Examination of Method Effects with a Social-Emotional Screening Instrument Across Parent and Teacher Raters

Christine DiStefano[1], Jungsun Go[1], Fred Greer[1], Dexin Shi[1] & Erin Dowdy[2]

[1] University of South Carolina
[2] University of California- Santa Barbara

Abstract:
With psychological studies, method effects are often encountered when items that differ in the direction of the latent construct (e.g., positively and negatively worded items) are included on the same survey. The presence of method effects has been studied; however, largely investigations have used self-report data. As parents and teachers often provide ratings in school-based settings, wording effects may impact scales used to assess children. This study investigated responses to the Behavioral and Emotional Screening System (BESS) Parent and Teacher Rating Scales-preschool level to determine the strength of method effects related to wording when the same child was rated by teachers and parents. Results showed method effects due to oppositely worded items were apparent in the BESS and effects were more pronounced for teachers than for parent raters. Accounting for method effects due to wording allowed higher concordance between scales to be observed.

Keywords:
Method effect, screening, parent rating, teacher rating, wording effect

Introduction

There is much evidence supporting the value of early prevention and intervention of emotional and behavioral problems in school-based settings (Durlak, Weissberg, Dymnicki, Taylor, & Schellinger, 2011). For children at-risk for behavioral and emotional problems, early intervention and assistance may help minimize long-term harm of mental disorders and reduce overall health care burden and costs (Aos, Lieb, Mayfield, Miller, & Pennucci, 2004). To identify problematic behaviors, current efforts in prevention science include initiatives such as Positive Behavior Intervention and Support (PBIS) and the Multitiered System of Support (MTSS) to provide support and assistance to children identified with or with emerging emotional and/or behavioral risk. While PBIS and MTSS differ, both strat-

egies identify children at-risk and provide interventions to reduce risk, prevent the onset, or minimize the effects of a disorder.

With PBIS and MTSS, one of the initial steps in the process is to provide a universal assessment to all students in a school, with the goal of identifying at-risk students. It is estimated that less than 15% of United States public schools currently engage in a systematic screening process for emotional and behavioral risk (Bruhn, Woods-Groves, & Huggle, 2014); however, as more schools are aware of children's mental health needs, the practice is on the rise (Kamphaus, Reynolds, & Dever, 2014). Although several methods are available (e.g., teacher nomination, pediatric referral, parent referral, discipline referrals), school-wide universal screening has been recommended as an optimal approach (Kamphaus et al., 2014). This process typically involves administration of short, targeted surveys to gauge if a child's behavior is outside of "normal" development ranges, considering factors such as age and gender of the child. Screening systems are typically created by reviewing items for evidence of racial and cultural bias, resulting in a method that reduces disproportionality in special education referrals for minority students (Raines, Dever & Kamphaus, 2012).

While screening is becoming more popular across all grade levels, assessing emotional and behavioral risk may be especially critical at the preschool level for many reasons (DiStefano & Kamphaus, 2007; Dowdy et, 2013). Enrollment in pre-kindergarten has increased dramatically in the past decade, with estimates indicating that more than 1.3 million children (32% of all 3- and 4-year olds) attend state-funded preschools (Barnett, Carolan, Squires, Clarke-Browne, 2013). Correspondingly, the number of children entering preschools with emerging social, behavioral, or emotional difficulties is also increasing. For example, among children ages 1-6, approximately 10-13% have emotional or behavioral disorders (Conroy & Brown, 2004). Further, research suggests that behavioral and emotional problems that arise in early childhood are relatively stable and also predictive of negative educational and social outcomes (e.g., Lane, Little, Menzies, Lambert & Wehby, 2010). On a positive note, younger children are thought to be more malleable than older children and may respond better to intervention activities. Thus, prevention and early intervention services for social-emotional and behavioral problems have been recommended for young children based on evidence supporting the positive outcomes following intervention (Brophy-Herb, Lee, Nievar, & Stollak, 2007; Conroy & Brown, 2004).

As preschool is typically when young children in the U.S. begin their schooling, it may be the first setting where at-risk children have access to behavioral support services (DiStefano & Kamphaus, 2007). Given that preschoolers are between 3 and 5 years of age, screening information about a child's behavioral and emotional tendencies is gathered from parents or teachers. As informants, both teachers and parents are generally well-equipped to discuss a child's emotional and behavioral functioning (Smith, 2007) and, as raters, both groups have advantages and disadvantages. Parents are able to provide screening information on their child's functioning earlier than teachers and could provide information prior to the school year while teachers need four- to six-weeks of familiarization with a child to accurately evaluate behavior. Teachers, however, may be better judges of what constitutes "normal" development, as they deal with many students and may have years of prior experience (DiStefano & Kamphaus, 2007).

Concordance Across Behavior Ratings

An extensive literature base has noted differences in children's social-emotional/behavior ratings when different raters are assessing the same children (e.g., Achenbach, McConaughy, & Howell, 1987; De Los Reyes & Kazdin, 2005; Gresham, Elliott, Cook, Vance & Kettler, 2010; Greenbaum, Dedrick, Prange, & Friedman, 1994; Efstratopoulou, Janssen, & Simons 2012). Studies have consistently found low to moderate Pearson Product Moment correlations (r = .20-.30) between parent and teacher ratings on similar emotional-behavioral scales (Achenbach et al., 1987; Gresham et al., 2010).

Various reasons have been proposed for why a low concordance between ratings may occur, such as setting (e.g., home, school, Efstratopoulou, et al., 2012) scale used (Gresham et al., 2010), or rater characteristics (e.g., anxiety, depression, De Los Reyes & Kazdin, 2005). Another area that has been suggested is scale construction, including issues of scaling systems used in the questionnaire and how these scales are utilized (Mandal, Olin, Wilczynski, 1999).

As screening measures are rating scales, these forms are often constructed using commonly accepted principles. Screening instruments typically use a Likert-type item format with relatively few categories (i.e., between 2- and 5-categories) and have all scale-points anchored to assist with defining context for informants (Greer & Liu, 2016). Readability is a consideration for some respondents, with many parent forms written at an elementary grade reading level (Reynolds & Kamphaus, 2015). In addition, screeners may include items with wording that varies in "direction" of a construct focused on risk behaviors.

Test developers typically use items varying in wording direction for a variety of reasons. This procedure is often used to make sure respondents are carefully reading the questions and providing responses which are in line with the direction of the wording (DeVellis, 2016). In addition, screening instruments that measure "risk" may include positively framed questions focusing on a child's strengths (Reynolds & Kamphaus, 2015). However, including items of differing wording direction together on a questionnaire has itself been identified as problematic.

If differences in responses are found to be due to wording direction, then the resulting data are likely confounded by a method effect caused by the mechanism used to collect the information, even after recoding to place all items in the same direction (Weems, Onwuegbuzie, Schreiber, & Eggers, 2003).

Modeling Method Effects

When using a questionnaire to collect data, items in the "opposite" direction of the focal construct may result in a method effect. Even reverse coding items to follow the same direction may allow for method effects related to item wording to remain (e.g., Barnette, 2000; Horan, DiStefano, & Motl, 2003). The presence of method effects due to wording has been consistently observed in situations such as: scales measuring different content (e.g., Motl, Conroy & Horan, 2000; Rauch, Schweizer, Moosbrugger, 2007; Ye & Wallace, 2014), with very different populations of individuals (e.g., Tomás & Oliver 1999, Roszkowski & Soven, 2010; Wang, Siegal, Falck & Carlson, 2001) and across time (Motl & DiStefano, 2002). Including items of mixed direction on the same scale

increases the response burden and, in turn, increases the measurement error associated with the responses. Consequences due to mixing item formats generally result in lower reliability estimates and validity coefficients than if items were worded in only one direction (Dalal & Carter, 2015; Castro-Schilo, Widaman & Grimm, 2013). Integrating method effects into analyses allows for clearer study of latent constructs of interest by removing variance that is related to the way in which information was obtained (Schweizer & Troche, 2016) from the construct itself.

Various methods have been suggested to model method effects. For example, several researchers have linked a multitrait-multimethod (MTMM) conceptual framework (Campbell and Fiske, 1959) with a confirmatory factor analysis (CFA) analysis strategy (Jöreskog, 1974) to assist in the separation and empirical estimation of substantive and methods components in investigations of method effects (Marsh, 1989; Marsh & Grayson, 1995; Bagozzi, 1993). This MTMM theoretical framework has been applied with CFA in analyses of method effects using correlated traits, correlated uniquenesses (CTCU) models, or correlated traits, correlated methods (CTCM) models, or a combination of the models. Such models allow for the separation of substantive content from the mechanism or method used to gain the responses.

Typically, self-report data has been used to study method effects due to item wording; however, the effects may arise regardless of respondent. As screening information may be used to identify young children at-risk for behavioral problems, it is of interest to examine if a method effect exists in teacher and parent ratings of behavioral screening data. The current study investigated if method effects due to item wording were present for parents and/or teachers in a widely used behavioral/emotional screening instrument. Using a MTMM-CFA approach, we investigated the presence of method effects to determine if parents or if teachers reported higher levels of association. After accounting for method effects, concordance across scales and raters was examined. In addition, we investigated characteristics of the child being rated to determine if demographic characteristics affected the presence of method effects.

Method

Data

Data were collected as part of a funded project investigating universal screening information provided by parents and teachers of preschool students. Behavioral ratings were collected from public schools in the fall of three academic years (2016, 2017, 2018) from nine primary schools in two U.S. states (California and South Carolina). Institutional Review Board approval was obtained by both institutions prior to conducting the study. Both parents' and teachers' participation in the project was voluntary, and teachers received small stipends ($25) for completing screener information for all students in their classroom.

The sample used consisted of 1,007 ratings of preschoolers by both teacher and parent raters. Preschool children were between the ages of 3 years, 0 months and 5 year, 11 months with average age of 4 years, 2 months (SD = 6 months) and roughly even across female (n = 482, 47.9%) and male children (n = 498, 49.5%) with 2.7% (n = 27) children missing gender information. The

sample of children rated was primarily Caucasian (42.6%) but included children from many different backgrounds including Hispanic (29.6%), African American, (19.7%), Asian, (1.0%) Multiracial, (3.9%) and "Other" (0.5%); race was not reported for 2.8% of the children.

Instrumentation

We collected data using the Behavioral and Emotional Screening System (BESS) Parent Rating Scale-Preschool (PRS-P) and Teacher Rating Scale-Preschool (TRS-P) instruments (Kamphaus & Reynolds, 2015). Both forms assess of three broad behavioral dimensions of behavioral and emotional risk: however, item wording differs for most items to allow the unique perspective of a rater. The Externalizing Risk dimension consists of items associated with externalizing behaviors, such as hyperactivity, aggression, and conduct problems (e.g., "Hits other children"). The Internalizing Risk dimension includes items assessing anxiety, depression, and somatization, which are characteristic of internalizing behaviors (e.g., "Is easily upset"). The Adaptive Skills Risk dimension assesses core characteristics of adaptive behavior including adaptability, social skills, and activities of daily living important for functioning at home and school, and in the community (e.g., "Responds appropriately when asked a question"). Adaptive Skills questions are "positively" stated and are recoded prior to analysis.

The TRS-P includes 20 items, with six items ascribed to each of the three (Externalizing Risk, Internalizing Risk, Adaptive Skills Risk) subscales. Per the manual recommendations, two reverse coded items measuring Attention Problems do not belong to one of the three subscales at the preschool level and are not included in the analysis. The PRS-P screener includes 29 items, with nine items measuring each of the three scales. Again, two items measuring Attention Problems are not included in the analyses; however, these items are not in the opposite direction on the parent form. For both parent and teacher scales, all Adaptive Skills items are in the opposite direction and are recoded prior to analysis; the latent construct associated with these items is a risk variable. For all latent variables on the BESS forms, higher scores denote higher levels of risk.

As part of universal screening, teachers completed the TRS-P form for all students in their classroom six weeks after the start of the academic year. At the same time, screening forms were sent home for parents to complete. All items on the BESS were rated on a four-point Likert scale, with anchors of "Never" = 1, "Sometimes" = 2, "Often" = 3, and "Always" =4. Raters assessed the frequency with which a specific behavior was observed over the past few weeks.

Analyses

Analyses were conducted using MPlus (version 8.4, Muthén & Muthén, 2004). Prior to analysis, distribution characteristics were examined. The weighted least squares mean- and variance-adjusted (WLSMV) estimator was chosen to accommodate the ordinal level of the data (e.g., Finney & DiStefano, 2013). Upon investigation, the item-level Likert data were found to be asymmetric, with relatively few ratings identifying a high level of risk for the specified behaviors (i.e., Sometimes or Almost Always). As WLSMV may pose estimation problems when there is sparse data in some of the categories, the four-point Likert scale

was collapsed to a three-point scale (DiStefano, Shi & Morgan, 2020). Even after collapsing the number of scale points from four to three, Cronbach's alpha reliability estimates were acceptable for the TRS-P (Externalizing Problems Risk = .91, Internalizing Problems Risk = .82, Adaptive Skills Risk = .87) and PRS-P (Externalizing Problems Risk = .84, Internalizing Problems Risk = .72, Adaptive Skills Risk = .83) scales, with values approximating reliability estimates noted in the BESS technical manual (Reynolds & Kamphaus, 2015).

Three models were tested. First, a six-factor CFA was run, including parent and teacher information together without the presence of method effects. Second, a method effect factors were included for reverse coded items, with separate factors modeled for parent and for teacher raters. Method effect factors were allowed to correlate, however, there were no associations allowed between content and method factors. The final model added characteristics of age and gender to the method effects model. Here, age was reported as a continuous variable, reported in years and months. Figure 1 illustrates the CTCM model (Model 2) and describes additional models tested.

Models were evaluated using a selection of six fit indices, which focus upon different aspects of model fit. The fit indices were chosen on the basis of recommendations from Gerbing and Anderson (1993), Hu and Bentler (1999), and Tanaka (1993): (1) Chi-square statistic; (2) Tucker Lewis index (TLI); (3) comparative fit index (CFI), (4) root mean square error of approximation and associated 90% confidence interval (RMSEA); (5) standardized root mean residual (SRMR) and the (6) weighted root mean square (WRMR). In addition, we recognize that the robust nature of the WLSMV estimator results in adjustments to standard errors of parameter estimates, the chi-square fit statistic, and as indices which use the chi-square in calculations (Finney & DiStefano, 2013).

The chi-square statistic assesses absolute fit of the model to the data, but the statistic is sensitive to sample size and assumes the correct model is tested (Bollen, 1989; Jöreskog, 1993; Jöreskog & Sörbom, 1996). Both the Tucker-Lewis Index (TLI) and Comparative Fit Index (CFI) are incremental fit indices and test the proportionate improvement in fit between the tested model and a baseline model with no correlations among observed variables (Bentler, 1990; Bentler & Bonett, 1980). NNFI and CFI values approximating 0.95 were indicative of good fit (Hu & Bentler, 1999). The root mean square error of approximation (RMSEA) represents closeness of fit (Browne & Cudeck, 1993). The RMSEA value should approximate or be less than 0.05 to demonstrate close fit of the model (Browne & Cudeck, 1993). The 90% confidence interval (CI) around the RMSEA point estimate should contain 0.05 to indicate the possibility of close fit (Browne & Cudeck, 1993). The Standardized Root Mean Square Residual (SRMR) provides an estimate of the amount error remaining; values of .05 or lower indicate little residual error (DiStefano, 2016). Finally, the Weighted Root Mean Square value was calculated. This value provides an assessment of model misfit when categorical data are used. Values should close to 1.0 to illustrate a well-fitting model (DiStefano, Liu, Jiang, & Shi, 2018).

After model evaluation, local fit was examined. To evaluate local fit, standardized item parameter values were examined for significance and strength and associated standard errors of parameter estimates were examined for precision. Model modi-

fication indices and the standardized residual matrix were also evaluated to identify possible areas of misfit.

Results

Fit indices for the tested models are reported in Table 1. For the sample of children with parent and teacher BESS scores, the six-factor model fit the data reasonably well. While CFI, TLI, and SRMR are below the recommended cutoff values, RMSEA is acceptable. The addition of method factors for teacher and parents greatly improved fit. This model demonstrated most fit indices within acceptable levels. WRMR is a bit higher than its recommended cutoff value of 1.0; however, this index is sensitive to large models and large sample sizes (DiStefano et al., 2018). Thus, adding a method factor to help account for wording with 'reverse coded' items had a great improvement upon model-data fit.

Investigating local fit of the method effects model showed differences by rater. As the goal was to examine method effects due to opposite wording, parameter information is available upon request. For teacher ratings, all parameter estimates for items relating to behavioral trait factors and the method effects factor were significant. Considering parent ratings, all parameter estimates for items relating to behavioral trait factors were significant; however, selected items on the method effect factor were not significant. One last point of interest is that standard error of parameter estimates were lower for TRS-P ratings than for PRS-P, illustrating less variability in estimating parameters for teacher raters than for parent raters. Modification indices and residual estimates did not offer any item deletions or parameter additions that were substantive in nature.

Table 2 provides information about recoded items and variance shared between the target factor and the methods effect factor. As the BESS consists of copywrite materials, we are not allowed to reproduce exact item content. Instead, we refer to the item number on the respective form (TRS-P or PRS-P) and the underlying dimension. As the method and trait factors are orthogonal

Model	Chi-sq	df	RMSEA 90%CI	CFI	TLI	SRMR	WRMR
No Method Effects	3917.42	1021	.053 .051-.055	.919	.914	.073	2.21
Including Method Effects for Opposite Items	2460.80	914	.041 .039-.043	.956	.952	.058	1.53
Including Method Effects for Opposite Items and Covariates of Age & Gender	2460.33	988	.039 .037-.041	.957	.957	.067	1.48

Table 1 Fit indices of Tested Models, BESS TRS-P and PRS-P Screening Instruments (N = 1,007)

Table 2 Fit indices of Tested Models, BESS TRS-P and PRS-P Screening Instruments (N = 1,007)

Form	Recoded Item	Adaptive Skills Construct Variance	Method Variance
TRS-P	Item3	.750	.029
	Item6	.188	.733
	Item7	.689	.043
	Item13	.821	.037
	Item18	.340	.540
	Item20	.309	.175
PRS-P	Item1	.494	.010^{ns}
	Item4	.256	.318
	Item5	.225	.534
	Item10	.433	.047
	Item13	.389	.263
	Item18	.424	.009^{ns}
	Item22	.623	.008^{ns}
	Item28	.585	.012
	Item29	.301	.180

Note: ^{ns} denotes a non-significant parameter value

in a CTCM model, the standardized parameter estimate may be squared to denote the amount of variance shared between the item and the respective factor. Reverse worded items were concentrated on the Adaptive Skills dimension for both parent and teacher forms. Investigating the parameter estimate values for the method effects factor showed that teachers were more effected by opposite wording than parents. All items on the method effect factor were significant for teacher raters while roughly half of the method effects parameters were significant for parent raters.

Concordance Across Raters

To examine concordance among raters, correlations between scales were examined. Information for the three sets of correlation values are provided in Table 3. Three sets of correlation evidence were reported. First, correlations between scale scores (as commonly provided in technical manuals) was

Table 3 Correlations between Raters

	External-Teacher	Internal – Teacher	Adaptive-Teacher	External–Parent	Internal - Parent	Adaptive-Parent
External – Teacher		.376	.537	.339	.141	.217
Internal – Teacher	.500 (.500)		.481	.197	.160	.212
Adaptive– Teacher	.789 (.693)	.684 (.596)		.185	.102	.224
External – Parent	.390 (.389)	.253 (.253)	.286 (.238)		.709	.446
Internal – Parent	.178 (.178)	.192 (.192)	.160 (.133)	.899 (.899)		.419
Adaptive– Parent	.292 (.270)	.293 (.270)	.232 (.294)	.629 (.572)	.545 (.491)	

Notes: External = Externalizing Risk, Internal = Internalizing Risk, Adaptive = Adaptive Skills Risk. Top half of the matrix consists of Pearson Product Moment correlations between BESS TRS-P and PRS-P subscale scores. Bottom half of matrix reports correlations between correlations with latent variables when method effects were removed, values in parenthesis are when method effects were not modeled.

computed using Pearson Product Moment (PPM) correlation estimates. Second, relations among content factors were examined at the latent level when method effects were included and when no method effects were included.

Examining PPM estimates showed positive, yet moderate to low concordance across parent and teacher raters, with values in line with previous research. As expected, correlations among scales for the same rater (e.g., parents) or across raters were highest. Ratings same subscale (e.g., Externalizing Risk) were roughly .20, while ratings across different subscales and raters were roughly .10.

Values were higher when concordance between raters was considered at the latent level. As expected, correlations followed a similar pattern as with the correlation estimates. As expected, correlations across latent variables were higher than PPM values when error was removed. Comparing the correlations among latent variables when method effects were removed, these values were higher than correlations with the other two approaches. Finally, there was a positive, moderate correlation of .54 between parent and teacher method effect factors. Thus, ratings for recoded items across parent and teacher raters were related, even when accounting for the Adaptive Skills Risk scale across raters.

Covariates and Method Effects

Building off of the method effects model, covariates of age and gender were included in the model to determine if these characteristics influenced observation of method effects. Table 1 indicated acceptable global fit. As the purpose was to investigate the influence of covariates on method effects, we focus on the relation between demographic characteristics and the presence of method effects for parent and teacher raters.

Gender demonstrated a significant influence (β = -.19, $p < .001$) on the presence of method effects for parent raters. As males were dummy coded "1" and the parameter value negative, the interpretation is that responses to opposite worded questions were more pronounced for male children. There was no statistically significant effect of gender observed for teachers. When considering age of the child being rated, parents did not demonstrate any effect of age. Teachers, however, noted a negative relation with age (β = -.11, p = .005) and the method effect factor noting that method effects were more often observed with younger children.

Discussion

When creating questionnaires, survey constructors may use items that are in the "opposite" direction of the focal construct. However, items in the opposite direction may contribute to the presence of method effects due to wording of items which is in direct opposition of the focal construct (DeVellis, 2016; Horan et al., 2003). The purpose of the current study was to determine if method effects due to opposite worded items were present in a behavioral screening instrument administered to parents and teachers. The form used, the Behavioral and Emotional Screening System (BESS) Teacher Rating Scale-Preschool (TRS-P) and Parent Rating Scale-Preschool (PRS-P) were used with ratings of the same preschool children, collected at the start of an academic year. Using roughly 1,000 ratings, the BESS TRS-P and PRS-P screening instruments were examined to determine if method effects were present using a CTCM model. In addition, the amount of variance accounted for by item values were compared to method effects.

In general, the addition of a method effects factor for items worded in the opposite direction greatly improved model-data fit. Accounting for wording in the opposite direction has been found to improve model data fit (e.g., DiStefano & Motl, 2006; Rauch et al. 2007; Tomás & Oliver 1999; Wang et al., 2001); however, a difference with the current study is that the raters are proxy ratings for young children rather than self-report ratings. For teachers, age of the child rated predicted the presence of method effects; for parents, gender was significantly related to the presence of a method effect. Previous studies investigating demographic characteristics of the BESS TRS-P identified age and gender characteristics of the students were related to teacher ratings of the constructs (DiStefano, Ene, & Leighton, 2016). This study found that these same demographic characteristics also impacted the presence of method effects, yet what predicted the occurrence of a method effect differed by teacher or parent rater.

With the BESS TRS-P and PRS-P instruments, Adaptive Skills items denote positive behaviors, and thus are in the opposite direction of the focal construct (i.e., Adaptive Skills Risk). While test developers consider it a benefit to include items mea-

suring behavioral strengths as well as items measuring behavioral risk (Reynolds & Kamphaus, 2015), the positively worded exhibited a method effect, which was present even after item recoding. Including items within the same survey has been found to be difficult for raters to process (Roszkowski & Soven, 2010; Salzar, 2015) and adversely affect the measurement of the construct. Further, including scales of combined items may lead to inconsistent responses (Salazar, 2015). As the goal of school-based screening is to assist young children, inconsistency in responding may result in a child not receiving a score that would flag them as needing assistance.

When considering the method effects with the BESS scales, parents reported less of a method effect than teachers for most of the recoded items. Considering magnitude of the method variance relative to trait variance, there were two items on the TRS-P form which illustrated larger variance associated with the method factor than to the underlying factor. These two items may illustrate higher method variance due to ambiguity in item content. For example, on the teacher form, Item3 asked teachers to rate if the student is a "good sport" where Item 18 asks if the student "responds appropriately to questions". Additionally, parent ratings demonstrated more variance associated with the method factor for Item 5, regarding student's ability to "communicate clearly". Some of the variance related to the negative wording effect may be due to the ambiguity of the item wording, which, in combination with the recoded nature of the items, causes additional problems with respondents. (e.g., Barnett, 2000; Roszkowski & Soven 2010).

While rater concordance was similar acoss different methods, removing variance due to items of opposite wording helped to provide a clearer view of the relations between constructs on the BESS and different raters. Latent correlation values when method effects due to opposite wording were removed showed higher relations between scales as compared to observed correlations, or even correlations among latent variables without removing method effects. Prior research studies have noted low correlations among different raters for different scales (e.g., Achenbach et al, 1987; De Los Reyes & Kazdin, 2005; Gresham et al 2010; Greenbaum et al 1994; Efstratopoulou et al, 2012). While correlations between scales are still similar (i.e., positive, low-moderate values) when measurement error and method effects are removed, the values are higher. Thus, removing method effect variance and estimating relations on the latent level may better illustrate the "true" relations between raters and scales. As screener information may be important for school personnel to provide interventions to students found to be at-risk for behavioral and or emotional problems, using items worded in one direction may provide the best estimate of children's behavioral development.

Limitations and Future Research

As with any study, we recognize that there are limitations with the current research. The data were collected from selected locales in the U.S. Future studies may examine results with larger samples. In addition, self-reports for young children may be incorporated into future work to examine method effects with children on screening instruments.

Another limitation we recognize is that the data for teacher raters is nested, as teachers completed screening instruments

for all students in their classroom. Given that the data were matched (teacher – parent pairs of ratings), we could not include a design effect to accommodate nesting for teachers as this was apparent only for teachers. We recognize that not accounting for nesting may affect standard errors of parameter estimates (Stapleton, 2013). Future studies may incorporate nesting into the models to see the effect upon results.

Finally, alternative methods have been suggested to model method effects. For example, the correlated traits, correlated methods model less one [CT(M-1)], has been proposed to overcome estimation problems observed with the CTCM framework and still allows for decomposition of variance into trait and method components (Eid, Lischetzke, Nussbeck, & Trierweiler, 2003). In addition, the CTC(M-1) framework may be useful to untangle effects of different raters on variance. The fixed effect model (Maydeu-Olivares & Coffman, 2006) incorporates methods effects as an intercept factor (i.e., fixed values at 1) and analyzes the specific effects of different groups that are not randomly chosen to contrast their rating patterns. Finally, the trifactor model (von der Embse, Kim, Kilgus, Dedrick, & Sanchez, 2017) has been used with screener ratings to decompose variance into construct, rater, and item sources. Future studies may consider alternative methods of estimating method effects to compare differences among raters completing behavioral scales.

As MTSS procedures are becoming more commonplace in the school environment, parent and/or teacher raters may be asked to complete emotional/behavioral screening scales to proactively identify risk and allow for interventions to take place. The BESS scale is a popular instrument used across the U.S. for school-based universal screening. The results here found the presence of method effects due to opposite wording were apparent and were more pronounced for teachers than for parents. Incorporating method effects can assist with screener information in many ways, such as: clearly demonstrating the relations between constructs, provide greater support for validity studies, and could even contribute to scoring procedures for risk identification. These changes may help researchers, parents, and school personnel to provide a clearer view of the constructs measuring child behavior.

References

Achenbach, T. M., McConaughy, S. H., & Howell, C. T. (1987). Child/adolescent behavioral and emotional problems: implications of cross-informant correlations for situational specificity. *Psychological Bulletin*, *101*(2), 213-323.

Aos, S., Lieb, R., Mayfield, J., Miller, M., & Pennucci, A. (2004). Benefits and costs of prevention and early intervention programs for youth. *Washington State Institute for Public Policy – Technical Appendix*, Olympia, WA.

Bagozzi, R. P. (1993). Assessing construct validity in personality research: Applications to measures of self-esteem. *Journal of Research in Personality*, *27*(1), 49-87.

Barnett, S. W., Carolan, M. E., Squires, J. H., & Clarke Brown, K. (2013). *The state of preschool 2013*. New Brunswick, NJ: National Institute for Early Education Research.

Barnette, J. J. (2000). Effects of stem and Likert response option reversals on survey internal consistency: If you feel the need, there is a better alternative to using those negatively worded stems. *Educational and Psychological Measurement*, *60*(3), 361-370.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, *88*(3), 588-606.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*(2), 238-246.

Bollen, K. A. (1989). *Structural Equations with Latent Variables*. New York, NY: John Wiley.

Brophy-Herb, H. E., Lee, R. E., Nievar, M. A., & Stollak, G. (2007). Preschoolers' social competence: Relations to family characteristics, teacher behaviors and classroom climate. *Journal of Applied Developmental Psychology*, *28*(2), 134-148.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136-162). Newbury Park, CA: Sage.

Bruhn, A. L., Woods-Groves, S., & Huggle, S. (2014). A preliminary investigation of emotional and behavioral screening practices in K–12 schools. *Education and Treatment of Children*, *37*, 611–634. doi:10.1353/etc.2014.0039.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81-105.

Castro-Schilo, L., Widaman, K. F., & Grimm, K. J. (2013). Neglect the structure of multitrait-multimethod data at your peril: implications for associations with external variables. *Structural Equation Modeling: A Multidisciplinary Journal*, *20*(2), 181-207.

Conroy, M. A., & Brown, W. H. (2004). Early identification, prevention, and early intervention with young children at risk for emotional or behavioral disorders: Issues, trends, and a call for action. *Behavioral Disorders, 29,* 224-236.

Dalal D. K., & Carter N. T. (2015). Negatively worded items negative impact survey research. In L. E. Lance & R. J. Vandenberg (Eds.), *More Statistical Myths & Urban Legends* (pp. 112–132). New York, NY: Routledge.

De Los Reyes, A., & Kazdin, A. E. (2005). Informant discrepancies in the assessment of childhood psychopathology: A critical review, theoretical framework, and recommendations for further study. *Psychological Bulletin, 131,* 483–509.

DeVellis, R. F. (2016). *Scale development: Theory and applications* (Vol. 26). Thousand Oaks, CA: Sage.

DiStefano, C. (2016). Examining fit with structural equation models. *Principles and methods of test construction. Principles and methods of test construction: Standards and recent advancements* (pp.166-193), Göttingen, Germany: Hogrefe Publishers.

DiStefano, C. A., & Kamphaus, R. W. (2007). Development and validation of a behavioral screener for preschool-age children. *Journal of Emotional and Behavioral Disorders*, *15*(2), 93-102.

DiStefano, C., Ene, M., & Leighton, E. (2016). Teacher ratings of child behavior in preschool: A MIMIC investigation of the BESS TRS-P. *Psychological Assessment*, *28*(8), 1015-1019.

DiStefano, C., Liu, J., Jiang, N., & Shi, D. (2018). Examination of the weighted root mean square residual: Evidence for trustworthiness? *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(3), 453-466.

DiStefano, C., Shi, D., & Morgan, G. B. (2020). Collapsing categories is often better than modeling sparse data. Structural Equation Modeling: A Multidisciplinary Journal, DOI: 10.1080/10705511.2020.1803073

Dowdy, E., Furlong, M., Raines, T. C., Bovery, B., Kauffman, B., Kamphaus, R. W., ... & Murdock, J. (2015). Enhancing school-based mental health services with a preventive and promotive approach to universal screening for complete mental health. *Journal of Educational and Psychological Consultation*, *25*(2-3), 178-197.

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

Efstratopoulou, M., Janssen, R., & Simons, J. (2012). Agreement among physical educators, teachers and parents on children's behaviors: A multitrait–multimethod design approach. *Research in Developmental Disabilities*, *33*(5), 1343-1351.

Eid, M., Lischetzke, T., Nussbeck, F. W., & Trierweiler, L. I. (2003). Separating trait effects from trait-specific method effects in multitrait-multimethod models: A multiple-indicator CT-C (M-1) model. *Psychological Methods*, *8*(1), 38-60.

Finney, S. J. & DiStefano, C. 2013. Nonnormal and categorical data in structural equation models. In *A Second Course in Structural Equation Modeling*, 2nd Edition, Edited by: Hancock, G. R. and Mueller, R. O. 439–492. Charlotte, NC: Information Age.

Gerbing, D. W. & Anderson, J. C. (1993). Monte Carlo evaluations of goodness-of-fit indices for structural equation models. In K. A. Bollen and J. S. Long (Eds.), *Testing Structural Equation Models,* (pp. 4-65). Newbury Park, CA: Sage.

Greenbaum, P. E., Decrick, R. F., Prange, M. E., & Friedman, R. M. (1994). Parent, teacher, and child ratings of problem behaviors of youngsters with serious emotional disturbances. *Psychological Assessment*, *6*(2), 141-148.

Greer, F. W., & Liu, J. (2016). Creating short forms and screening measures. In K. Schweizer & C. DiStefano (Eds.), *Principles and methods of test construction: Standards and recent advancements* (pp. 272-287). Göttingen, Germany: Hogrefe Publishers.

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*(1), 157-166.

Horan, P. M., DiStefano, C., & Motl, R. W. (2003). Wording effects in self-esteem scales: Methodological artifact or response style? *Structural Equation Modeling: A Multidisciplinary Journal*, *10*(3), 435-455.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*, 1–55

Jöreskog, K. G. (1993). Testing structural equation models. *Sage focus editions*, *154*, 294-294.

Jöreskog, K. G. (1974). Analyzing psychological data by structural analysis of covariance matrices. In D. H. Krantz, R. C. Atkinson, R. D. Luce, & P. Suppes (Eds.), *Contemporary Developments in Mathematical Psychology* (Vol 2, pp. 1-56). San Francisco: Freeman.

Jöreskog, K. G., & Sörbom, D. (1996). *LISREL 8: User's reference guide*. Scientific Software International.

Kamphaus, R. W., & Reynolds, C. R. (2007). *BASC-2 Behavioral and Emotional Screening System Manual*. Circle Pines, MN: Pearson.

Kamphaus, R. W., Reynolds, C. R., & Dever, B. V. (2014). Behavioral and mental health screening. In R. J. Kettler, T. A. Glover, C. A. Albers, & K. A. Feeney-Kettler (Eds.), *Universal screening in educational settings: Evidence-based decision making for schools* (pp. 249–273). Washington, DC: American Psychological Association. doi: 10.1037/14316-010.

Lane, K. L., Little, A., Menzies, H., Lambert, W., & Wehby, J. (2010). A comparison of students with behavior challenges educated in suburban and rural settings: Academic, social, and behavioral outcomes. *Journal of Emotional and Behavioral Disorders*, *18*(3), 131-148.

Mandal, R. L., Olmi, D. J., & Wilczynski, S. M. (1999). Behavior rating scales: Concordance between multiple informants in the diagnosis of attention-deficit/hyperactivity disorder. *Journal of Attention Disorders, 3,* 97–103.

Marsh, H. W., & Grayson, D. (1995). Latent variable models of multitrait-multimethod data. In R. Hoyle (Ed.), *Structural equation modeling: Concepts, issues and applications*, (pp. 177–198). Thousand Oaks, CA: Sage.

Marsh, H. W. (1994). Confirmatory factor analysis models of factorial invariance: A multifaceted approach. *Structural Equation Modeling: A Multidisciplinary Journal*, *1*(1), 5-34.

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

Motl, R. W., & DiStefano, C. (2002). Longitudinal invariance of self-esteem and method effects associated with negatively worded items. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(4), 562-578.

Motl, R. W., Conroy, D. E., & Horan, P. M. (2000). The Social Physique Anxiety Scale: an example of the potential consequence of negatively worded items in factorial validity studies. *Journal of Applied Measurement,* 1, 327–345.

Muthén, L. K., & Muthén, B. O. (2004). *Mplus user's guide: Statistical analysis with latent variables: User's guide*. Muthén & Muthén.

Raines, T. C., Dever, B. V., Kamphaus, R. W., & Roach, A. T. (2012). Universal screening for behavioral and emotional risk: A promising method for reducing disproportionate placement in special education. *The Journal of Negro Education*, *81*(3), 283-296.

Rauch, W. A., Schweizer, K., & Moosbrugger, H. (2007). Method effects due to social desirability as a parsimonious explanation of the deviation from unidimensionality in LOT-R scores. *Personality and Individual Differences*, *42*(8), 1597-1607.

Reynolds, C. R., & Kamphaus, R. W. (2015). *BASC-3 behavioral and emotional screening system manual.* Circle Pines, MN: Pearson.

Roszkowski, M. J., & Soven, M. (2010). Shifting gears: Consequences of including two negatively worded items in the middle of a positively worded questionnaire. *Assessment & Evaluation in Higher Education*, *35*(1), 113-130.

Salazar, M. S. (2015). The dilemma of combining positive and negative items in scales. *Psicothema*, *27*(2), 192-199.

Schweizer, K., & Troche, S. (2016). Method effects in psychological assessment due to item wording and item position. In K. Schweizer & C. DiStefano (Eds.), *Principles and methods of test construction: Standards and recent advancements*. Göttingen, Germany: Hogrefe Publishers.

Smith, S. R. (2007). Making sense of multiple informants in child and adolescent psychopathology: A guide for clinicians. *Journal of Psychoeducational Assessment*, *25*(2), 139-149.

Stapleton, L. (2013). Multilevel structural equation modeling with complex sample data. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (2nd ed., pp. 521–562). United States of America: Information Age Publishing Inc.

Tanaka, J. S. (1993) Multifaceted conceptions of fit in structural equation models. In K. A. Bollen & J S. Long (Eds.) *Testing structural equation models* (pp. 1-39), Newberry Park, CA: Sage.

Tomás, J. M., & Oliver, A. (1999). Rosenberg's self-esteem scale: Two factors or method effects. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 84-98.

von der Embse, N., Kim, E. S., Kilgus, S., Dedrick, R., & Sanchez, A. (2019). Multi-informant universal screening: Evaluation of rater, item, and construct variance using a trifactor model. *Journal of School Psychology*, *77*, 52-66.

Wang, J., Siegal, H. A., Falck, R. S., & Carlson, R. G. (2001). Factorial structure of Rosenberg's self-esteem scale among crack-cocaine drug users. *Structural Equation Modeling: A Multidisciplinary Journal, 8*, 275–286.

Weems, G. H., Onwuegbuzie, A. J., Schreiber, J. B., & Eggers, S. J. (2003). Characteristics of respondents who respond differently to positively and negatively worded items on rating scales. *Assessment & Evaluation in Higher Education*, *28*(6), 587-606.

Ye, F., & Wallace, T. L. (2014). Psychological sense of school membership scale: method effects associated with negatively worded items. *Journal of Psychoeducational Assessment*, *32*(3), 202-215.

Author Correspondence:
Christine DiStefano
138 Wardlaw Hall
College of Education
Columbia, SC 29208
distefan@mailbox.sc.edu

Acknowledgement:
The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A150152 to the University of South Carolina. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education