINA model. Some exceptions were noted to item performance. For example, a review of parameter estimates indicated students had a high probability of receiving an *Often/Almost Always* rating even if they had not mastered Attribute A4 (relationship skills). There are several possible explanations for such a finding. For instance, a student may *Often/Almost Always* join classroom activities, not because of his or her superior relationship skills, but because participation is mandatory or well supported within the classroom. Rather than demonstrating the skill of initiating or joining activities, the student may have simply been demonstrating compliance. Alternately, this finding could be due to Q-matrix misspecification: in this case, there may be unmodeled attributes that are involved in initiating or joining activities.

Further review of descriptive statistics suggested mastery prevalence estimates, as defined by DCM, were aligned with developmental expectations founded in theory and prior research. Results suggested that per the ISP-Skills, certain skills were more difficult than others, as indicated by lower mastery rates. This was particularly true of responsible decision making (mastery = 21.3%), study skills (mastery = 42.3%), and motivation (45.8%). Responsible decision making is indeed one of the higher executive functions, often developing in late childhood and adolescence, and involving the evaluation of social norms, adaptive goal setting, and appreciation of moral and ethical standards (Payton et al., 2000; Ross & Tolan, 2018). Similarly, study skills are likely to build over time as students encounter increasingly difficult academic tasks, which they must complete with greater independence through the intentional and skillful application of learning strategies (DiPerna, 2006).

It should be noted that skill classifications of “mastery” or “non-mastery” derived through DCM are model-based and should not be treated as an indicator of the need for supplemental instruction. Rather, the classification accuracy results discussed below afford the most actionable guidelines by which to use ISP-Skills findings to inform decisions regarding mastery status. More specifically, ROC curve findings suggest which DCM-based probability cut scores should potentially be used (pending cross-validation in subsequent research) in determining whether a skill (a) has been acquired, (b) represents a strength for a student, or (c) has not been acquired, thus supporting the need for supplemental instruction.

Research Purpose 2

The second research purpose was to evaluate if the ISP-Skills yielded valid scores as compared to previously validated measures, including the DESSA, SSIS, and the ACES. These criterion measures were chosen for their strong psychometric defensibility, as well as their correspondence to similar constructs targeted through the ISP-Skills via a smaller number of items (DiPerna & Elliott, 1999; Gresham, Elliott, Cook, Vance, & Kettler, 2010). The ISP-Skills showed promising criterion-related validity as compared to the DESSA, SSIS, and the ACES scales. Convergent validity was high for all eight ISP-Skills scales, with coefficients ranging between .70 and .86, exceeding the threshold for a “large” effect. Considering the strong support of the DESSA, SSIS, and ACES as well as their popular usage in schools, these high convergent relations are promising in the context of using the ISP-Skills as an alternative to these measures for more efficient assessment at Tier 2.

In contrast to convergent findings, discriminant validity results were somewhat mixed. Encouragingly, the ISP-Skills social-emotional skills were more strongly related to criterion social-emotional scales than criterion academic enablers scales. Similarly, ISP-Skills academic enablers scales were more strongly related to criterion academic enabler scales than criterion social-emotional scales. These findings therefore speak to the ISP-Skills scales' capacity to differentiate between broader skill domains. In contrast, evidence of ISP-Skills capacity to further differentiate between skills within these domains was lacking. Proposed discriminant relations were higher than hypothesized and exceeded the "large" effect threshold in all cases except one. Furthermore, for certain ISP-Skills items, one or more hypothesized discriminant relations exceeded those of hypothesized convergent relations. This particular finding aligns with the DCM classification accuracy findings, which suggested that the accuracy of the ISP-Skills was optimal when identifying students who had mastered all or none of the attributes.

Discriminant validity findings are somewhat concerning in the psychometric sense, as subscale differentiation is required of a measure intended for skill assessment. With that said, previous research suggests such findings could be expected. Studies have demonstrated the interrelatedness of various social-emotional and academic enabling skills, suggesting that while these skills can be differentiated, an individual's performance of one skill is predictive of their performance of others (Doromal, Cottone, & Kim, 2019). Additional peer-reviewed studies and technical manuals reveal high inter-scale correlations within measures of social-emotional functioning, again suggesting the challenges associated with differentiation (e.g., Gresham & Elliott, 2008). Such findings then suggest that while some level of differentiation among skills is possible, one should not expect the absence of any inter-skill relationships. Regardless, future research should continue to examine this issue, while also considering whether any alterations to the ISP-Skills might enhance its skill differentiation capacity. The importance of this capacity should not be understated, as it might be considered a prerequisite for any measure intended for use in skill assessment to inform instructional interventions.

Research Purpose 3

The final research aim was to evaluate the classification accuracy of ISP-Skills scales. That is, the ability of the ISP-Skills to identify which students have a skill deficit (below average) and are in need of instructional support, as well as identifying areas of strength (i.e., where the student skill is above average). This classification and usage of cut scores can help schools to interpret the results of the ISP-Skills more easily and better match students to skill instruction aligned with their unique needs. Analyses indicated that the ISP-Skills yielded overall acceptable classifications, differentiating students with below average skills from those with normal to above average skills. Self-Awareness, Engagement, and Motivation showed to be moderate predictors of below average classification while the remaining skills (Social Awareness, Self-Management, Relationship Skills, Responsible Decision Making, and Study Skills) were strong predictors of below average skill classification. Furthermore, all eight ISP-Skills scales were moderate predictors of differentiating students with above average skills from students with average to below average skills. Taken together, the cut scores that were generated to differentiate below average students and above average students consistently demonstrated acceptable performance indicating the ISP-Skills accurately classifies students as compared to previously validated measures. The use of cut scores and classification is critical in schools given the spectrum of skill abilities across students. It is essential that schools are able to identify students with a true fluency or acquisition deficit in order to prioritize the implementation of their instructional intervention over students who do not show a skill deficit and fall within the normal range (Stichter et al., 2018). Schools are notoriously low on resources and ensuring the allocation of services is provided to the appropriate students is critical to the success of all students. Skills assessment is intended to assist schools with this and classification within these assessments eases school's ability to interpret the results.

Implications

The results of this study provide preliminary support for the use of the ISP-Skills as an abbreviated alternative to more lengthy skill assessments, as results from this study support the G-DINA model and suggest strong criterion-related validity and accurate risk classification. These results are particularly promising in the context of usage within a multi-tiered system of supports (MTSS). MTSS, which is founded upon a data-based decision-making framework, emphasizes the use of data to make informed decisions regarding intervention decisions for students. Further, MTSS highlights the importance of prevention and targeting students for intervention before their problems are so severe, they necessitate intensive individualized intervention requiring extensive assessments and district resources to ameliorate the problem. Unfortunately, many assessments that allow schools to individualize interventions to student's specific skills deficits lack feasibility at a Tier 2 level due to the time required to complete the lengthy forms.

The ISP-SS was developed to improve upon the time requirements for social skills assessments. The ISP-Skills builds upon the ISP-SS, which was limited to the assessment of social skills, by assessing the broader domain of social-emotional skills, as well as academic enablers. With only 14 items, this tool can be used with a larger number of students than traditionally used tools (e.g. DESSA) in order to tailor Tier 2 interventions. A range of curricula have been developed to date, including those targeting academic enablers, social skills, and the broader domain of social-emotional skills. Within these curricula there are various domains and lessons to target various skills. For example, a social skills curriculum may target both relationship skills in one section, conflict resolution skills in another, and classroom survival skills in another (e.g., Skillstreaming the Elementary School Child; McGinnis, 2012). In our experience, the specific sections applied to skills instruction groups are often determined by the implementer and based upon clinical judgment, rather than being informed by problem analysis data. This is particularly concerning given that many students may be receiving skills instruction that does not address their specific area of need. For example, a student who is referred for anger or conflict may receive emotion regulation instruction when in fact the student would benefit more from skills regarding relationship skills and compromise. Prior research has underscored the importance of matching instructional interventions with specific deficits (Barreras, 2009; Gresham et al., 2006). The conciseness of the ISP-Skills allows for teachers to quickly and efficiently determine likely deficits so that Tier 2 interventions can be tailored to more specific domains. Of course, the limited number of ISP-Skills items restricts its potential for use in item-level analysis, such as what is possible via longer assessments that afford information regarding a wider range of narrow skills (e.g., SSIS and DESSA). Thus, while the ISP-Skills might not suggest which specific narrow skills a student is lacking (e.g., introducing oneself to others), through its DCM-generated subscale scores, it might still afford information regarding the broader skill domain and thus the subset of narrow skills that should be targeted for intervention (e.g., relationship skills).

Limitations and Future Directions

Certain limitations to this investigation should be noted. First, study findings are subject to mono-method and mono-informant biases, as the ISP-Skills and all three criterion measures were completed by the same teacher. It is thus likely that the current findings represent overestimates of true ISP-Skills score reliability, validity, and classification accuracy. It is further likely that the lack of observed discriminant validity at least partially reflects the presence of shared method variance, which inflated the relation among all examined scales. In the interest of gathering even stronger evidence of ISP-Skills construct validity, future research should examine the ISP-Skills relative to alternate informant reports, as well as ecologically-valid outcomes that are commonly of interest to schools and predictive of social and academic success (e.g., office discipline referrals, attendance, suspensions, academic benchmark scores).

Second, all data were collected at a single time point. Thus, while findings are indicative of concurrent relations between the ISP-Skills and criterion measures, no conclusions can be drawn relative to ISP-Skills capacity to predict future outcomes. Once again, in the interest of expanding construct validity evidence, future research should employ a longitudinal design, thereby supporting examination of ISP-Skills predictive validity and classification accuracy, as well as temporal stability. Third, this study examined only a single sample of students. Although a single sample can support calibration of a model-driven scoring protocol, the protocol must be applied to a separate sample to support cross-validation. Accordingly, at this time, support for the ISP-Skills is considered preliminary. Applied use of the measure will only be justified once the DCM-based scoring protocol can be examined within an additional independent sample.

Fourth, though this study was conducted with a broad and normative sample, the ISP-Skills is intended for use with students who are exhibiting SEB risk and thus require Tier 2 intervention. We should note that the use of such samples is common when informing the development, refinement, and initial testing of skill assessment tools (e.g., see technical manuals for the SSIS [Gresham & Elliott, 2008] and DESSA [LeBuffe et al., 2014]). Determination of whether a measure is capable of identifying those exhibiting skill deficits necessitates evaluation of whether the measure can accurately rule out those who are not exhibiting such deficits. Nevertheless, research suggests that when evaluating a measure intended for use with a subsample of individuals (e.g., students exhibiting SEB risk), the use of a broad and general sample may result in misestimation of the measure's psychometric properties (Briesch, Swaminathan, Welsh, & Chafouleas, 2014). Accordingly, future research should recruit samples of only students exhibiting SEB risk to yield a more realistic depiction of ISP-Skills validity and diagnostic accuracy.

As noted within the Introduction to this paper, a final direction for future research pertains to the examination of ISP-Skills treatment utility. To be truly useful and effective as a skill assessment, a measure must yield accurate decisions that contribute to effective intervention. Treatment utility studies support the evaluation of a measure’s capacity in this regard, indicating whether interventions informed by the assessment are more effective than those that are not (Nelson-Gray, 2003). As part of the federally funded research project of which this study is a part, we will be conducting a series of single-case studies and randomized controlled trials that evaluate ISP-Skills treatment utility when used in isolation and as part of a broader Tier 2 process within an MTSS model.

References

Barreras, R. B. (2009). *An experimental analysis of the treatment validity of the social skills deficit model for at-risk adolescents* (Unpublished doctoral dissertation). University of California, Riverside, CA.

Chen, J., de la Torre, J., & Zhang, Z. (2013). Relative and absolute fit evaluation in cognitive diagnosis modeling. *Journal of Educational Measurement*, *50*(2), 123-140.

Christ, T. J., & Boice, C. (2009). Rating scale items: A brief review of nomenclature, components, and formatting to inform the development of Direct Behavior Rating (DBR). *Assessment for Effective Intervention*, *34*, 242-250.

Christ, T. J., Riley-Tillman, T. C., & Chafouleas, S. M. (2009). Foundation for the development and use of Direct Behavior Rating (DBR) to assess and evaluate student behavior. *Assessment for Effective Intervention, 34*, 201–213.

Collaborative for Academic, Social, and Emotional Learning. (2005). *Safe and sound: An educational leader's guide to evidence-based social and emotional learning programs—Illinois edition*. Chicago: Author.

Cook, C. R., Gresham, F. M., Kern, L., Barreras, R. B., Thornton, S., & Crews, S. D. (2008). Social skills training for secondary students with emotional and/or behavioral disorders: A review and analysis of the meta-analytic literature. *Journal of Emotional and Behavioral Disorders*, *16*, 131-144.

de la Torre J. (2011). The generalized DINA model framework. *Psychometrika, 76*, 179-199.

de la Torre J., & Chiu, C-Y. (2016). A general method of empirical Q-matrix validation. *Psychometrika, 81*, 253-273 de la Torre, J., van der Ark, L. A., & Rossi, G. (2018). Analysis of clinical data from a cognitive diagnosis modeling framework. *Measurement and Evaluation in Counseling and Development, 51,* 281-296.

DiPerna, J. C. (2006). Academic enablers and student achievement: Implications for assessment and intervention services in the schools. *Psychology in the Schools, 43,* 7-17.

DiPerna, J., Anthony, C., & Elliott, S. (2019, October 23). *It's about time* [Blog post]. Retrieved from https://measuringsel.casel.org/its-about-time/

DiPerna, J. C., & Elliott, S. N. (1999). The development and validation of the Academic Competence Evaluation Scales. *Journal of Psychoeducational Assessment, 17,* 207–225.

DiPerna, J. C., & Elliott, S. N. (2000). Academic Competence Evaluation Scales. San Antonio, TX: The Psychological Corporation

DiPerna, J. C., Volpe, R. J., & Elliott, S. N. (2001). A model of academic enablers and elementary reading/ language arts achievement. *School Psychology Review, 3,* 298–312.

Doromal, J. B., Cottone, E. A., & Kim, H. (2019). Preliminary validation of the teacher-rated DESSA in a low-income, kindergarten sample. *Journal of Psychoeducational Assessment, 37*, 40-54.

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*, 405-432.

EASEL Lab (2019a). Explore SEL. Retrieved from http://exploresel.gse.harvard.edu/

EASEL Lab (2019b). Taxonomy project. Retrieved from https://easel.gse.harvard.edu/taxonomy-project

Eklund, K., Kilpatrick, K., Kilgus, S. P., Haider, A. (2018). A systematic review of state-level social emotional learning standards: Implications for practice and research. *School Psychology Review, 47,* 316-326.

Elliott, S. N., Gresham, F. M., Frank, J. L., & Beddow, P. A. (2008). Intervention validity of social behavior rating scales: Features of assessments that link results to treatment plans. *Assessment for effective intervention*, *34*, 15-24.

Enders, C. K. (2010). *Applied missing data analysis*. New York, NY: Guilford Press.

Glover, T. A., & Albers, C. A. (2007). Considerations for evaluating universal screening assessments. *Journal of School Psychology, 45,* 117–135.

Gold, M. S., & Bentler, P. M. (2000). Treatments of missing data: A Monte Carlo comparison of RBHDI, iterative stochastic regression imputation, and expectation-maximization. *Structural Equation Modeling*, *7,* 319– 355.

Gresham, F. M., Van, M., & Cook, C. R. (2006). Social skills training for teaching replacement behaviors: Remediation of acquisition deficits for at-risk children. *Behavioral Disorders, 32,* 32-46.

Gresham, F. M., & Elliott, S. N. (2008). *Social Skills Improvement System—Rating Scales manual*. Minneapolis, MN: Pearson Assessments.

Gresham, F. M., Elliott, S. N., Cook, C. R., Vance, M. J., & Kettler, R. (2010). Cross-informant agreement for ratings for social skill and problem behavior ratings: An investigation of the Social Skills Improvement System—Rating Scales. *Psychological Assessment*, *22*, 157-166.

Gresham, F. M., Elliott, S. N., & Kettler, R. J. (2011). Base rates of social skill acquisition/performance deficits, strengths, and problem behaviors: An analysis of the

Social Skills Improvement System—Rating Scales. *Psychological Assessment, 22*, 809-815.

Hawken, L. S, Adolphson, S. L., MacLeod, K. S., & Schumann, J. M. (2009). Secondary tier interventions and supports. In G. Sugai, R. H. Horner, G. Dunlap, & W. Sailor (Eds.). *Handbook of Positive Behavior Support,* New York: Springer.

Henson, R. A., Templin, J. L., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika, 74*, 191-210.

Jaffery, R., Johnson, A. H., Bowler, M. C., Riley-Tillman, T. C., Chafouleas, S. M., & Harrison, S. E. (2015). Using consensus building procedures with expert raters to establish comparison scores of behavior for Direct Behavior Rating. *Assessment for Effective Intervention*, *40*, 195-204.

Jones, S., Bailey, R., Brush, K., & Nelson, B. (2019). *Introduction to the Taxonomy Project: Tools for Selecting & Aligning SEL Frameworks* (Measuring SEL Frameworks Briefs, Comparative Series #1). Retrieved from https://measuringsel.casel.org/wp-content/uploads/2019/02/Frameworks-C.1.pdf

Jones, S. M., McGarrah, M. W., & Kahn, J. (2019). Social and emotional learning: A principled science of human development in context. *Educational Psychologist*, *54*, 129-143.

Kilgus, S. P., Eklund, K., & von der Embse, N. P. (2019). Psychometric defensibility of the Intervention Selection Profile – Social Skills (ISP-SS) with students at risk for behavioral concerns. *Psychology in the Schools, 56,* 526-538.

Kilgus, S. P., von der Embse, N. P., & Eklund, K. (2018). *Intervention Selection Profile –Skills.* (Unpublished measure).

Kilgus, S. P., von der Embse, N. P., Scott, K., & Paxton, S. (2015). Use of the Intervention Selection Profile–Social Skills (ISP-SS) to identify social skill acquisition deficits: A preliminary validation study. *Assessment for Effective Intervention, 40,* 228-239.

LeBuffe, P. A., Shapiro, V. B., & Naglieri, J. A. (2014). *The Devereux Student Strengths Assessment (DESSA): Assessment, technical manual, and user's guide*. Charlotte, NC: Apperson, Inc. (Original work published 2009)

LeBuffe, P. A., Shapiro, V. B., & Robitaille, J. L. (2018). The Devereux Student Strengths Assessment (DESSA) comprehensive system: Screening, assessing, planning, and monitoring. *Journal of Applied Developmental Psychology*, *55*, 62-70. doi:10.1016/j.appdev.2017.05.002

Liu J., Xu, G., & Ying, Z. (2012). Data-driven learning of Q-matrix. *Applied Psychological Measurement, 36*, 548-564.

Maggin, D. M., Zurheide, J., Pickett, K. C., & Baillie, S. J. (2015). A systematic evidence review of the check-in/check-out program for reducing student challenging behaviors. *Journal of Positive Behavior Interventions, 17,* 197-208.

Martens, B. K., Daly, E. J., Begeny, J. C., & VanDerHeyden, A. (2011). Behavioral approaches to education. In W. Fisher, C. Piazza, & H. Roane (Eds.), *Handbook of applied behavior analysis* (pp. 385–401). New York: Guilford.

Martin-Raugh, M., Tannenbaum, R. J., Tocci, C. M., & Reese, C. (2016). Behaviorally anchored rating scales: An application for evaluating teaching practice. *Teaching and Teacher Education*, *59*, 414-419.

Masten, A. S., Roisman, G. I., Long, J. D., Burt, K. B., Obradović, J., Riley, J. R., … Tellegen, A. (2005). Developmental Cascades: Linking Academic Achievement and Externalizing and Internalizing Symptoms Over 20 Years. *Developmental Psychology, 41*(5), 733–746.

Maydeu-Olivares, A. (2013). Goodness-of-fit assessment of item response theory models. *Measurement: Interdisciplinary Research & Perspective*, *11*, 71–101.

McGinnis, E. (2012). Skillstreaming the elementary school child: A guide for teaching prosocial skills (3rd ed.). Champaign, IL: Research Press.

McIntosh, K., Campbell, A. L., Carter, D. R., & Dickey, C. R. (2009). Differential effects of a tier two behavior intervention based on function of problem behavior. *Journal of Positive Behavior Interventions, 11*, 82-93.

National Center for Education Statistics, U.S. Department of Education. (January 2020). *Common core of data*. Retrieved from https://nces.ed.gov/ccd/schoolsearch/index.asp

Newcomer, L. L., & Lewis, T. J. (2004). Functional behavior assessment: An investigation of assessment reliability and effectiveness of function-based interventions. *Journal of Emotional and Behavioral Disorders, 12*, 168-181.

Obradović, J., Burt, K. B., & Masten, A. S. (2009). Testing a Dual Cascade Model Linking Competence and Symptoms Over 20 Years from Childhood to Adulthood. *Journal of Clinical Child & Adolescent Psychology*, *39*, 90–102.

Payton, J. W., Wardlaw, D. M., Graczyk, P. A., Bloodworth, M. R., Tompsett, C. J., & Weissberg, R. P. (2000). Social and emotional learning: A framework for pro- moting mental health and reducing risk behaviors in children and youth. *Journal of School Health*, *70*, 1-8.

Preston, K., Reise, S., Cai, L., & Hays, R. D. (2011). Using the nominal response model to evaluate response category discrimination in the PROMIS emotional distress item pools. *Educational and Psychological Measurement, 71*, 523-550.

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

Rupp, A. A., & Templin, J. L. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement, 6*, 219-262.

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic assessment: Theory, methods, and applications*. New York: Guilford.

Sessoms, J., & Henson, R. A. (2018). Applications of Diagnostic Classification Models: A Literature Review and Critical Commentary. *Measurement: Interdisciplinary Research and Perspectives*, *16*, 1-17.

Smolkowski, K., & Cummings, K. D. (2015). Evaluation of diagnostic systems: The selection of students at risk of academic difficulties. *Assessment for Effective Intervention*, *41*, 41-54.

Stichter, J. P., Malugen, E. C., & Davenport, M. A. (2019). A Six-Step Decision-Making Process to Guide Social Skills Instruction. *Intervention in School and Clinic*, *54*, 149-159.

Streiner, D. L., & Cairney, J. (2007). What's under the ROC? An introduction to receiver operating characteristic curves. *The Canadian Journal of Psychiatry/ La Revue canadienne de psychiatrie, 52,* 121–128.

Sugai, G., & Horner, R. (2009). Response-to-intervention and school-wide positive behavior supports: Integration of multitiered systems approaches. *Exceptionality, 17,* 223-237.

Tatsuoka, K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20,* 345-354.

Wang, W., Song, L., Chen, P., Meng, Y., & Ding, S. (2015). Attribute-level and pattern-level classification consistency and accuracy indices for cognitive diagnostic assessment. *Journal of Educational Measurement, 52*, 457-476.

Zins, J. E., & Elias, M. J. (2007). Social and emotional learning: Promoting the development of all students. *Journal of Educational and Psychological Consultation*, *17*, 233-255.

Table 1

Student and teacher demographic statistics

Variable	Student	Teacher
Grade		
K	114 (13%)	
1	135 (15%)	
2	119 (14%)	
3	199 (23%)	
4	151 (17%)	
5	120 (14%)	
6	41 (5%)	
Gender		
Male	469 (53%)	14 (7%)
Female	402 (46%)	180 (92%)
Prefer not to say	6 (1%)	2 (1)
Race/Ethnicity		
White	396 (45%)	160 (82%)
Black	272 (31%)	16 (8%)
Hispanic/Latinx	148 (17%)	13 (7%)
Other	37 (4%)	4 (2%)
Multiracial	20 (2%)	1 (<1%)
Asian	4 (<1%)	1 (<1%)
American Indian/Alaska Native	1 (<1%)	0 (0%)
Missing	1 (<1%)	1 (<1%)
Years of Experience		
0-5 years		60 (31%)
6-10 years		43 (22%)
11-15 years		36 (18%)
16-20 years		33 (17%)
21+ years		23 (12%)
Teacher Degree		
Bachelor's		104 (53%)
Master's		87 (44%)
Professional or Doctorate		5 (3%)

Table 2

Percentage of students within each classification level across criterion measures

	Below Average	Average	Above Average
DESSA Self-Awareness	228 (26%)	460 (52%)	191 (22%)
DESSA Social Awareness	244 (28%)	390 (44%)	245 (28%)
DESSA Self-Management	228 (26%)	440 (50%)	211 (24%)
DESSA Responsible DM	225 (26%)	445 (51%)	209 (24%)
SSIS Total	278 (32%)	378 (43%)	223 (25%)
ACES Engagement	219 (25%)	559 (64%)	101 (11%)
ACES Motivation	252 (29%)	498 (57%)	129 (15%)
ACES Study Skills	348 (40%)	449 (51%)	82 (9%)

Table 3

Original and estimated Q-matrices for the social-emotional skills items

		Attribute				
		A1	**A2**	**A3**	**A4**	**A5**
Item		Self-Awareness	Social Awareness	Self-Management	Relationship Skills	Responsible Decision-making
1	Perceives, understands, and appreciates his/her own skills, interests, attitudes, thoughts, and emotions.	1	0	0	0	0
2	Perceives, understands, and appreciates others' emotions.	0	1	0	0	0
3	Monitors own emotions and controls his/her behavior.	1	0	1	0	0
4	Identifies problems and chooses socially acceptable solutions.	0	1	0	0	1
5	Speaks to others in a polite, courteous, and respectful manner.	0	1*	0	1	1
6	Helps others, shares possessions, and complies with rules.	0	0	1	1	1*
7	Treats objects with care; takes ownership for personal roles and actions.	0	0	1	0	1
8	Initiates or joins activities with peers.	0	0	0	1	0
9	Responds to others in an appropriate and safe manner within conflict and non-conflict situations.	0	0	1	1	1

Note. * indicates Q-matrix modifications suggested by the validation algorithm.
G-DINA model fit using the original Q-matrix: AIC = 6015.48; BIC = 6364.33.
G-DINA model fit using the estimated Q-matrix: AIC = 6008.07; BIC = 6318.69.

Table 4

Q-matrix for the Academic Enabler Items

		Attribute		
		A6	**A7**	**A8**
Item		Study Skills	Engagement	Motivation
10	Adequately prepares for quizzes, tests, and assignments.	1	0	0
11	Takes good notes; Effectively organizes materials and assignments.	1	0	0
12	Actively or passively participates in classroom instruction and activities.	0	1	0
13	Can complete assignments independently; Can work alone for an extended period of time.	0	0	1
14	Interested in and excited for academics; Produces quality work.	0	0	1

Note. The estimated Q-matrix was identical to the original Q-matrix, so no modifications were suggested.

Table 5

Parameter Estimates Obtained from Fitting the G-DINA Model to Social-Emotional Item Data

		G-DINA coefficients							
Item	Attribute(s)	δ_0	δ_1	δ_2	δ_3	δ_{12}	δ_{13}	δ_{23}	δ_{123}
1	A1	.08	.90						
2	A2	.05	.93						
8	A4	.38	.57						
3	A1, A3	.01	.24	.58	—	.12			
4	A1, A3	.00	.48	.25	—	.21			
7	A3, A5	.11	.86	.82	—	-.82			
5	A2, A4, A5	.08	.12	.45	.80	.31	-.03	-.45	-.31
6	A3, A4, A5	.02	.66	.61	.66	-.29	-.66	-.58	.58
9	A3, A4, A5	.01	.00	.37	.00	.45	.90	.09	-.83

Note. N = 879. Attribute A1 = self-awareness; A2 = social awareness; A3 = self-management; A4 = relationship skills; A5 = responsible decision-making. Standardized root mean square of the residuals = .031.

Table 6

Parameter Estimates Obtained from Fitting the G-DINA Model to Academic Enabler Item Data

		G-DINA coefficients		*P(*x = *Often/Almost Always*) if attribute is mastered
Item	Attribute	δ_0	δ_1	
10	A6	.03	.89	.94
11	A6	.05	.84	.89
12	A7	.07	.90	.97
13	A8	.09	.85	.94
14	A8	.06	.85	.91

Note. *N* = 879. Attribute A6 = study skills; A7 = academic engagement; A8 = motivation. Standardized root mean square of the residuals = .032.

Table 7

Attribute Mastery Prevalence

Attribute		**Non-mastery (%)**	**Mastery (%)**
Social-emotional skills			
A1	Self-Awareness	48.6	51.4
A2	Social Awareness	46.3	53.7
A3	Self-Management	45.3	54.7
A4	Relationship Skills	41.6	58.4
A5	Responsible Decision-making	78.7	21.3
Academic enablers			
A6	Study Skills	57.7	42.3
A7	Academic Engagement	43.7	56.3
A8	Motivation	54.2	45.8

Table 8

Often/Almost Always *Rating Probabilities on the Social-Emotional Items, Conditional on Mastery Pattern (Latent Class)*

Mastery Pattern					Item									Freq.	Classification Accuracy
A1	**A2**	**A3**	**A4**	**A5**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**		
0	0	0	0	0	.08	.05	.01	.00	.08	.02	.11	.38	.01	269	.97
1	0	0	0	0	.98	.05	.25	.00	.08	.02	.11	.38	.01	0	.00
0	1	0	0	0	.08	.98	.01	.48	.20	.02	.11	.38	.01	0	.00
0	0	1	0	0	.08	.05	.59	.00	.08	.68	.97	.38	.01	16	.73
0	0	0	1	0	.08	.05	.01	.00	.53	.62	.11	.94	.37	19	.54
0	0	0	0	1	.08	.05	.01	.26	.88	.68	.92	.38	.01	0	.00
1	1	0	0	0	.98	.98	.25	.48	.20	.02	.11	.38	.01	9	.85
1	0	1	0	0	.98	.05	.95	.00	.08	.68	.97	.38	.01	3	.86
1	0	0	1	0	.98	.05	.25	.00	.53	.62	.11	.94	.37	5	.51
1	0	0	0	1	.98	.05	.25	.26	.88	.68	.92	.38	.01	0	.00
0	1	1	0	0	.08	.98	.59	.48	.20	.68	.97	.38	.01	0	.00
0	1	0	1	0	.08	.98	.01	.48	.95	.62	.11	.94	.37	4	.81
0	1	0	0	1	.08	.98	.01	.95	.97	.68	.92	.38	.01	0	.00
0	0	1	1	0	.08	.05	.59	.00	.53	1.00	.97	.94	.82	0	.00
0	0	1	0	1	.08	.05	.59	.26	.88	.68	.97	.38	.91	26	.63
0	0	0	1	1	.08	.05	.01	.26	.88	.71	.92	.94	.46	30	.69
1	1	1	0	0	.98	.98	.95	.48	.20	.68	.97	.38	.01	12	.80
1	1	0	1	0	.98	.98	.25	.48	.95	.62	.11	.94	.37	4	.56
1	1	0	0	1	.98	.98	.25	.95	.97	.68	.92	.38	.01	0	.00
1	0	1	1	0	.98	.05	.95	.00	.53	1.00	.97	.94	.82	1	.63
1	0	1	0	1	.98	.05	.95	.26	.88	.68	.97	.38	.91	0	.00
1	0	0	1	1	.98	.05	.25	.26	.88	.71	.92	.94	.46	22	.69
0	1	1	1	0	.08	.98	.59	.48	.95	1.00	.97	.94	.82	24	.64
0	1	1	0	1	.08	.98	.59	.95	.97	.68	.97	.38	.91	5	.56
0	1	0	1	1	.08	.98	.01	.95	.97	.71	.92	.94	.46	0	.00
0	0	1	1	1	.08	.05	.59	.26	.88	1.00	.97	.94	.99	14	.56
1	1	1	1	0	.98	.98	.95	.48	.95	1.00	.97	.94	.82	0	.00
1	1	1	0	1	.98	.98	.95	.95	.97	.68	.97	.38	.91	7	.37
1	1	0	1	1	.98	.98	.25	.95	.97	.71	.92	.94	.46	16	.54
1	0	1	1	1	.98	.05	.95	.26	.88	1.00	.97	.94	.99	0	.00
0	1	1	1	1	.08	.98	.59	.95	.97	1.00	.97	.94	.99	0	.00
1	1	1	1	1	.98	.98	.95	.95	.97	1.00	.97	.94	.99	393	.99

Note. *N* = 879. Frequencies based on MAP estimation. Attribute and attribute-level classification accuracy: A1 = self-awareness (.97); A2 = social awareness (.98); A3 = self-management (.96); A4 = relationship skills (.94); A5 = responsible decision-making (.96). Test-level classification accuracy = .885.

Table 9

Often/Almost Always *Rating Probabilities on the Academic Enabler Items, Conditional on Mastery Profile (Latent Class)*

Mastery Pattern			**Item**					**Freq.**	**Classification Accuracy**
A6	**A7**	**A8**	**10**	**11**	**12**	**13**	**14**		
0	0	0	.03	.05	.07	.09	.06	354	.93
1	0	0	.92	.89	.07	.09	.06	2	.70
0	1	0	.03	.05	.97	.09	.06	127	.94
0	0	1	.03	.05	.07	.93	.90	0	.00
1	1	0	.92	.89	.97	.09	.06	0	.00
1	0	1	.92	.89	.07	.93	.90	0	.00
0	1	1	.03	.05	.97	.93	.90	25	.63
1	1	1	.92	.89	.97	.93	.90	371	.99

Note. *N* = 879. Frequencies based on MAP estimation. Attribute and attribute-level classification accuracy: A6 = study skills (.99); A7 = academic engagement (.96); A8 = motivation (.98). Test-level classification accuracy = .935.

Table 10

Item Fit Statistics

		Proportion					
	Item	Expected	Observed	Difference	*SE*	*Z*	*p*
Social-emotional skills	1	.550	.548	.001	.017	.077	.938
	2	.541	.540	.001	.017	.046	.963
	3	.520	.521	.001	.017	.065	.948
	4	.505	.503	.003	.017	.147	.883
	5	.642	.639	.002	.016	.144	.885
	6	.624	.623	.001	.016	.048	.962
	7	.659	.659	.000	.016	.008	.994
	8	.714	.712	.002	.015	.110	.912
	9	.576	.575	.001	.017	.063	.950
Academic enablers	10	.408	.407	.001	.017	.028	.978
	11	.400	.403	.003	.017	.178	.859
	12	.579	.580	.002	.017	.104	.917
	13	.474	.474	.001	.017	.039	.969
	14	.444	.446	.002	.017	.096	.924

Note. *N* = 879.

Table 11

Criterion-related Validity Coefficients, Comparing ISP-Skills Probability Scores and Criterion Scale Scores

	ISP-Skills Scales							
Criterion	Self-Awareness	Social Awareness	Self-Management	Relationship Skills	Responsible DM	Engagement	Motivation	Study Skills
DESSA Self-Awareness	**.70**	.70	.59	.68	.67	.72	.72	.71
DESSA Social Awareness	.81	**.83**	.78	.79	.81	.66	.66	.65
DESSA Self-Management	.80	.82	**.78**	.78	.81	.71	.72	.71
SSIS Total	.83	.85	.78	**.82**	.81	.69	.69	.69
DESSA Responsible DM	.80	.83	.78	.78	**.80**	.69	.70	.69
ACES Engagement	.61	.60	.46	.60	.55	**.74**	.73	.73
ACES Motivation	.70	.70	.60	.67	.67	.85	**.86**	.85
ACES Study Skills	.69	.69	.61	.66	.67	.83	.84	**.84**

Note. Bolded values correspond to hypothesized convergent relations, while non-bolded values correspond to hypothesized discriminant relations.

Table 12

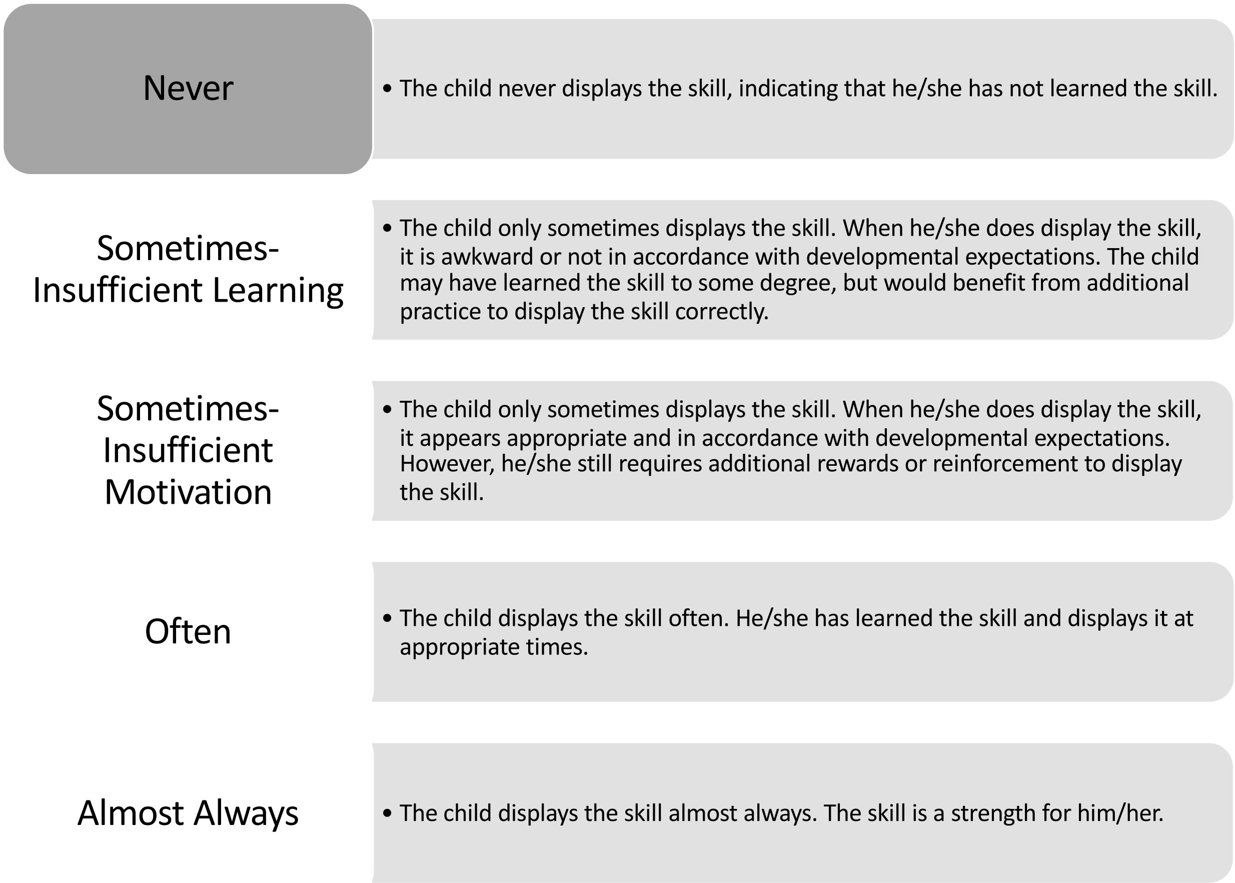
Diagnostic Accuracy of ISP-Skills Scales in Predicting Below Average Scores on Criterion Measures

	AUC	Cut Score	SE	SP	PPV	NPV	CC
Self-Awareness	.84 (.80-.87)	.06	.86	.70	.50	.93	.74
Social Awareness	.90 (.88-.92)	.01	.90	.81	.65	.95	.84
Self-Management	.90 (.88-.92)	.06	.89	.81	.62	.96	.83
Relationship Skills	.92 (.91-.94)	.41	.87	.86	.74	.93	.86
Responsible DM	.90 (.88-.92)	.04	.85	.83	.64	.94	.84
Study Skills	.91 (.89-.93)	.02	.91	.81	.76	.93	.85
Engagement	.85 (.82-.87)	.01	.82	.79	.56	.93	.80
Motivation	.89 (.87-.91)	.01	.94	.75	.61	.97	.81

Table 13

Diagnostic Accuracy of ISP-Skills Scales in Predicting Above Average Scores on Criterion Measures

	AUC	Cut	SE	SP	PPV	NPV	CC
Self-Awareness	.86 (.84-.88)	.99	.90	.78	.53	.97	.80
Social Awareness	.88 (.86-.90)	.99	.93	.74	.58	.97	.80
Self-Management	.84 (.82-.87)	.99	.94	.76	.55	.98	.80
Relationship Skills	.87 (.85-.89)	.99	.90	.81	.62	.96	.83
Responsible DM	.86 (.84-.88)	.99	.91	.78	.56	.96	.81
Study Skills	.87 (.86-.89)	.99	.96	.76	.29	1.00	.78
Engagement	.86 (.84-.88)	.99	.90	.77	.34	.98	.79
Motivation	.89 (.88-.91)	.99	.98	.79	.44	.99	.81



Never

- The child never displays the skill, indicating that he/she has not learned the skill.

Sometimes-Insufficient Learning

- The child only sometimes displays the skill. When he/she does display the skill, it is awkward or not in accordance with developmental expectations. The child may have learned the skill to some degree, but would benefit from additional practice to display the skill correctly.

Sometimes-Insufficient Motivation

- The child only sometimes displays the skill. When he/she does display the skill, it appears appropriate and in accordance with developmental expectations. However, he/she still requires additional rewards or reinforcement to display the skill.

Often

- The child displays the skill often. He/she has learned the skill and displays it at appropriate times.

Almost Always

- The child displays the skill almost always. The skill is a strength for him/her.

Figure 1. ISP-Skills behaviorally anchored rating scale (BARS)

Self-Awareness

- The ability to accurately recognize one’s own emotions, thoughts, and values and how they influence behavior. The ability to accurately assess one’s strengths and limitations, with a well-grounded sense of confidence, optimism, and a “growth mindset.”

Self-Management

- The ability to successfully regulate one’s emotions, thoughts, and behaviors in different situations — effectively managing stress, controlling impulses, and motivating oneself. The ability to set and work toward personal and academic goals.

Social Awareness

- The ability to take the perspective of and empathize with others, including those from diverse backgrounds and cultures. The ability to understand social and ethical norms for behavior and to recognize family, school, and community resources.

Relationship Skills

- The ability to establish and maintain healthy and rewarding relationships with diverse individuals and groups. The ability to communicate clearly, listen well, cooperate with others, negotiate conflict constructively, and seek and offer help when needed.

Responsible Decision Making

- The ability to make constructive choices about personal behavior and social interactions based on ethical standards, safety concerns, and social norms. The realistic evaluation of consequences of various actions, and a consideration of the well-being of oneself and others.

Study Skills

- Behaviors that facilitate the processing of new materials and enhance performance on academic tasks (e.g., preparing for tests, applying appropriate learning strategies, taking good notes, organization of materials).

Academic Engagement

- Passive or active engagement in academic activities (e.g., writing, raising hand, answering a question, talking about a lesson, listening to the teacher, reading silently, or looking at instructional materials).

Motivation

- Approach, persistence, and interest in academic subjects and activities (e.g., excitement for academics, ability to keep working when challenging, willingness to take on new tasks, capacity to generate quality work).

Figure 2. Conceptual definitions of each ISP-Skills attribute.

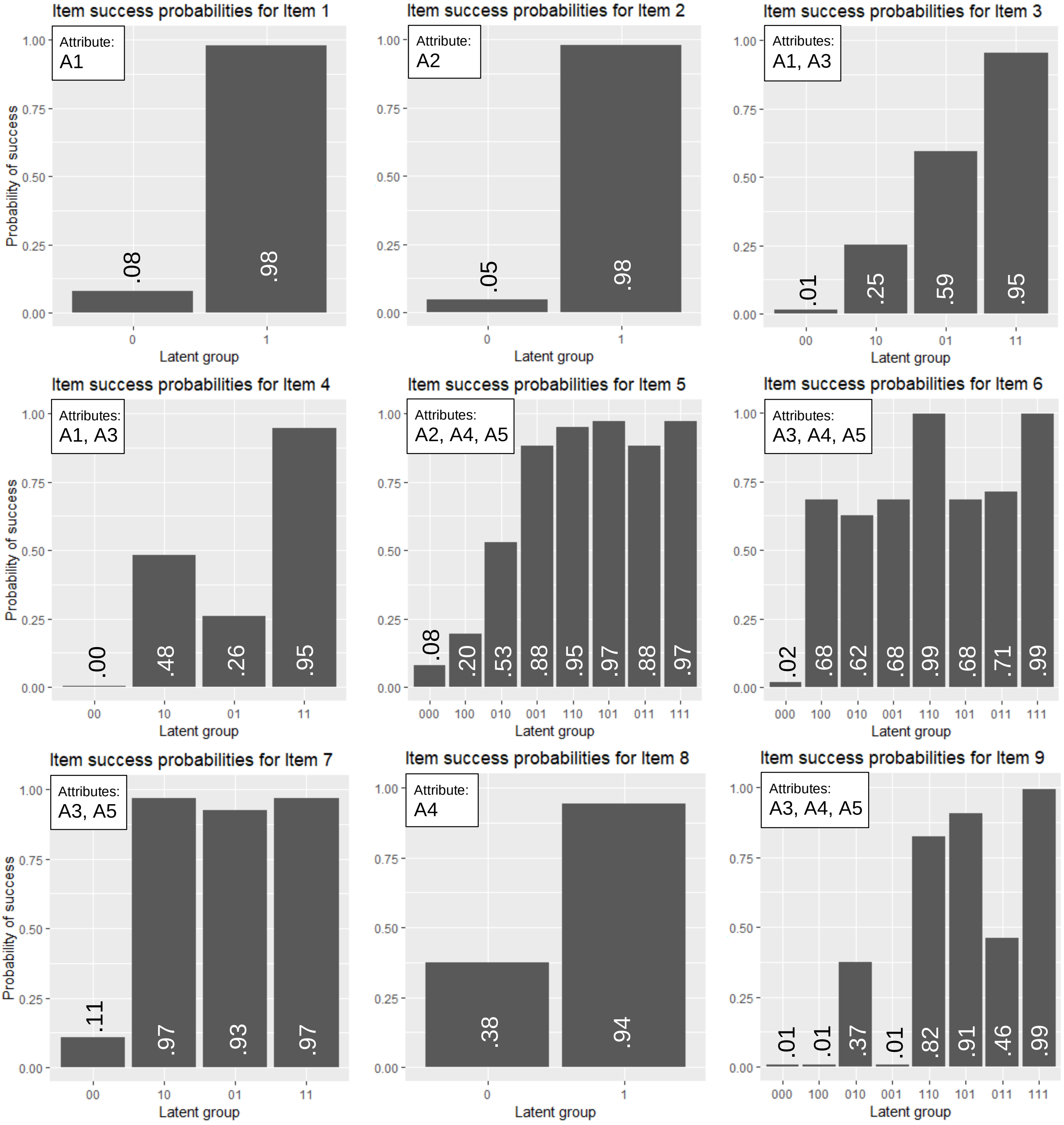


Figure 3. *Often/Almost Always* rating probabilities on the 9 social-emotional items, conditional on latent mastery classes. Attribute A1 = self-awareness; A2 = social awareness; A3 = self-management; A4 = relationship skills; A5 = responsible decision-making. 0 = attribute non-mastery; 1 = attribute mastery.