Title

Predictors of Rater Bias in the Assessment of Social-Emotional Competence

Journal

INTERNATIONAL JOURNAL OF EMOTIONAL EDUCATION, 8(2)

ISSN

2073-7629

Authors

Shapiro, Valerie B
Kim, BK Elizabeth
Accomazzo, Sarah
et al.

Publication Date

2016

Published:

Shapiro, V.B., Kim, B.K.E., Accomazzo, S. & Roscoe, J.N. (2016). Predictors of rater bias in the assessment of Social Emotional Competence. *International Journal of Emotional Education*, 8(2), 25-44.

Abstract

The *Devereux Student Strengths Assessment Mini* (*DESSA-Mini*; LeBuffe, Shapiro, & Naglieri, 2011/2014) efficiently monitors the growth of Social Emotional Competence (SEC) in the routine implementation of Social Emotional Learning programs. The *DESSA-Mini* is used to assess approximately .5 million children around the world. Since behavior rating scales can have "rater bias," this paper examines rater characteristics that contribute to *DESSA-Mini* ratings.

Rater characteristics and *DESSA-Mini* ratings were collected from elementary school classroom teachers (*n*=72) implementing TOOLBOX in a racially/ethnically diverse California school district. Teachers rated 1,676 students, who scored similarly to a national reference group.

Multilevel modeling was used. Only 16% of variance in DESSA-mini ratings was attributable to raters. Relationships between teacher characteristics and ratings were estimated to examine rater variance. Collectively, four characteristics of teachers (perceived barriers to student learning, sense of their "typical" student's level of Social Emotional Competence, anticipation of SEL program implementation challenges, and intentions to fully implement a newly adopted SEL program) accounted for bias in teacher-generated DESSA scores, leaving only 10% of the variance unexplained. Identified sources of "rater bias" can be controlled for in research and addressed through thoughtful program selection, training, and implementation.

Keywords: Rater bias, social emotional competence, social emotional learning, DESSA, TOOLBOX

Predictors of Rater Bias in the Assessment of Social Emotional Competence

Nearly 20% of youth in the United States have a mental, emotional, or behavioral disorder (Kessler et. al, 2012; O'Connell, Boat, & Warner, 2009). The presence of a mental, emotional, or behavioral problem makes it less likely that a young person will reach important developmental and social milestones of adolescence, which in turn increases the likelihood of problems in adulthood (Copeland, Wolke, Shanahan, & Costello, 2015). Mental, emotional, and behavioral disorders and their consequences to society costs the United States roughly $247 billion annually (O'Connell, et al., 2009). This cost does not include the personal hardship experienced by each individual child and family challenged to navigate a complex social environment without the tools to do so successfully.

Longitudinal research has identified reliable predictors of youth mental, emotional, and behavioral problems (Coie et al., 1993; Catalano et al., 2012). These predictors serve as clues as to what characteristics and experiences disrupt typical youth development and what skills and supports children need to succeed. To promote positive youth development, communities act intentionally ("intervene") in hopes of reducing children's experiences of adversity (reducing "risk factors") while augmenting children's strengths (increasing "protective factors"). Findings from resilience research have revealed that most children have both intrinsic and learned capacities to overcome the adversities they face (Masten, 2014). Social Emotional Learning (SEL) interventions in schools are intended to uncover, recognize, and nurture these endemic capacities in children, disrupting trajectories toward problem occurrence, and strengthening their prospects for school and life success. An emerging science demonstrates that SEL programs can impact a broad array of important child outcomes, such as preventing aggression, anxiety, bullying, conduct problems, delinquency, drug use, and truancy, while promoting emotional regulation, prosocial skills, and academic achievement (Abbott, et al., 1998; Domitrovich, Cortes, & Greenberg, 2007; Durlak, Weissberg, Dymnicki, Taylor, & Schellinger, 2011; Espelage, Rose, & Polanin, 2015; Flay & Allred, 2003; Greenberg et al., 2003).

In order to progress our knowledge about whether specific SEL programs work, for whom, and under what conditions, we need psychometrically sound assessment tools that allow us to observe the

impact of the intervention on the growth of protective factors (Naglieri, LeBuffe, & Shapiro, 2013). Such tools, if practical enough for routine use, can also facilitate the high-quality implementation of SEL programs in multiple ways. For example, initial assessment can help teachers and student service personnel identify students with the greatest strengths and needs to target with interventions (Naglieri, LeBuffe, & Shapiro, 2011). Also, repeated assessment can determine whether the SEL intervention is having its intended impact on students in real-time, or if changes to the nature, intensity, or implementation quality of the intervention need to be made (Simmons, Shapiro, Accomazzzo, & Manthey, 2016). This type of monitoring is particularly useful in regions where SEL programs tend to be either untested or imported from other contexts and adapted for local populations and service settings (Perez-Gomez, Mejia-Trujillo, Brown, & Eisenberg, 2016).

The *Devereux Student Strengths Assessment* (*DESSA*) *Mini* (Naglieri, LeBuffe, & Shapiro, 2011/2014) was designed to overcome obstacles to screening and monitoring the growth of Social Emotional Competence in the routine implementation of SEL programs (Maras, 2015). With only 8 items, the *DESSA-Mini* is a behavior rating scale that can be completed by teachers and out-of-school time program staff in just one minute (Shapiro, Kim, Fleming, & LeBuffe, 2016). This strength-based assessment system, which includes four interchangeable brief forms and a longer full assessment (LeBuffe, Shapiro, & Naglieri, 2009/2014), is now being used to assess approximately a half million children each year in the United States, and in countries such as Australia, Canada, Mexico, Qatar, South Africa, and the United Kingdom. The English and Spanish language instruments, normed on a representative sample of youth aged 5 to 14 in the United States, have also been translated (e.g. Italian Edition; LeBuffe, Shapiro, & Naglieri, 2009/2015), normed, and culturally adapted (e.g. Dutch Adaptation; LeBuffe, Shapiro, Naglieri, Pont, & Punt, 2013) for use in other countries. The instruments are being used by researchers in various regions of the globe to determine the effectiveness of SEL interventions; examples include the *Random Acts of Kindness* Curriculum (PI: Kimberly Schonert-Reichl) in Canada and the *Cool to Be Me* Programme (PI: Linda Bruce) in South Africa (SEL Consulting, 2015).

The *DESSA-Mini* uses a format that is common to many behavior rating scales, measuring the frequency of a student's behavior relative to a standardized reference group. The *DESSA-Mini* is completed by indicating, for each item, how often in the past four weeks the student performed a specific positive behavior (*Never* = 0, *Rarely* = 1, *Occasionally* = 2, *Frequently* = 3, *Very Frequently* = 4). Items are summed to a Raw Score Total which is converted to a *T*-score ($M = 50$, $SD = 10$), referred to as the Social Emotional Total.

Behavior rating scales like the *DESSA-Mini* have many perceived benefits (Shapiro, Accomazzo, Claassen, & Fleming, 2015). They can be used to efficiently collect information about behavior performance across settings, from multiple informants, and over multiple time points. They tend to have broad coverage and are somewhat more practical to administer, score, and interpret compared to other data collection options (e.g., direct observation) (McKown, 2015). These are important advantages for supporting primary prevention programs (LeBuffe & Shapiro, 2004). Yet, behavioral rating scales have also been criticized for their potential to incorporate rater bias into assessment scores because each item requires interpretation, reflection, and judgment by the rater (Elliot, Frey, & Davies, 2015; Hoyt & Kerns, 1999). In other words, behavior rating scale scores are likely to reflect characteristics of the rater as well as the student being rated (Hoyt, 2000).

Rater bias is a form of non-random measurement error, or systematic variance, that is attributable to the rater (Hoyt & Kerns, 1999). Rater bias may artificially inflate or suppress assessment scores relative to the actual frequency of behavior. A large amount of rater bias is problematic in practice settings because scores could be less precise than are desired for clinical and educational decision making. A large amount of rater bias is also problematic in research because it reduces the capacity to fully estimate (and ultimately detect) relationships between variables. Although systematic variance is difficult to observe in routine practice, it can be corrected once it has been identified (Mason, Gunersel, & Ney, 2008). Sources of rater bias are important to uncover to ensure scores that inform decision-making in both the research and practice realms are reliable, valid, and equitable.

Rater bias comes in two forms: dyad-specific variance and rater-specific variance (Hoyt & Kerns, 1999). Dyad-specific variance is inherently about the interaction between the rater and the student being rated, which occurs when a rater has a different reaction to particular students. Specifically, a characteristic of the student (e.g., disability), unrelated to the construct being measured, influences the way the adult rates the student. Studies have experimentally-induced rater bias by varying the gender or diagnostic label assigned to children in the assessment processes (e.g., Foster & Ysseldyke, 1976; Kelter & Pope, 2011), but the randomized design was used to eliminate individual rater differences rather than examine them.

Rater-specific variance reflects rater differences that are consistent across targets (Hoyt, 2000). Put simply, different raters can react to the same questions differently, regardless of the student they are rating. The average rating from one teacher may deviate from the average rating across all teachers in ways that are predictable, reflecting how the rater generally perceives students in the domain being assessed (e.g., Social Emotional Competence) or reacts to the assessment prompt, items, or response choices. If a teacher's average rating trends positive, the rater is said to be lenient (Ford, 1931). If a teacher's average rating trends negative, the rater is said to be severe. Understanding the nature and source of leniency and severity errors could inform score interpretation in research and practice.

Rater variance in the assessment of Social Emotional Competence is difficult to explore in routine practice where there is usually only one rater per student, and the obtained score is treated as the "true" score. Assessment developers often conduct small inter-rater reliability studies to broadly understand the extent to which a pair of raters agree in their assessment of the same child (Gresham, Cook, Vance, Elliott, & Kettler, 2010). Inter-rater reliability studies of the *DESSA-Mini*, for example, have shown correlations that range from .70-.81 across the 4 forms, and scores that differ, on average, by 0-.60 *T*-scores points (Naglieri, et al., 2011/2014). Studies like these provide evidence that behavior rating scale scores do reflect characteristics of the rater, to some extent, in addition to the student being rated. On the other hand, these studies do not reveal the source of the bias, or clarify how one might address it when interpreting or using the scores.

Given that there is no consensus indicator for "true" levels of student Social Emotional Competence, validity studies that attempt to discover which teacher has the more "correct" perception are not tenable at this time. Alternatively, we can use statistical techniques to determine the extent to which a given teacher's ratings of his or her students, on average, deviate from all other teachers' ratings of their students. Teacher characteristics that predict these deviations can be considered sources of rater bias.

Although questions about how teacher perceptions impact ratings arise frequently in practice settings, a recent review (Schultz & Evans, 2012) found only 37 articles on the topic. Mason and colleagues (2014) note that most of the articles written about rater bias are conceptual and "do not offer quantifiable evidence of mean differences directly attributable to teacher characteristics of beliefs" (p.1019). Additionally, they argue that, given the large number of teacher variables that may influence behavior ratings, initial inquiries need to look through one lens at a time. The current paper examines potential sources of rater-specific biases in rating student Social Emotional Competence through the lens of implementation science (the systematic study of implementation).

Implementation is a term used to describe the activities designed to put an intervention into practice. A central lesson from implementation science is the importance of the program implementers to the ultimate success of an intervention (Elias, Zins, Graczyk, & Weissberg, 2003). Examining "rater bias" through the lens of implementation science encourages us to understand the contributions of the rater as an essential part of the assessment, intervention, and evaluation process rather than overlooking or isolating them as "noise" in measurement.

There is a burgeoning literature on contextual variables that impact the implementation of SEL programs in schools (Elias, 2007; Fagan, Hawkins, & Shapiro, 2015; Greenberg, Domitrovich, Graczyk, & Zins, 2005). Durlak & Depre's (2008) systematic review of this literature identified teacher characteristics consistently associated with implementation success: perceptions that the intervention is needed, expectations that the intervention will be beneficial, and having the requisite skills and confidence to do what is expected. Additional organizational factors identified with implementation success included a positive work climate and staff norms regarding change. It may be that the same

characteristics that predict implementation success also predict the way in which teachers rate student behavior.

This paper seeks to determine whether teacher characteristics that impact the successful implementation of SEL programs are similar to those that explain rater bias in the assessment of student Social Emotional Competence. Specifically, this study examines the extent to which *DESSA-Mini* ratings are affected by teacher attitudes, capacities, and expectations, perceptions of implementation, impact, and school climate, and finally, what they generally perceive to be the levels of Social Emotional Competence within themselves and others. Each of these teacher characteristics was hypothesized to reveal a potential leniency or severity error in the completion of the *DESSA-Mini*.

Methods

Study and Data Description

The TOOLBOX Implementation Research Project is a quasi-experimental study of TOOLBOX (Collins, 2015), a commonly used Social Emotional Learning (SEL) program aimed at enhancing Social Emotional Competence among students in Kindergarten through 6th grade. TOOLBOX provides a common language to guide school and family support for children's social and emotional development through the instruction and reinforcement of 12 tools (e.g., the Breathing Tool, the Garbage Can Tool). Developed to be an inherently practical SEL intervention, TOOLBOX strives to augment approaches that are natural to teachers and caregivers to reveal tools endemic to children. Specifically, TOOLBOX seeks to foster self-awareness, social-awareness, self-management, decision-making, and relationship skills in children through explicit lesson plans, classroom and school-wide strategies, and integration/reinforcement at home. TOOLBOX has been widely implemented in Northern California school districts and has been explored in two studies. The West Costa County Unified School District Evaluation (Dovetail Learning, 2013) found that teachers reported using and valuing TOOLBOX on a post-intervention survey. The Sonoma County Collaboration for Resilient Children pre/post evaluation found that, after just four months of using TOOLBOX, teachers and yard aids perceived significantly

higher emotional and behavioral strengths in children, including various interpersonal, intrapersonal, and affective strengths (DeLong-Cotty, 2011).

The current study features the implementation of TOOLBOX within one Northern California School District. This district served 10,982 students during the 2014-2015 academic year (District, 2016). Six elementary schools were each assigned to one of the following conditions: (1) the TOOLBOX "standard" implementation - a high-dosage condition which included full TOOLBOX lesson plans, a compendium of TOOLBOX strategies and practices, and a full complement of material resources, (2) the TOOLBOX "primer" implementation - a low-dosage condition which included only the most essential TOOLBOX strategies and practices, and only brief introductory lessons to the TOOLBOX tools, without the benefit of full lesson plans or material resources, and (3) a measurement-only comparison condition.

The four elementary schools assigned to implement the Standard or Primer versions of TOOLBOX serve a racially and ethnically diverse student body (53% Hispanic/Latino(a), 16% Asian/Asian American, 13% Black/African American, 8% White/European American, 7% Filipino, and 3% Other) with 42% of students primarily speaking a language other than English (e.g., Spanish, Cantonese, Mandarin, Tagalog, Vietnamese, Arabic) in their homes (District, 2016). Close to 70% of students had a household income of less than $44,123 annually for a family of 4. In 2015, students meeting or exceeding the state educational standards in Language Arts/Literacy was 27% and in Mathematics was 24%.

Five days prior to the start of instruction for the 2015-2016 school year, teachers and staff from the four elementary schools using the Standard or Primer versions of TOOLBOX received a six-hour training to prepare them to implement TOOLBOX. Data were collected before and after the training to learn about the teachers and their teaching environment, collect their feedback on the training, and understand their expectations for program implementation. Of the 101 classroom teachers in schools implementing TOOLBOX, 95 classroom teachers attended the training (94%). Of the classroom teachers in attendance, 72 (76%) completed a pre-training survey and 71 (75%) completed a post-training survey. A total of 73 classroom teachers responded to at least one of these surveys, with 72 of the 73 (99%)

survey participants consenting for their responses to be used in research. Thus, the analysis sample for this paper included 72 classroom teachers.

During October of 2015 (29-34 days of instruction into the school year), classroom teachers assessed their students' Social Emotional Competence using the *Devereux Student Strengths Assessment* (*DESSA*) *Mini* (Naglieri et al., 2011/2014), a brief 8-item universal screening and progress monitoring tool. The 72 teachers in this study, each completed the *DESSA-Mini* on an average of 23 students (range 4-31), collectively completing *DESSA-Minis* on 1,676 students. At this time, teachers also completed an SEL Programming Survey. Of the 72 teachers in the analysis sample, 70 completed the October SEL Programming Survey. The university human subjects Institutional Review Board approved all research processes.

Sample

The sample in this study is described with valid percentages (see Table 1). It includes 72 credentialed teachers who taught students in transitional-kindergarten through 5th grade. The majority provided general education instruction in English. Although 13% of teachers were new to the district this year, 46% had worked in the district for more than 10 years. Of those who provided a response to the question about their racial/ethnic identity (69 teachers, or 96%), 59% identified as White/European American, 12% as Asian/Asian American, 12% as Hispanic/Latino(a), 7% as multi-race, 6% as Black/African American, and 4% as other.

Teachers in this sample reported that they are generally eager or very eager (77%) to adopt new initiatives at school. When asked about their general preferences for rolling out a new initiative at school, 40% of teachers preferred initial structure with increasing flexibility, 31% preferred initial flexibility with increasing structure, 18% preferred highly flexible at all times, and 5% preferred highly structured at all times. About 6% of teachers reported no preference. Prior to the August TOOLBOX training, no teacher had ever used TOOLBOX, but 13% had observed TOOLBOX in practice and 8% had attended a previous TOOLBOX training. At the end of the training, 90% of teachers rated the training quality as good (53%) or excellent (37%).

Measures

Social Emotional Competence. The *DESSA-Mini* Form 1 (Naglieri et al., 2011/2014) includes eight items that ask the raters the frequency (never = 0, rarely = 1, occasionally = 2, frequently = 3, very frequently = 4) of observed positive behaviors of the child in the past four weeks. The 8-items ($\alpha = 0.95$) are summed to create a Raw Score Total. The Raw Score Total is then converted into a standardized *T*-score based on the national norms, yielding the Social Emotional Total (SET). High SET scores (*T*-scores of 60 and above) are a *strength*, SET scores between 41 and 59 (inclusive) are *typical*, and low SET scores (*T*-scores of 40 and below) point to a *need for instruction*. The U.S. norm sample has been independently reviewed and judged as representative and sufficiently large for interpretation of this nature (Merrell & Gueldner, 2010).

Teacher attitudes. At training, teachers reported the importance of Social Emotional Competence to school success (unimportant = 1 to essential = 5), and their eagerness to use TOOLBOX (not eager at all =1 to very eager =5).

Teacher capacities. At training and in October, teachers reported the extent to which they felt informed (uninformed = 1 to very informed = 5) about TOOLBOX and confident (no confidence = 1 to very confident = 5) in their capacity to implement TOOLBOX. Teachers were also asked to state the mantra associated with 1 of the 12 Tools to directly assess their knowledge about TOOLBOX (0 = incorrect; 1 = correct).

Teacher expectations. At training, teachers reported to what extent they: (1) personally anticipated implementing TOOLBOX relative to others at school (least fully = 1, most fully = 10), (2) believed TOOLBOX would benefit students (no benefit = 1, very beneficial =5), and (3) anticipated challenges in implementing TOOLBOX (low challenge = 1 to high challenge = 5).

Teacher perceptions of implementation and impact. In October, teachers reported to what extent they: (1) were implementing TOOLBOX relative to others at school (least fully = 1, most fully = 10), and (2) believed TOOLBOX has benefited their students (no benefit = 1, very beneficial =5)

School climate. In October, teachers reported the extent to which they perceived barriers to student learning, experienced barriers to providing effective instruction, experienced stress at work, and experienced conflict at work (very low = 0 to very high = 4). Furthermore, teachers reported (very poor = 0 to great = 4) on the overall learning ("I would describe our school as a ____ place for students to learn") and working ("I would describe our school as a ____ place for adults to work") environment of the school.

Social Emotional Competence (SEC) in self, a typical colleague, and a typical student. In October, teachers reported (very low = 0 to very high =4) their own SEC ("Social Emotional Competence refers to an awareness of, and ability to manage emotions in, a context-appropriate manner. How do you think your colleagues would rate your social emotional competence, as it shows up at school?"); that of a typical colleague ("How would you rate the SEC of the typical colleague you work with at school?") and that of a typical student they teach ("Think of a child that is fairly representative of the children with whom you work. How would you rate the Social Emotional Competence of this child?").

Analysis

In order to account for clustering in the data and to address missing data (2.8%-26% across all predictor variables, see Table 2), hierarchical linear modeling with maximum likelihood estimation (Rabe-Hesketh & Skrondal, 2012) was used to estimate the relationship between teachers' ratings of student Social Emotional Competence (*DESSA-Mini* scores; level one) and teachers' self-reported characteristics and perceptions (from pre-training, post-training, and October SEL Programming surveys; level two). First, to identify specific teacher characteristics and perceptions that contribute to teachers' *DESSA-Mini* ratings, each predictor was added to the null model individually. Then, the significant predictors of teacher ratings were included in the final model to estimate their joint contribution to explaining rater bias in this data. Correlations and paired t-tests were used to examine relationships between variables across time points. All analyses were conducted using Stata 7 (Statacorps, 2001).

Results

Social Emotional Competence

Student Social Emotional Competence is indicated through *DESSA-Mini* Social Emotional Total (SET) scores. The average score was 50.88 (*SD* = 11.74). One fourth of the students received scores of 60 and above (strength), 57% received scores between 41 and 59 (typical), and 18% received scores of 40 and below (need for instruction). Approximately 16% (ICC = .16) of the variance in scores was attributable to teacher raters.

Bivariate Relationships between Student Social Emotional Competence and Teacher Characteristics

Teacher attitudes. Before training, nearly all teachers believed that Social Emotional Competence was very important (39%) or essential (60%) to school success and 60% were eager or very eager to implement TOOLBOX. After training, all teachers believed that Social Emotional Competence was very important (33%) or essential (67%) to school success and 83% were eager or very eager to implement TOOLBOX. While teachers' attitude towards Social Emotional Competence started high and remained statistically unchanged ($t(60) = -1.63$, $p = .11$), their eagerness to implement TOOLBOX was significantly higher after training ($t(60) = -4.44$, $p < .001$). No measure of teacher attitudes significantly predicted *DESSA-Mini* ratings (see Table 2).

Teacher capacities. Before training, very few teachers (3%) felt "sufficiently" or "very" informed about TOOLBOX. After training this was significantly higher; most teachers (82%) felt "sufficiently" or "very" informed about TOOLBOX ($t(61) = -16.72$, $p < .001$). In addition, after training, 88% of teachers felt confident or very confident in their capacity to implement TOOLBOX. However, when asked to state the mantra associated with 1 of the 12 Tools, only 35% of teachers provided a correct answer. There was no detectable relationship between feeling informed ($r = .11$, $p = .45$) or confident ($r = .11$, $p = .41$) and knowledge of the TOOLBOX mantra. Approximately 7 weeks after training, with implementation underway, significantly fewer teachers (20%) felt confident or very confident in their capacity to use TOOLBOX ($t(61) = 11.33$, $p < .001$), but a comparable number of teachers (46%) provided the correct response to the TOOLBOX knowledge question ($t = -.42$, $p = .81$). In October, there still was no detectable relationship between confidence and knowledge ($r = .09$, $p = .49$). No teacher

capacities (the extent to which teachers were informed, confident, or knowledgeable) measured at training or in October significantly predicted *DESSA-Mini* ratings in October.

Teacher expectations. At the end of training, teachers had high expectations to fully implement TOOLBOX (M = 7.28, SD = 1.45). Teachers expected, on average, a moderate degree of challenge implementing TOOLBOX (M = 1.84, SD = .77). Before training, 34% of teachers expected TOOLBOX to be very beneficial to their students. After training, teacher expectations were significantly higher; 71% of teachers expected TOOLBOX to be very beneficial to their students ($t(57) = -5.76$, $p < .001$).

At the end of training, the extent to which teachers expected to fully implement TOOLBOX significantly predicted their *DESSA-Mini* ratings (b = 1.00, p = .04**).** In addition, the extent to which teachers anticipated challenges to TOOLBOX implementation predicted *DESSA-Mini* ratings (b = -2.46, p = .003). The extent to which teachers expected TOOLBOX would benefit students at the end of training marginally predicted their *DESSA-Mini* ratings (b = 1.90, p = .09).

Teacher perceptions of implementation and impact. In October, teachers reported moderate levels of implementation (M = 5.95, SD = 2.18), significantly lower than their expectation at the end of training ($t(56) = 5.01$, $p < .001$). Seventy seven percent of teachers agreed or strongly agreed that TOOLBOX had benefited their students. The benefits they perceived during implementation were significantly higher than the benefits they expected at the end of training ($t(55) = 6.88$, $p < .001$). Neither teacher perceptions of implementation nor teacher perceptions of impact significantly predicted their concurrent *DESSA-Mini* ratings (See Table 2).

School climate. In October, 80% of teachers perceived their school to be a good or great place for students to learn. However, 68% reported that the barriers to student learning were high or very high. Similarly, 76% of teachers reported that their school was a good or great place to work, but 64% reported that their stress level at work was high or very high (although only 24% experienced high or very high levels of conflict or tension at work). In October, teachers' perception of barriers to student learning significantly predicted concurrent teachers *DESSA-Mini* ratings (b = -1.60, p = .04). No other concurrent measures of school climate significantly predicted teachers' *DESSA-Mini* ratings.

Social Emotional Competence (SEC) in self, colleagues, and students. In October, teachers reported, on a scale from 0-4, their own SEC as they imagined others perceived it, the SEC of a "typical" colleague, and the SEC of a "typical" student. On average, teachers reported their own SEC (M = 2.76, SD = .69) to be higher than that of a typical colleague (M = 2.49, SD = .68; $t(68) = 2.71$, p = .008). In fact, 70% of teachers reported themselves as having high or very high SEC, while 54% of teachers reported their colleagues as having high or very high SEC. When teachers reported the SEC of their "typical" student (M = 1.69, SD = .69), only 7% reported their students as having high or very high SEC.

Neither teachers' reports of their own SEC nor their reports of their typical colleagues' SEC significantly predicted *DESSA-Mini* ratings. However, teacher reports of their "typical" student's SEC, predicted *DESSA-Mini* ratings (b = 3.35, $p < .001$).

Multivariate Relationship between Student Social Emotional Competence and Teacher Characteristics

Four teacher characteristics (each statistically significant in the bivariate models) were included together in a model to estimate teachers' *DESSA-Mini* ratings (see Table 3). Teachers' higher post-training intent to fully implement TOOLBOX (b = .94, p = .03), and teachers' higher October reports of their "typical" student's SEC (b = 2.79, p = .004) continued to significantly predict higher *DESSA-Mini* ratings. Teacher's higher post-training anticipation of challenge in implementing TOOLBOX (b = -1.90, p = .02) and higher October perception of barriers to student learning (b = -1.70, p = .02) significantly predicted lower *DESSA-Mini* ratings. A likelihood ratio test confirmed that the full multivariate model better fit the data than the null model ($\chi^2(4) = 28.64, p < .001$).Together, these predictors explained nearly 6% of variance in *DESSA-Mini* ratings, as about 10% (ICC = .10) of the variance remained unexplained in the final model.

Discussion

This study explored the extent to which teacher characteristics predicted teacher ratings of student Social Emotional Competence on the *DESSA-Mini*. The *DESSA-Mini* is a behavior rating scale being used to assess approximately a half million children worldwide. Despite their popularity, behavioral rating

scales are known to incorporate rater bias into assessment scores because each item requires interpretation, reflection, and judgment by the rater. We found that only a small amount of the variance in *DESSA-Mini* scores was attributable to raters and that ⅓ of the rater bias could be explained by four rater characteristics: teachers' expectations about their own level of SEL program implementation, anticipation of implementation challenges, perceptions of the barriers their students face, and perceptions of Social Emotional Competence among their students.

Teacher Attitudes, Capacities, and Expectations

Teachers at training felt that Social Emotional Competence was important to school success, and were eager to implement TOOLBOX, but these attitudes did not bias subsequent *DESSA-Mini* ratings. Teachers felt sufficiently informed and confident in their capacity to implement TOOLBOX, but these capacities also did not bias teacher ratings. It is possible that demand characteristics limited the variance in teacher reports of their attitudes and capacities, which attenuated the relationships between these teacher attributes and their *DESSA-Mini* ratings. On the other hand, this potential was minimized by having a third party collect the data, rather than the district or the SEL program developer.

After training, teachers had high expectations for their personal implementation of TOOLBOX. Higher expectations for implementation at training predicted more lenient *DESSA-Mini* ratings in October. Teacher reports of their actual implementation were lower than their earlier expectations, and did not predict their concurrent *DESSA-Mini* ratings. Furthermore, teachers had high expectations for the impact of TOOLBOX after training, and had even higher expectations for benefit seven weeks into implementation, but neither predicted *DESSA-Mini* scores.

After training, teachers anticipated TOOLBOX implementation to be only moderately challenging. Higher anticipation of implementation challenges at training predicted more severe *DESSA-Mini* ratings in October. Teachers were not asked about the extent of the actual challenges that they faced when surveyed in October.

Teacher Perceptions of the School Climate and the Students

Teachers felt the school learning environment was positive, although they perceived high student barriers to learning. Higher teacher perceptions of student barriers to learning predicted more severe *DESSA-Mini* ratings. Teachers also felt the school working environment was positive, despite the high levels of teacher stress. Teachers perceived their colleagues to have above-average Social Emotional Competence, and their own Social Emotional Competence to be even higher. Neither teacher perception of their own reputation for Social Emotional Competence, nor teacher perceptions of their colleagues' Social Emotional Competence, biased concurrent teacher *DESSA-Mini* ratings.

Although teachers generally perceived themselves and their colleagues to have above-average Social Emotional Competence, they perceived their typical student as having below-average Social Emotional Competence. The *DESSA-Mini* scores, however, were fairly comparable in this district to the normative sample. More students in this district (25%) had strengths, relative to the national norm (16%), and a similar number of students in this district (18%) had a need for instruction, relative to the national norm (16%). These data suggest that, before doing a formal assessment, the students were underestimated! Given the barriers that students in this district face (e.g., 70% low-income status), the level of protective factors is impressive as well as important. Higher teacher perceptions of general levels of student Social Emotional Competence predicted more lenient *DESSA-Mini* ratings.

Although teachers' broad-based impressions of their "typical" student explains some variance in *DESSA-Mini* scores, the current study design does not enable us to determine whether perceptions of the "typical" student (a) shape every *DESSA-Mini* rating completed, revealing rater bias, or (b) reflects the actual amount of Social Emotional Competence in his or her students astutely and accurately observed by teachers. To the extent that this is interpreted as an undesirable bias, replication in other samples could be done to determine if the routine collection of this information and a score adjustment is warranted.

Study Limitations

This study was limited in several respects. First, it was conducted in a single Northern California school district, which limits the generalizability of the findings. However, the district includes a diverse student body and the sample is described at length to facilitate judgments about the transferability of the

findings. Second, teachers provided the *DESSA-Mini* ratings and information about themselves, which could create method-bias. However, we think this is appropriate given our research question and frame that teachers are central to the assessment and intervention process. Third, requirements for survey brevity prevented us from using multi-items scales to assess teacher characteristics. This could increase (a) missingness, potentially heightening the risk of sampling error, and (b) measurement error, reducing power to detect effects. Finally, studies have reported that timing in the school year matters; bias is higher when raters are unfamiliar with assessment tools and when the students are less known to the rater (Evans, Allen, Moore, & Strauss, 2005; Hoyt & Kerns, 1999). Future studies should examine the extent of rater bias that exists at the end of the school year, and provide guidance for practitioners and evaluators using the behavior rating scale to measure change over time.

Contributions

An important contribution of this study is clarifying the extent to which DESSA-Mini scores reflect characteristics of the rater, in addition to characteristics of the individual student being rated. We find that approximately 16% of the variance in *DESSA-Mini* scores is attributable to the teacher rater. To the best of our knowledge, this is the first study to report this information about the *DESSA-Mini*, which is somewhat less biased than the 20%-50% of the variance attributable to the rater reported on other tools (Molina, Pelham, Blumenthal, Galiszewski, 1998; Phillips & Lonigan, 2010; Schultz & Evans, 2012). It should be noted, however, that variance attributed to the teacher could also be attributed to the classroom or school environment. Students in the same environment are likely to perform behaviors more similarly to each other than students in different environments. Future research with more schools may want to use a 3-level statistical model to analyze the variance that can be attributable to the school.

Implications & Future Directions

This study has important implications for practice. Overall, TOOLBOX training was well received in this district. Training increased the extent to which teachers felt informed, confident, and eager to use the program, resulting in high intent to fully implement the program and the expectation for student benefit. Once implementation had begun, teachers perceived a greater benefit than they expected,

but they reported lower levels of capacity and were implementing the program less fully than they planned. This suggests that despite the benefits of initial training, potentially due to implementation challenges encountered in routine practice, a booster-training or ongoing technical assistance might be necessary.

Furthermore, it may be that the teacher characteristics identified in this study as sources of rater bias can be remediated through training, intervention planning, and implementation supports in order to shrink the systematic error in student assessments. For example, providing rater training for assessment tools has been shown to reduce, although not eliminate, rater bias. In this study raters were provided with as-needed technical assistance in the completion of the *DESSA-Mini*, but did not participate in any of the *DESSA* trainings available through the Devereux Center for Resilient Children. Future studies should explore whether training to use the *DESSA* specifically, or training about rater bias in general (including the sources identified in this study), would shrink the extent of rater bias in the assessment process. Biases may be unconscious or implicit; the co-creation of such a training with educators may be a particularly generative project.

Interesting implications to guide future research also emerged. In this study, the extent to which teachers felt informed and confident in their capacity to use TOOLBOX was not associated with the extent to which they correctly recalled essential information about the TOOLBOX program. As the best way to assess the knowledge of prevention program implementers remains unresolved in literature (Shapiro, Oesterle, & Hawkins, 2015), it would be important to understand how teacher perceptions of capacity relate to their actual knowledge. Finally, future studies should look for sources of rater bias beyond the field of implementation science to explain remaining (unmeasured sources of) variance. Some studies have found that raters' fixed characteristics (e.g., age, gender) can bias ratings (Schultz & Evans, 2012). Although we only explored theoretically malleable characteristics in the current study, fixed characteristics may be useful for researchers who wish to approximate true levels of student Social Emotional Competence. Future studies should also examine student-level characteristics to see if dyad-specific variance, or interactions between student and teacher characteristics, systematically bias scores.

Psychometrically sound assessment tools may facilitate the discovery and implementation of effective Social Emotional Learning (SEL) programs. Findings of this study help us unpack student assessment scores into their component parts, which may increase our responsible use of behavior rating scales like the *DESSA-Mini* for the rating of student Social Emotional Competence. Responsible use of such tools in research and practice has the potential to facilitate the routine implementation and evaluation of SEL programs to help ameliorate mental, emotional, and behavioral problems in young people.

References

Abbott, R.D., O'Donnell, J., Hawkins, J.D., Hill, K.G., Kosterman, R., & Catalano, R.F. (1998). Changing teaching practices to promote achievement and bonding to school. *American Journal of Orthopsychiatry, 68,* 542-552. doi: 10.1037/h0080363

California Department of Education (n.d.). *Common Core State Standards*. Retrieved from http:// www.cde.ca.gov/re/cc/

Catalano, R.F., Fagan, A.A., Gavin, L.E., Greenberg, M.T., Irwin, C.E., Ross, D.A., & Shek, D.T. (2012). Worldwide application of prevention science in adolescent health. *The Lancet, 379*(9826), 1653-1664. doi: 10.1016/S0140-6736(12)60238-4

Chang, D. F., & Sue, S. (2003). The effects of race and problem type on teachers' assessments of student behavior. *Journal of Consulting and Clinical Psychology, 7*, 235 – 242. doi: 10.1037/0022-006X.71.2.235

Coie, J. D., Watt, N. F., West, S. G., Hawkins, J. D., Asarnow, J. R., Markman, H. J., et al. (1993). The science of prevention. A conceptual framework and some directions for a national research program. *American Psychologist, 48,* 1013-1022. doi: 10.1037/0003-066X.48.10.1013

Collins, M.A. (2015). The TOOLBOX™ *Curriculum Guide (Primer, Lesson Plans K-3, & Lesson Plans 4-6)*. Sebastopol, CA: Dovetail Learning, Inc.

Copeland, W., Wolke D., Shanahan, L., & Costello E. (2015). Adult functional outcomes of common childhood psychiatric problems: A prospective, longitudinal study. *JAMA Psychiatry*, *72*(9), 892-899. doi:10.1001/jamapsychiatry.2015.0730

De Long-Cotty, B. (2010). Report on the TOOLBOX/Sonoma County Collaboration for Resilient Children. Unpublished report, WestEd, San Francisco.

District. (2016). *Data Dashboard.* Retrieved on 5/13/16 from [masked URL].

Domitrovich, C. E., Cortes, R., & Greenberg, M. T. (2007). Improving young children's social and emotional competence: A randomized trial of the preschool PATHS curriculum. *Journal of Primary Prevention, 28*, 67-91. doi: 10.1007/s10935-007-0081-0

Dovetail Learning, Inc. (2013). West Contra Costa Unified School District (WCCUSD) Toolbox Evaluation (Unpublished report).

Durlak, J. A., & DuPre, E. P. (2008). Implementation matters: A review of the research on the influence of implementation on program outcomes and the factors affecting implementation. *American Journal of Community Psychology, 41*, 327-350. doi: 10.1007/s10464-008-9165-0

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

Elias, M. J. (2007). From model implementation to sustainability: A multisite study of pathways to excellence in social-emotional learning and related school programs. In A. M. Blankstein, P. D. Houston, & R. W. Cole (Eds.), *Sustaining professional learning communities: The soul of educational leadership series* (pp. 59–95). Thousand Oaks, CA: Corwin Press.

Elias, M. J., Zins, J. E., Graczyk, P. A., & Weissberg, R. P. (2003). Implementation, sustainability, and scaling up of social-emotional and academic innovations in public schools. *School Psychology Review, 32*(3), 303-319.

Elliott, S.N., Frey, J.R. & Davies, M. (2015). Systems for assessing and improving students' social skills to achieve academic competence. In J.A. Durlak, C.E. Domitrovich, R.P. Weissberg, & T.P. Gullotta (Eds.), *Handbook of social and emotional learning: Research and practice*. New York: Guilford.

Espelage, D.L., Rose, C.A., & Polanin, J.R. (2015). Social-emotional learning program to reduce bullying, fighting, and victimization among middle school students with disabilities. *Remedial and special education, 36*(5), 299-311. doi: 10.1177/0741932514564564

Evans, S.W., Allen, J., Moore, S., & Strauss, V. (2005). Measuring symptoms and functioning of youth with ADHD in middle schools. *Journal of Abnormal Child Psychology*, *33*(6), 695-706. doi: 10.1007/s10802-005-7648-0

Fagan, A.A., Hawkins, J.D., & Shapiro, V.B. (2015). Taking SEL to Scale in Schools: The Role of Community Coalitions. In J.A. Durlak, C.E. Domitrovich, R.P. Weissberg, & T.P. Gullotta (Eds.), *Handbook of social and emotional learning: Research and practice* (pp. 468-481). New York: Guilford.

Feldman, K. A. (1989). Instructional Effectiveness of College Teachers as Judged by Teachers Themselves, Current and Former Students, Colleagues, Administrators, and External (Neutral) Observers. *Research in Higher Education, 30*(2), 137-194. doi:10.1007/BF00992716

Flay, B. R. & Allred, C. G. (2003). Long-term effects of the Positive Action program. *American Journal of Health Behavior, 27*(Supplement 1), 6-21.

Ford, A. (1931). Neutralizing inequalities in rating. Personnel Journal, 9. 466-469.

Foster, G., & Ysseldyke, J. (1976). Expectancy and Halo Effects as a Result of Artificially Induced Teacher Bias. *Contemporary Educational Psychology, 1*(1), 37-45. doi: 10.1016/0361-476X(76)90005-9

Greenberg, M. T., Domitrovich, C. E., Graczyk, P. A., & Zins, J. E. (2005). *The study of implementation in school-based preventive interventions: Theory, research, and practice*. Washington, DC: U.S. Department of Health and Human Services, Substance Abuse and Mental Health Services Administration, Center for Mental Health Services. Final Project Report.

Greenberg, M.T., Weissberg, R.P., O'Brien, M.U., Zins, J.E., Fredericks, L., Resnik, H., et al. (2003). Enhancing school-based prevention and youth development through coordinated social, emotional, and academic learning. *American Psychologist*, *58*(6-7), 466-474. doi: 10.1037/0003-066X.58.6-7.466

Gresham, F. M., Cook, C. R., Vance, M. J., Elliott, S. N., & 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 (English). *Psychological Assessment, 22*(1), 157-166. doi: 10.1037/a0018124

Hoyt, W. T. (2000). Rater bias in psychological research : When is it a problem and what can we do about it? (English). *Psychological Methods, 5*(1), 64-86. doi: 10.1037/1082-989X.5.1.64

Hoyt, W. T., & Kearns, M. D. (1999). Magnitude and moderators of bias in observer ratings: A meta-analysis. *Psychological Methods, 4*, 403-424. doi:10.1037/1082-989X.4.4.403

Kessler R.C, Gruber M.J, Petukhova M, Sampson N.A, Zaslavsky A.M, McLaughlin K.A, … Costello J. (2012). Severity of 12-month DSM-IV disorders in the National Comorbidity Survey Replication Adolescent Supplement. *Arch. Gen. Psychiatry Archives of General Psychiatry*, *69*, 381–389.

Kelter, J. D., & Pope, A. W. (2011). The effect of child gender on teachers' responses to oppositional defiant disorder. *Child & Family Behavior Therapy*, *33*, 49 – 57. doi: 10.1080/07317107.2011.545013

LeBuffe, P.A. & Shapiro, V.B. (2004). Lending 'strength' to the assessment of preschool social-emotional health. *California School Psychologist, 9*, 51-61. doi:10.1007/BF03340907

LeBuffe, P.A., Shapiro, V.B., & Naglieri, J.A. (2009/2014). *The Devereux Student Strengths Assessment (DESSA): Assessment, Technical Manual, and User's Guide.* Charlotte, NC: Apperson, Inc.

LeBuffe, P.A., Shapiro, V.B., Naglieri, J.A. (2015) *DESSA Vragenlijst over Sociaal-Emotionele Competenties - Edizione italiana [Italian Edition]*. (I. Ardizzone, R. Ranaldi, F. Santoro, & S. Galosi, Trans.) Florence: Hogrefe Publisher. (Original work published 2009).

LeBuffe, P.A., Shapiro, V.B., Naglieri, J.A., Pont, S., & Punt, D.J. (2013). *Assessment delle Competenze Socio-emotive legate alla Resilienza Handleiding – Nederlandse Bewerking [Dutch Adaptation]*. Amsterdam: Hogrefe Publisher.

Maras, M.A., Thompson A.M., Lewis, C. Thornburg, K. & Hawks, J. (2015). Developing a tiered response model for social-emotional learning through interdisciplinary collaboration. *Journal of Educational and Psychological Consultation, 25,* 1-26. doi: 10.1080/10474412.2014.929954

Mason, B.A., Gunersel, A.B., & Ney, E.A. (2014). Cultural and ethnic bias in teacher ratings of behavior: A criterion-focused review. *Psychology in the Schools, 51*(10), 1017-1030. doi:10.1002/pits.21800

Masten, A. S. (2014). *Ordinary magic: Resilience in development*. New York: Guilford Press.

McKown, C. (2015). Challenges and opportunities in the direct assessment of Children's Social and Emotional Comprehension. In J.A. Durlak, C.E. Domitrovich, R.P. Weissberg, & T.P. Gullotta (Eds.), *Handbook of social and emotional learning: Research and practice*. New York: Guilford.

Merrell, K. W., & Gueldner, B. A. (2010). *Social and Emotional Learning in the Classroom: Promoting Mental Health and Academic Success*. New York: Guilford Press.

Molina, B. S. G., Pelham, W. E., Blumenthal, J., & Galiszewski, E. (1998). Agreement Among Teachers' Behavior Ratings of Adolescents With a Childhood History of Attention Deficit Hyperactivity Disorder. *Journal of Clinical Child Psychology*, *27*(3), 330–339. doi:10.1207/s15374424jccp2703_9

Naglieri, J.A., LeBuffe, P.A., & Shapiro, V.B. (2011). Universal screening for social emotional competencies: A study of the reliability and validity of the DESSA-mini. *Psychology in the Schools, 48*(7), 660-671. doi: 10.1002/pits.20586

Naglieri, J.A., LeBuffe, P.A., & Shapiro, V.B. (2011/2014). *The Devereux Student Strengths Assessment - Mini (DESSA-Mini): Assessment, Technical Manual, and User's Guide.* Charlotte, NC: Apperson, Inc.

Naglieri, J.A., LeBuffe, P.A., & Shapiro, V.B. (2013). Assessment of social-emotional competencies related to resilience. In S. Goldstein & R. Brooks (Eds.), *Handbook of resilience in children* (pp. 261-272). New York, NY: Kluwer/Academic Press.

Nation, M., Crusto, C., Wandersman, A., Kumpfer, K. L., Seybolt, D., Morrissey-Kane, E., & Davino, K. (2003). What works in prevention: Principles of effective prevention programs. *American Psychologist, 58*, 449–456. doi: 10.1037/0003-066X.58.6-7.449

O'Connell, M.E., Boat, T., & Warner, K.E. (Eds.). (2009). *Preventing mental, emotional, and behavioral disorders among young people: Progress and possibilities*. Washington, DC: National Academies Press.

Pérez-Gómez, A., Mejía-Trujillo, J., Brown, E. C. and Eisenberg, N. (2016). Adaptation and implementation of a science-based prevention system in Colombia: Challenges and achievements. *Journal Community Psychology, 44*: 538–545. doi: 10.1002/jcop.21781

Phillips, B. M., & Lonigan, C. J. (2010). Child and Informant Influences on Behavioral Ratings of Preschool Children. *Psychology in the Schools*, *47*(4), 374–390. doi:10.1002/pits.20476

Rabe-Hesketh, S. & Skrondal, A. (2012). *Multilevel and Longitudinal Modeling using Stata.* Stata Press.

SEL Consulting (2015). Cool to Be Me: Presentation of first year assessment results. Unpublished report, WestEd, San Francisco.

Schultz, B. K., & Evans, S. W. (2012). Sources of bias in teacher ratings of adolescents with ADHD. *Journal of Educational and Developmental Psychology*, *2*(1), 151. doi:10.5539/jedp.v2n1p151

Simmons, C.A., Shapiro, V.B., Accomazzo, S., & Manthey, T.J. (2016). Strengths-Based Social Work: A Meta-Theory to Guide Social Work Research and Practice. In N. Coady & P. Lehmann (Eds.), *Theoretical perspectives for direct social work practice,* (3rd edition, pp. 131-154). New York: Springer Publishing Company.

Shapiro, V.B., Accomazzo, S., Claassen, J., & Fleming, J.L. (2015). The choices, challenges, and lessons learned from a multi-method social-emotional / character assessment in an out of school time setting. *Journal of Youth Development: Bridging Research and Practice, 10*(3): 32-45. Retrieved from: http://jyd.pitt.edu/ojs/index.php/jyd/article/view/5

Shapiro, V.B., Kim, B.K.E., Fleming, J.L., & LeBuffe, P.A. (2016). Protective Factor Screening for Prevention Practice: Sensitivity and Specificity of the DESSA-Mini. *School Psychology Quarterly*.

Shapiro, V.B., Oesterle, S., & Hawkins, J.D. (2015). Relating coalition capacity to the adoption of science-based prevention in communities: Evidence from a randomized trial of Communities

That Care. *American Journal of Community Psychology, 55*(1-2): 1-12. doi: 10.1007/s10464-014-9684-9

StataCorp. 2001. *Stata Statistical Software: Release 7*. College Station, TX: StataCorp LP.