PACE
Policy Analysis for California Education
CORE-PACE RESEARCH PARTNERSHIP

WORKING PAPER

An IRT Mixture Model for Rating Scale Confusion Associated with Negatively Worded Items in Measures of Social-Emotional Learning

We illustrate the application of mixture IRT models to evaluate the possibility of respondent confusion due to the negative wording of certain items on a social-emotional learning (SEL) assessment. Using actual student self-report ratings on four social-emotional learning scales collected from students in grades 3-12 from CORE districts in the state of California, we also evaluate the consequences of the potential confusion in biasing student- and school-level scores as well as correlational relationships between SEL and student-level variables. Models of both full and partial confusion are examined. Our results suggest that (1) rating scale confusion due to negatively-worded items does appear to be present; (2) the confusion is most prevalent at lower grade levels (3rd-5th); and (3) the occurrence of confusion is positively related to reading proficiency and ELL status, as anticipated, and bias estimates of SEL correlations with these student-level variables. For these reasons, we suggest future iterations of the SEL measures use only positively oriented items. To maintain measurement continuity, we suggest bias corrections based on the studied mixture model may be useful, although the precision of such corrections is sensitive to the nature of confusion (e.g., full versus partial).

Daniel M. Bolt
University of Wisconsin-Madison

Yang Caroline Wang
Robert H. Meyer
Libby Pier
Education Analytics

VERSION: October 2019

Acknowledgements
This paper was produced as part of the CORE-PACE Research Partnership, which is focused on producing research that informs continuous improvement in the CORE districts (Fresno, Garden Grove, Long Beach, Los Angeles, Oakland, Sacramento City, San Francisco, and Santa Ana unified school districts) and policy and practice in California and beyond. We thank the CORE Districts for partnering with us in this research, providing the data for this study, and giving feedback on earlier drafts of this study. PACE working papers are circulated for discussion and comment purposes and have not undergone the peer-review process that accompanies official PACE publications.

Introduction

Measurement of social-emotional learning (SEL) constructs has emerged as an important component of K–12 assessment. SEL measures increasingly play a role in school accountability, student progress monitoring, evaluation of post-secondary preparedness, and continuous improvement planning among networked improvement communities, among other uses (Durlak, Domitrovich, Weissberg, & Gullotta, 2015; Marsh et al., 2018). While alternative response formats are increasingly considered, the predominant format used in measuring such SEL constructs is the self-report rating scale format (West, Buckley, Krachman, & Bookman, 2018). It is widely known that such a format has various limitations. In particular, the validity of such measures can be undermined by factors associated with the idiosyncratic use of rating scales, including situations in which the negative wording of items may lead some students to use the rating scale in the reverse direction to what is intended. The consequences of such disorientation can impact not only the interpretations of individual respondent scores, but also assessments of the psychometric properties of the SEL measures, including evaluations of how the SEL measures correlate with other variables. When such disorientation is related to other student characteristics it also has the potential to significantly bias score reports at the student, school or district levels.

A significant amount of prior research has examined issues related to the use of reverse (i.e., negatively worded) items on self-report survey instruments (e.g., Barnette, 2000; Schriescheim, Eisenbach & Hill, 1991; Weijters & Baumgartner, 2012). Often the inclusion of such items is viewed as beneficial in addressing concerns related to acquiescent response style bias or social desirability (Roszkowski & Soven, 2010). Further, where it is apparent to the respondent that such items exist on a measure, a benefit of their inclusion is that it encourages the respondent to carefully read each item before responding. In this respect, reverse-worded items potentially provide protection against careless (or overly casual) responding to items. Others have noted potential benefits of negatively worded items in reducing the floor or ceiling effects sometimes seen with positively oriented items (Anastasi, 1982; Lin, Strong, Tsai & Lee, 2017; Nunnally, 1978). Indeed, for the assessment studied in this paper, this latter advantage was highlighted as a primary reason for the use of negatively worded items in one of the SEL scales (e.g., Dweck, 2000). Working against these potential benefits, however, are the psychometric complications that can emerge when inappropriate responses are given to the negatively worded items. The inclusion of negatively-worded items frequently reduces the reliability of scales and often yields an artificially complex factorial structure (Kam & Meyer,2015; Magazine, Williams & Williams, 1996; Meade & Craig, 2012; Schmitt & Stults, 1985; Sonderen, Sanderman, & Coyne, 2013; Woods, 2006). It is not uncommon to observe a separate statistical dimension emerge related to the directionality of items, dimensions that upon further examination are frequently interpreted as methodological artifacts related to incorrect interpretations and use of the rating scale.

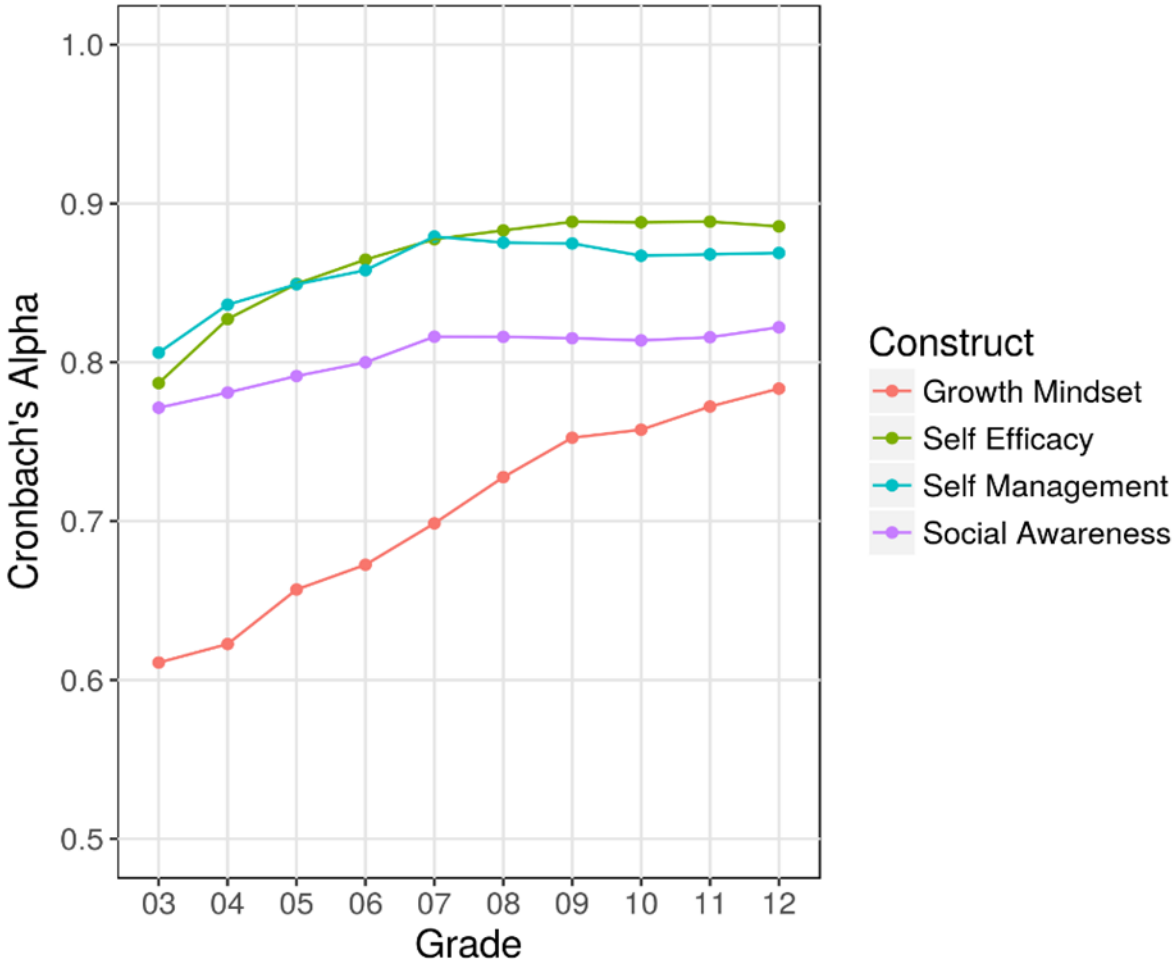
A general challenge in addressing the likely presence of rating scale confusion is that such confusion will likely exist only for a subsample of the respondents. Prior approaches have

emphasized factor analytic strategies for modeling such effects (e.g., Weijters, Baumgrtner, & Schillewaeert, 2013). However, given the qualititative nature of such effects, a mixture model provides an appealing alternative. Jin, Chen, & Wang (2017) considered an item response theory (IRT) mixture model for modeling inattentive response behavior, a related but distinct phenomenon. In this paper, we apply an IRT mixture model that uses latent classes to tease out students whose responses are affected by negatively worded items. We examine the application of the method using an assessment battery administered to students in grades 3–12 in the CORE districts – a consortium of eight California school districts[1] who collectively serve over one million students attending roughly 1,800 schools in the state. CORE districts are the first in the country to initiate a large-scale panel survey measuring students' social-emotional learning skills. Since the initial piloting in 2014, over 430,000 students have participated in SEL assessment each school year. Six districts[2], comprising around 436,000 students from over 1,100 schools, participated in the SEL survey in the 2014–15 school year. The assessment includes a total of 25 self-report items measuring four constructs: Self-Management (SM), Growth Mindset (GM), Self-Efficacy (SE), and Social Awareness (SA). All items are scored on a scale ranging from 1 to 5. Appendix A displays each of the items associated with the four SEL scales. Concern has been expressed over the poorer psychometric properties of the GM scale, especially at lower grade levels (e.g., reduced intercorrelations with other scales, poorer internal consistency). Figure 1 illustrates alpha coefficients of internal consistency observed across grade levels for each of the four SEL scales, where it is apparent that the internal consistency of GM items is considerably lower (especially at lower grade levels) than that of the other scales. As seen in Appendix A, a unique aspect of the GM scale is that its items have stems that consistently use negative wording such as "If I am not naturally smart in a subject, I will never do well in it." For such items, the failure to endorse items (e.g., selecting "Not at All True") is indicative of a high level of the GM construct, and thus the items are reverse scored in measuring the construct. By contrast, each of the other scales consist of items that are positively worded, implying endorsement (e.g., selecting "Almost All the Time") consistently reflects a positive orientation on the construct.

[1] The eight school districts are Fresno, Garden Grove, Long Beach, Los Angeles, Oakland, Sacramento City, San Francisco, and Santa Ana Unified School Districts.

[2] Two districts, Garden Grove Unified School District and Sacramento City Unified School District, did not participate in the SEL survey in 2014–15.

Figure 1. Cronbach Alpha Estimates by SEL Construct, Grades 3–12, CORE Districts



The same GM items in Appendix A have been used in other contexts and in turn to understand relationships between GM and other student variables. One finding of relevance to the current study is the tendency to observe lower levels of GM among students of both English language learner status and of lower academic achievement levels. In their report regarding students in the Clark County district in the state of Nevada, for example, Snipes and Tran (2017) noted significantly negative relationships between GM and prior academic achievement (as defined by a state math assessment), as well as significantly lower scores for English language learners (mean item score = 3.5) compared to non-English language learners (mean item score = 4.0).

The current application has several unique features relative to prior studies that have studied issues related to reverse- or negatively-worded items. First, in the current setting, one scale (i.e., GM) has all of its items negatively worded (as opposed to including items that are both positively and negatively worded), while the other three scales have entirely positively worded items. Second, there are very few GM items (i.e., 4) with which to evaluate the

presence of confusion. Each of these first two features makes it important that the existence of confusion in GM responses be evaluated in relation to the scores on non-GM scales. Third and finally, although the SEL constructs are measured at the student level, primary interest lies in the assessment of constructs at the school, and potentially district levels. As a result, we are particularly interested in any methodological tool that could be applied to address confusion in the assessment of SEL at the school level.

A unique aspect of the current assessment design was the decision to orient all items for one of the constructs (i.e., GM) in a negative direction. The psychometric capacity to detect rating scale confusion relies on the observation of GM responses that substantially defy the positive intercorrelations expected among all four SEL constructs. Table 1 illustrates the interscale correlations observed for the four scales across grade levels 3–12 for data collected in 2014–2015. The positive direction of the interscale correlations are as anticipated, given that all four constructs are likely underpinned by a higher-order factor. At the same time, the pattern of intercorrelations also makes apparent the weaker interscale correlations consistently observed for the GM scale. Such effects are likely related in part to the poorer reliability observed for the GM scale, but also go beyond what can be explained by effects of reliability, as even corrections for attenuation keep the GM intercorrelations well below those observed amongst the other scales. While such lower intercorrelations may occur for various reasons (including a GM construct that just happens to be more unique in relation to the other SEL constructs, or possibly general confusion as to what the GM items are asking), it is also consistent with the anticipated rating scale confusion among certain respondents.

Table 1. Interscale Correlations (Lower Triangle: Raw Correlations; Upper Triangle: Disattenuated Correlations), Grades 3–12, CORE District

Mean	SM	*GM*	SE	SA
Self-Management	1	*0.257*	0.512	0.637
Growth Mindset	*0.201*	1	*0.300*	*0.166*
Self-Efficacy	0.443	*0.237*	1	0.552
Social Awareness	0.531	*0.126*	0.465	1

Mean	SM	*GM*	SE	SA
Self-Management	1	*0.285*	0.520	0.587
Growth Mindset	*0.221*	1	*0.356*	*0.236*
Self-Efficacy	0.446	*0.279*	1	0.507
Social Awareness	0.486	*0.178*	0.420	1

In this paper, we use IRT mixture models to examine a theory that the poorer psychometric properties of the GM scale, especially at lower grade levels, might be attributed to rating scale confusion due to the negative wording of the GM items. We further seek to use the models to correct for bias due to such confusion by evaluating psychometric properties of the GM scale with respect to only an "unconfused" latent class. To this end, we develop a two-class multidimensional graded response model attending to item responses from all four scales. We consider two such mixture models: a full confusion model, in which classes are distinguished by a completely correct interpretation of the negatively worded items versus a complete reversal on the GM items, and a partial confusion model in which respondents of a confusion latent class are confused in their use of the rating scale on some GM items but not others.

Regardless of the model used, our analyses are designed to serve several purposes. First, by applying the models across grade levels 3–12, we anticipate findings to both support the existence of rating scale confusion among a subset of respondents and are suggestive of greater confusion at the lower grade levels relative to the higher grade levels. Observing higher proportions of students in lower grade levels within a confused class would support a theory that the poorer psychometric performance of the GM scale is due to cognitive confusion, and also question the validity of the GM scale for use at earlier grade levels. Second, through use of a Bayesian estimation procedure, we anticipate that application of the model should allow us to quantify bias in the psychometric properties of the GM scale by attending only to members of the "unconfused" class in quantifying those properties. We examine such effects by contrasting

the interscale correlations observed between the GM and other SEL scales both for the whole sample, and only the "unconfused" class, as identified through the mixture model. Third, we seek to document that some of the associations previously observed between the growth mindset construct and student characteristics (e.g., Snipes & Tran, 2017) are likely biased and possibly fully explained by the correlations of these same student characteristics with rating scale confusion. Fourth and finally, by studying the decomposition of class composition across schools, we can evaluate whether effects of rating scale confusion likely interfere with school-level assessments of the SEL constructs. This latter issue is especially relevant to the extent that school level SEL measures are a part of school accountability metrics and also inform decisions related to school-level SEL interventions.

Full Confusion Model

As noted above, both the full and partial confusion models are based on applications of a mixture version of a multidimensional graded response model (Samejima, 1969). Items within each SEL scale are modeled as unidimensional with a distinct unidimensional trait for each scale. Each item has one item discrimination and four boundary curve threshold parameters. The item response probabilities for the "unconfused" class (g = 1) can be written:

$$P\left(U_{ij} = k \middle| \boldsymbol{\theta}_i; a_j, b_{j1}, \ldots, b_{j4}; g = 1\right) = P^*_{ij,k-1}(\boldsymbol{\theta}_i) - P^*_{ij,k}(\boldsymbol{\theta}_i) \ ,$$

for score categories k=1, … 5, and $\boldsymbol{\theta}_i$ is a four-dimensional trait vector representing latent levels on the 4 SEL constructs for student *i*. Consistent with the graded response model of Samejima (1969), the $P^*_{ij,k}(\boldsymbol{\theta}_i)$ define boundary characteristic curves;
in this case $P^*_{ij0}(\boldsymbol{\theta}_i) = 1, P^*_{ij5}(\boldsymbol{\theta}_i) = 0$, and

$$P^*_{ij,k}(\boldsymbol{\theta}_i) = \frac{\exp\left[a_j\left(\theta_{i,s(j)} - b_{jk}\right)\right]}{1 + \exp\left[a_j\left(\theta_{i,s(j)} - b_{jk}\right)\right]}$$

for k = 1, ... 4, where s indexes the scale to which item j belongs. Note that each item measures just the one latent trait corresponding to the scale it represents; the four latent traits corresponding to the four scales are assumed to correlate. We fix the a and b parameters at estimates observed when a single class multidimensional GRM is applied to the 12th grade students, as the rating scale confusion is speculated to be minimal at this grade level. Similarly, the correlation matrix among the latent traits is set at estimates observed in the 12th grade analysis.

For the "confused" class (g = 2) we assume the same model as for class 1 holds for the items of all scales except the GM scale, for which the item response probabilities are exactly reversed. Specifically,

$$P\left(U_{ij} = k \middle| \boldsymbol{\theta}_i; a_j, b_{j1}, \ldots, b_{j4}; g = 2\right) = P\left(U_{ij} = 5 - k \middle| \boldsymbol{\theta}_i; a_j, b_{j1}, \ldots, b_{j4}; g = 1\right)$$

for all GM items, but

$$P(U_{ij} = k|\boldsymbol{\theta}_i; a_j, b_{j1}, \ldots, b_{j4}; g = 2) = P(U_{ij} = k|\boldsymbol{\theta}_i; a_j, b_{j1}, \ldots, b_{j4}; g = 1)$$

for the three other scales.

Such constraints imply that examinees in the confused class provide psychometrically equivalent responses to the GM items except for use of the rating scale in exact reverse direction. That is a rating of '1' actually reflects a '5', a '2' reflects a '4', and so on.

Partial Confusion Model

The partial confusion model takes the same structure as the full confusion model, but with the distinction that the confused class will demonstrate confusion with respect to half (2/4) of the GM items. Specifically,

$$P(U_{ij} = k|\boldsymbol{\theta}_i; a_j, b_{j1}, \ldots, b_{j4}; g = 2) = .5 * P(U_{ij} = 5 - k|\boldsymbol{\theta}_i; a_j, b_{j1}, \ldots, b_{j4}; g = 1) + .5 * P(U_{ij} = k|\boldsymbol{\theta}_i; a_j, b_{j1}, \ldots, b_{j4}; g = 1)$$

for all GM items, and again that

$$P(U_{ij} = k|\boldsymbol{\theta}_i; a_j, b_{j1}, \ldots, b_{j4}; g = 2) = (U_{ij} = k|\boldsymbol{\theta}_i; a_j, b_{j1}, \ldots, b_{j4}; g = 1)$$

for the three other scales. The use of a partial confusion model addresses a likelihood that many respondents may only demonstrate rating scale disorientation on a subset of GM items, and thus the full confusion model may underestimate the actual proportion of confused respondents. We anticipate that the partial confusion model will lead to a higher estimated proportion in the "confused" class than the full confusion model, and thus a better correction for bias if, in fact, partial confusion provides a more accurate characterization of the nature of confusion produced by the negatively worded items.

Model Estimation

Each of the full and partial confusion models is fit using WINBUGS 1.4 (Spiegelhalter, Thomas, & Best, 2003) with priors for the mixture proportions specified as $\pi = (\pi_1, \pi_2)$ ~ Dirichlet (.1,.1). At the individual student level, we assume class membership parameters where the probability of membership in a confusion class is g~Bernoulli(π_2), and a 4-dimensional trait parameter $\boldsymbol{\theta}_i$~MultNormal(0,Σ), where Σ is defined by covariance estimates observed for a single-class multidimensional graded response model applied to the SEL measures at the 12th grade. Specification of such priors leads to application of an adaptive rejection sampling algorithm in WINBUGS 1.4. As our primary purpose in this paper is exploration of the methodology, we applied the models to a random sample of 5000 students at each grade level, including only student response patterns that had no missing responses. The Markov chains

were simulated out to 5000 iterations, and convergence was monitored using the Gelman-Rubin (1992) criterion. Importantly, at each stage of the simulated Markov chain, we observe a partitioning of the sample into unconfused and confused classes. This partitioning allows us estimate interscale correlations between scale scores conditional upon membership in the unconfused class. When these interscale correlations are averaged across iterations, we obtain estimates of the interscale correlations that are sensitive to the relative likelihoods of individual students being in the unconfused versus confused classes. We use these estimated interscale correlations for the unconfused as bias-corrected estimates, as described below.

Simulation Analyses

To evaluate the performance of the models and corresponding analysis procedures, we also conducted simulation analyses. In these analyses, we simulated data from each of the partial confusion and full confusion models. We in turn applied each of the two models (full confusion, partial confusion) to each dataset, evaluating the recovery of both the true mixing proportions and the interscale correlations across the four subscales. As for the real data, each of the generated datasets involved 5000 respondents and 25 items, with a four-dimensional structure and item parameter estimates identical to the estimates observed for the real data (see below). The purpose of the preliminary simulation analyses was (1) to confirm that the proposed models and analytic procedure can in fact recover both the true mixing proportions and interscale correlations when the data conform to the model being specified; and (2) to give a preliminary indication of how the models perform in the presence of misspecification (i.e., assuming full confusion when only partial confusion is present; or assuming partial confusion when full confusion is present). For data generated under the full confusion conditions, we used generating mixing proportions of .8 and .2 for the unconfused and confused classes, respectively. For the partial confusion condition, the corresponding mixing proportions were .7 and .3.

As with the real data analyses described below, we fixed the item parameter values, in this case at the generating parameter values used for the simulation. We estimate the mixing proportions for the two classes, as well as the class membership parameters and the four latent trait parameters for each respondent.

Table 2 reports results with respect to both the mixing proportion estimates and interscale correlation estimates for the simulation analyses conducted. Tables 2(a) and (b) illustrate results for data generated according the full confusion model, while Table 2(c) and (d) for the partial confusion data. The three entries in each cell of the table show the (1) true correlations among scale scores (as defined by students generated to be in the unconfused class), (2) the corresponding estimated correlations between scale scores in the unconfused class based on application of the mixture model, and (3) the corresponding correlations when estimated from the generating data across both classes (thus containing bias due to rating scale confusion). Bias is thus observed in comparing the first and third entries in each cell; the correction for bias is seen by comparing the second entry against the first and third. The closer

the second entry is brought to the first and away from the third, the greater the correction for bias.

Table 2. Simulation Analyses Based on Full/Partial Confusion Models

(a) Simulation 1 True/Estimated(UnConfused)/Est(Both) Interscale Correlations, Full Confusion Generated, Full Confusion Estimated

Scale	1	2	3	4
1	1			
2	.166/.161/.069	1		
3	.391/.386/.387	.309/.311/.163	1	
4	.503/.507/.500	.172/.167/.069	.349/.350/.354	1

Note: True mixing proportions: .811, .189; Estimated mixing proportions: .809, .191

(b) Simulation 2 True/Estimated(UnConfused)/Est(Both) Interscale Correlations, Full Confusion Generated, Partial Confusion Estimated

Scale	1	2	3	4
1	1			
2	.166/.133/.069	1		
3	.391/.383/.387	.309/.275/.163	1	
4	.503/.499/.500	.172/.138/.069	.349/.349/.354	1

Note: True mixing proportions: .811, .189; Estimated mixing proportions: .854, .146

(c) Simulation 3 True/Estimated(UnConfused)/Est(Both) Interscale Correlations, Partial Confusion Generated, Partial Confusion Estimated

Scale	1	2	3	4
1	1			
2	.210/.210/.142	1		
3	.368/.365/.370	.337/.339/.233	1	
4	.502/.498/.497	.216/.213/.148	.362/.361/.354	1

Note: True mixing proportions: .698, .302; Estimated mixing proportions: .706, .294

Table 2. Simulation Analyses Based on Full/Partial Confusion Models (concluded)

(d) Simulation 4 True/Estimated(UnConfused)/Est(Both) Interscale Correlations, Partial Confusion Generated, Full Confusion Estimated

Scale	1	2	3	4
1	1			
2	.210/.175/.142	1		
3	.368/.368/.370	.337/.289/.233	1	
4	.502/.498/.497	.216/.178/.148	.362/.356/.354	1

Note: True mixing proportions: .698, .302; Estimated mixing proportions: .892, .108

With this interpretation in mind, it is seen from each of the tables that the primary bias occurs for the intercorrelations involving the second scale (corresponding to GM). This is as expected, as it was only for the second scale where confusion was simulated. In Tables 2(a) and (c), it is further seen that whether full or partial confusion is simulated, application of the correct model yields mixture proportion estimates and corrected interscale correlation estimates that appear largely unbiased. Such results suggest that where the nature of confusion can be correctly defined (full versus partial), the method performs well, and yields estimates that are accurate. However, misspecification of the model (i.e. specifying partial confusion in the presence of full confusion in Table 2(b), or full confusion in the presence of partial confusion in Table 2(d)) results in a reduced correction. Regardless of the direction of misspecification, the estimated proportion in the confusion class is underestimated; consequently, it is also seen that the interscale correlation estimates, while improved, do not completely remediate the bias. In both cases, we see the interscale correlation estimates involving the second scale (GM) are increased, but not to the level corresponding to the true generating values.

As correct specification of the model (and in particular, the nature of confusion) appears important to the success of the method, we can also examine the degree to which application of the correct model can be statistically determined. As the full and partial confusion models share the same number of parameters, we compared directly the mean log-likelihood observed for each model when fit to each dataset. In addition, we attend here to results observed for a third dataset in which no respondents were members of the confusion class (mixing proportions of 1 of 0 for the unconfused and confused classes, respectively). This no confusion model can naturally be viewed as a special case of both the partial and full confusion models (In each case where the proportion in the confused class is 0). This third dataset, as well as the fitting of a model without a confused class, provides opportunity to examine whether the model can correctly identify a confused class.

Table 3 provides a comparison of the mean log-likelihoods across datasets and fitted models. For each of the partial and fully confusion datasets, we find the correctly corresponding model to yield the highest mean log-likelihood, and in both cases also a higher log-likelihood than the no confusion model. For the no confusion data, we find an equivalent

mean log-likelihood for the partial confusion class, and a lower log-likelihood for the full confusion class. Although there is clearly more that can be explored here (including the need for additional replications) taken together the results suggest a potential to learn both about the presence and nature of rating scale confusion through a statistical comparison of models that make different assumptions about the nature of confusion.

Table 3. Mean Log-Likelihood (95% Interval), Full/Partial/No Confusion Data Estimated Under Full/Partial/No Confusion Models, Simulation Analyses

Generating\Fitted	Full	Partial	None
Full	**-120100 (-120300, -119900)**	-121100 (-121300, -120900)	-120600 (-120800, -120400)
Partial	-124400 (-124600, -124200)	**-122600 (-122800, -122400)**	-124800 (-12500, -124600)
None	-122600 (-122800, -122400)	**-120600 (-120800, -120400)**	**-120600 (-120800, -120400)**

Note: **BOLD** identifies fitted models with highest mean log-likelihood for generating condition (rows).

In summary, it would appear that application of the mixture model generally yields more accurate interscale correlation estimates, but that the degree of improvement is sensitive to correctly capturing the nature of confusion (full versus partial) represented in the data. While it appears that it is possible to statistically distinguish between different models on confusion in terms of their relative statistical fit, there are naturally still other models (beyond the full and partial confusion models considered here). We consider the implications of such findings further in discussion.

Real Data Analyses

In the past few years, researchers in the CORE-PACE[3] Research Partnership have undertaken several studies involving CORE's SEL survey and launched a series of working papers summarizing the research findings to date (PACE, 2018). One working paper (Meyer, Wang, & Rice, 2018) examined the measurement properties of CORE's SEL survey items using unidimensional IRT models. The authors compared three polytomous IRT models (i.e., partial credit model, generalized partial credit model, and nominal response model) for each of the four SEL construct at each grade and found generally better model-data fit with a more general

[3] PACE – Policy Analysis for California Education – is an independent, nonpartisan research center led by faculty directors at Stanford University, the University of Southern California, the University of California Davis, the University of California Los Angeles, and the University of California Berkeley.

model and higher grades. To better understand the degree to which the meanings of SEL items remain the same across different grade levels and student demographic groups, these researchers also conducted differential item functioning (DIF) analyses. They identified a few items exhibiting "moderate to large" DIF between gender and race/ethnicity subgroups and most items exhibiting "moderate to large" DIF across grades, especially among three grade bands – grades 3–6, 7–9, and 10–12.

In another research paper (Bolt, Wang, Meyer, & Pier, 2019), the authors applied two-level bifactor, second-order factor, and correlated factor models to CORE's SEL data to evaluate measurement invariance among the four SEL constructs at both the school and student levels. Their results suggested that a multilevel bifactor approach provided a superior comparative fit to the data. Bifactor analysis results further revealed that the four SEL constructs generally appeared more differentiated at the student than at the school levels, especially at later grades and for growth mindset and self-efficacy.

In this paper, we examine data from students who participated the first operational SEL survey administration in the 2014–15 school year. IRT mixture model analyses results reported in this section are based on a random sample of 5,000 students from each grade level between grades 3 and 12 from the six districts included in this paper (i.e., a total of 50,000 students across grades) who completed all SEL items on the survey.

Table 4 reports the estimated mixture proportions across each grade from 3–12 for the SEL data. For the full confusion model, consistent with our theory, there appears to be an increasing proportion of students in the "confused" class as grade becomes lower. The maximum proportion in the confused class is .13, which occurs in Grade 3, while the minimum proportion (.02) occurs in Grade 12.

Table 4. Estimated Latent Class Proportions for Confused and Unconfused Classes, CORE SEL data Grades 3–12, Full Confusion Model (Sample of 5000 students per grade level)

Grade	P(Confused)	P(Unconfused)
3	.13	.87
4	.10	.90
5	.05	.95
6	.03	.97
7	.04	.96
8	.03	.97
9	.03	.97
10	.02	.98
11	.02	.98
12	.02	.98

Table 5 shows the inter-scale correlations between the GM scale and the other scales when estimated within the unconfused class only (leftmost columns) as compared to when estimated for all students. As anticipated, the correlations consistently increase (as expected) when evaluated only with respect to the unconfused class. Despite this increase, however, the interscale correlations involving the GM scale still appear consistently below those seen among the other three scales as shown in Table 1.

Table 5. Estimated Correlations Between Growth Mindset Scores and Other SEL Scales, Overall and Only Unconfused Class, CORE SEL data Grades 3–12, Full Confusion Model (Sample of 5000 students per grade level)

	Only Unconfused Class			All Students		
Grade	Self-Management	Self-Efficacy	Social Awareness	Self-Management	Self-Efficacy	Social Awareness
3	.19	.25	.15	.11	.14	.04
4	.20	.27	.18	.14	.19	.10
5	.24	.35	.21	.21	.28	.13
6	.23	.31	.17	.21	.28	.13
7	.23	.34	.19	.20	.28	.15
8	.25	.38	.22	.21	.34	.18
9	.20	.35	.16	.17	.30	.12
10	.24	.34	.22	.22	.31	.18
11	.20	.36	.22	.17	.30	.18
12	.20	.32	.20	.16	.26	.16

Tables 6 and 7 show the corresponding results for the partial confusion model analyses, again applied to each grade level. As seen in Table 6, a similar pattern to that observed for the full confusion model emerges, but with considerably higher proportions in the confused class. The estimated proportion in the confused class is as high as .53 (Grade 3) and drops to .05 at higher grades (Grades 11 and 12). It would thus appear that the application of the partial confusion substantially increases the proportion of students identified as confused.

Table 6. Estimated Latent Class Proportions for Confused and Unconfused Classes, CORE SEL data Grades 3–12, Partial Confusion Model (Sample of 5000 students per grade level).

Grade	P(Confused)	P(Unconfused)
3	.53	.47
4	.46	.54
5	.32	.68
6	.26	.74
7	.24	.76
8	.15	.85
9	.10	.90
10	.07	.93
11	.05	.95
12	.05	.95

Table 7. Estimated Correlations Between Growth Mindset Scores and Other SEL Scales, Overall and Only Unconfused Class, CORE SEL data Grades 3–12, Partial Confusion Model (Sample of 5000 students per grade level)

	Only Unconfused Class			All Students		
Grade	Self-Management	Self-Efficacy	Social Awareness	Self-Management	Self-Efficacy	Social Awareness
3	.24	.33	.22	.11	.14	.04
4	.25	.35	.25	.14	.19	.10
5	.28	.41	.26	.21	.28	.13
6	.26	.38	.22	.21	.28	.13
7	.26	.39	.22	.20	.28	.15
8	.26	.40	.23	.21	.34	.18
9	.21	.36	.39	.17	.30	.12
10	.24	.36	.36	.22	.31	.18
11	.20	.26	.34	.17	.30	.18
12	.20	.32	.36	.11	.14	.04

Thus, our real data findings appear quite consistent with our observations from the simulation analyses across grade levels. Naturally an important consideration is which of the two models appears most consistent with the data. As Grade 3 appears most affected by confusion, we focus here on the results observed for Grade 3. Following the same approach as with the simulation, at the Grade 3 level, we estimated a mean log-likelihood under each of the partial confusion and full confusion models. We observed a higher log-likelihood for the partial confusion model (mean = -145800, 95% interval of [-146000, -145600]) than the full confusion model (mean = -148200, 95% interval of [-148400, -148000]). The superiority of the partial confusion model is consistent with the poorer internal consistency seen for the GM subscale, which would suggest more contradictory responses among the GM items as is implied by the partial confusion model.

To further validate our application of the partial confusion mixture model with the real data, as well as to better understand its potential implications for bias, we examined associations between class membership and other student characteristics. Specifically, we considered correlations with a student's Smarter Balance Assessment Consortium (SBAC) English language arts/literacy (ELA/literacy) score, English Language Learner (ELL) status, gender, special education status, and race (Caucasian, African American, Asian, and Hispanic) using the full confusion analysis. Table 8 illustrates correlations observed between the posterior probability of membership in the confusion class and each of the SBAC ELA/literacy score and ELL status variables by grade level. For each correlation, we also report a merge rate proportion, which reflects the proportion of 5000 students in the confusion analysis for which the SBAC score or ELL status variables were available. We observe consistently negative correlations between SBAC ELA/literacy and confusion, and consistently positive correlations between ELL status and confusion, across grades. Such effects are consistent with theoretical expectations to the extent that we anticipate lower levels of reading and/or language

proficiency to yield a higher likelihood of confusion. Figure 2 provides an illustration of the relationship between SBAC achievement score and posterior probability of membership in the confusion class (weight) for the Grade 3 sample. Higher confusion class weight values imply a higher posterior probability of membership in the confusion class. The average relationship between SBAC ELA/literacy and confusion is shown by a kernel-smooth regression curve, showing students with lower SBAC ELA/literacy scores having a higher likelihood of membership in the confused class.

Table 8. Estimated Correlations between Confusion Class Membership and English Language Learner (ELL) Status and SBAC ELA/Literacy Scores (Sample of 5000 students per grade level), CORE SEL Data, Partial Confusion Model

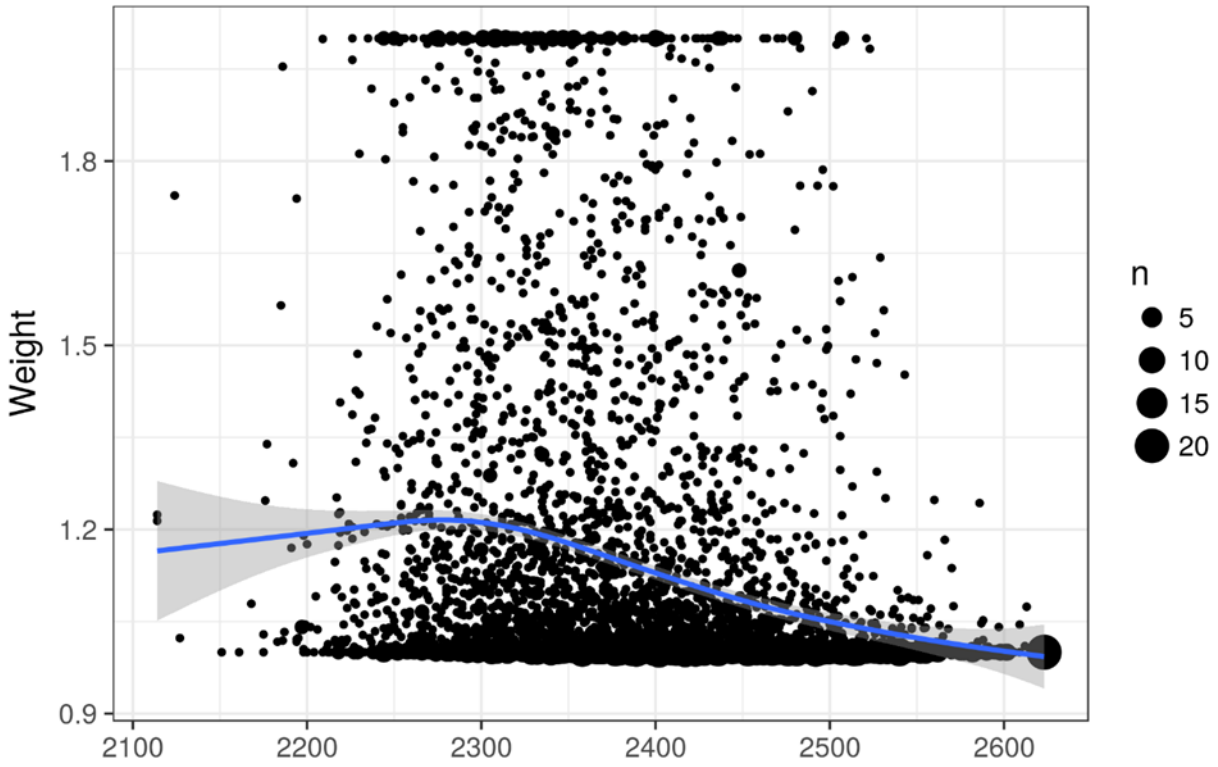
(a) SBAC ELA/Literacy

	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 11
Corr.	-.25	-.24	-.21	-.19	-.20	-.17	-.11
Merge Rate	97%	97%	98%	96%	93%	96%	90%

(b) ELL Status

	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 9	Grade 10	Grade 11
Corr.	0.18	0.18	0.19	0.17	0.13	0.13	0.17	0.13	0.13
Merge Rate	99.9%	99.9%	99.9%	99.9%	100%	100%	100%	100%	100%

Figure 2. Illustration of Relationship Between SBAC ELA/Literacy Scores and Probability of Membership in Confusion Class (Weight), Grade 3, CORE SEL Data (Sample of 5000 students).



We also performed multiple regression analyses by grade predicting the posterior probability of membership in the confusion class as a function of each of the student characteristics listed above. Tables 9 and 10 display the resulting regression coefficients and summary R-squared measures from these analyses. The results again display a consistent pattern across grade levels, but with somewhat stronger effects emerging at earlier grade levels. The strongest effects appear present for the SBAC and ELL variables. Despite these detectable effects, the overall prediction is relatively modest, as evidenced from the lower R-squared values in Table 10. Nevertheless, the relationships demonstrate the potential for confusion to render bias that might relate to student characteristics, and thus to school bias to the extent that such student characteristics vary across schools.

Table 9. Multiple Regression Coefficient Estimates Predicting Probability of Membership in Confusion Class as Related to SBAC ELA/Literacy Scores and Demographic Variables, CORE SEL Data (Sample of 5000 students per Grade Level), Partial Confusion Model

Grade	SBAC ELA/ Literacy	ELL	Female	SPED	African American	Hispanic	Asian
3	-.0006*	.032*	.009	-.013	.015	.021	.005
4	-.0005*	.023*	.018*	.006	.002	.006	-.004
5	-.0003*	.023*	.001	.013	.023	.004	-.008
6	-.0002*	.028*	.004	.010	.004	.001	-.003
7	-.0002*	.005	.004	.034*	.008	.007	.008
8	-.0002*	.007	.001	.001	-.012	-.000	.005
11	-.0001*	-.007	.002	.005	-.002	.001	.019
All	-.0004*	.022*	.007*	-.003	.002	.002	.003

Note: *$p < .01$

Table 10. Multiple Regression R-Square Estimates, CORE SEL Data (Sample of 5000 students per Grade Level), Partial Confusion Model

Grade	R-squared	df
3	.066	4836
4	.060	4831
5	.050	4884
6	.046	4760
7	.046	4618
8	.029	4742
11	.015	4282
All	.079	33048

As noted earlier, prior work (e.g., Snipes & Tran, 2017) has found GM to be notably lower among ELL students, as well as students with lower achievement. The above relationships of these same variables with confusion raise the prospect that such estimates may be biased due to the presence of confusion. Table 11 shows the zero-order correlations between GM and ELL status as well as the SBAC ELA/literacy score. These results are consistent with Snipes & Tran (2017) suggesting that students of ELL status and lower achievement show lower GM. However, when controlling for confusion, we see notable decreases in these associations, particularly for ELL status at the earlier grade levels. In particular, when attending to confusion, it appears that the association between ELL status and GM is not present at the early grade levels, and only emerges later in child development. In addition, even at higher grade levels, the statistical relationship between ELL status and GM appears weaker when accounting for confusion.

Table 11. Estimated Correlations Between GM and Student Variables of ELL Status and SBAC ELA/Literacy Scores Before and After Controlling for Confusion, Partial Confusion Model

Grade	Corr(GM,ELL)	Corr(GM,ELL), controlling confusion	Corr(GM,SBAC)	Corr(GM,SBAC), controlling confusion
3	-.17*	.01	.29*	.17*
4	-.18*	.00	.32*	.22*
5	-.21*	-.06*	.35*	.29*
6	-.17*	-.03*	.35*	.30*
7	-.16*	-.07*	.38*	.33*
8	-.14*	-.04*	.38*	.35*
9	-.20*	-.09*	NA	NA
10	-.18*	-.10*	NA	NA
11	-.16*	-.09*	.29*	.27*
12	-.18*	-.11*	NA	NA

Note: *$p < .01$

Finally, it might be speculated that application of the mixture model can be used to address the effects of confusion. Specifically, we can estimate at the respondent level a corresponding θ; under the MCMC approach described above, we use the mean sampled θ across 5000 iterations as a respondent level θ estimate, which can then be converted to a true score on the GM scale using the GRM item parameter estimates for the unconfused class from the mixture analysis. Figure 3 provides a scatterplot illustrating the original GM scores and their resulting corrections for the Grade 3 analysis at both (a) the student level and (b) the school level. It is clear from the figure that the amounts of bias appear more substantial at the student compared to school levels. Naturally, students confused in their use of the rating scale for GM items will often have dramatically changed scores. In actual practice, where scores might be the basis for interventions, it will likely prove beneficial to attend to the likelihood of student confusion in making student-level intervention decisions. While the school-level corrections seen in Figure 3(b) may seem less substantial than at the student level (due to the stronger positive relationship between uncorrected and corrected GM scores), the differences shown are in many cases still quite substantial. It is not uncommon to see pairs of schools reporting the same uncorrected GM score differing by as much as .4 units or more after making the adjustment for confusion. As an illustration of the magnitude of such adjustments, we consider two schools for which the 3rd grade students show a large difference in the proportion of students in the confused class. School 5, which also showed a higher prior proportion of ELL students, was observed to have 35% in the confusion class among its 3rd graders; for school 6 the proportion in the confused class was only 2%. Following the same procedure used to remove bias in the interscale correlations, we can similarly estimate a mean GM score for respondents in the unconfused class, yielding a school-level GM score corrected for bias. Figure 4 thus displays not only the observed mean GM scores for Schools 5 and 6, but also the bias corrected scores once accounting for the presence of confusion. While the original GM

scores would appear to suggest that school 5 is quite low in GM, when corrected for confusion, its GM score becomes much higher, and comparable that of school 6. Thus it would seem that the implications of the methodology for school-level estimates of GM could appear meaningful. We intend to explore more carefully the school-level consequences of the confusion-adjustments in future analyses.

Figure 3. Illustration of Effects of Student and School Level Correction of GM Scores Due to Confusion, CORE SEL Data, Grade 3

(a) Scatterplot of Original and Corrected Student GM scores

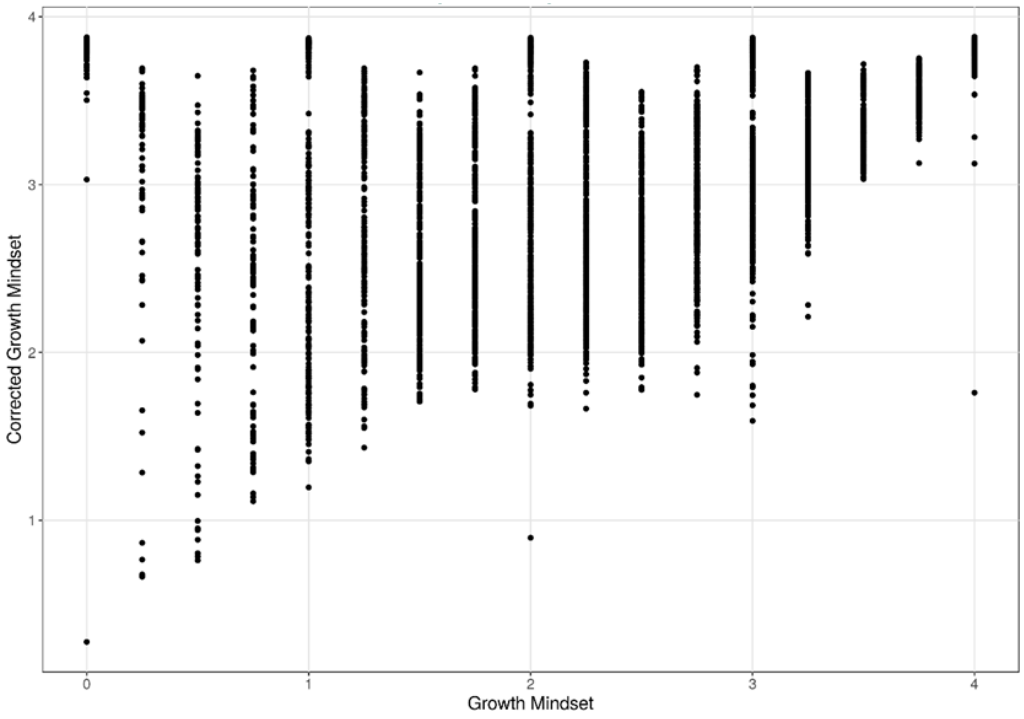


Figure 3. Illustration of Effects of Student and School Level Correction of GM Scores Due to Confusion, CORE SEL Data, Grade 3 (continued)

(b) Scatterplot of Original and Corrected School-Level GM Scores

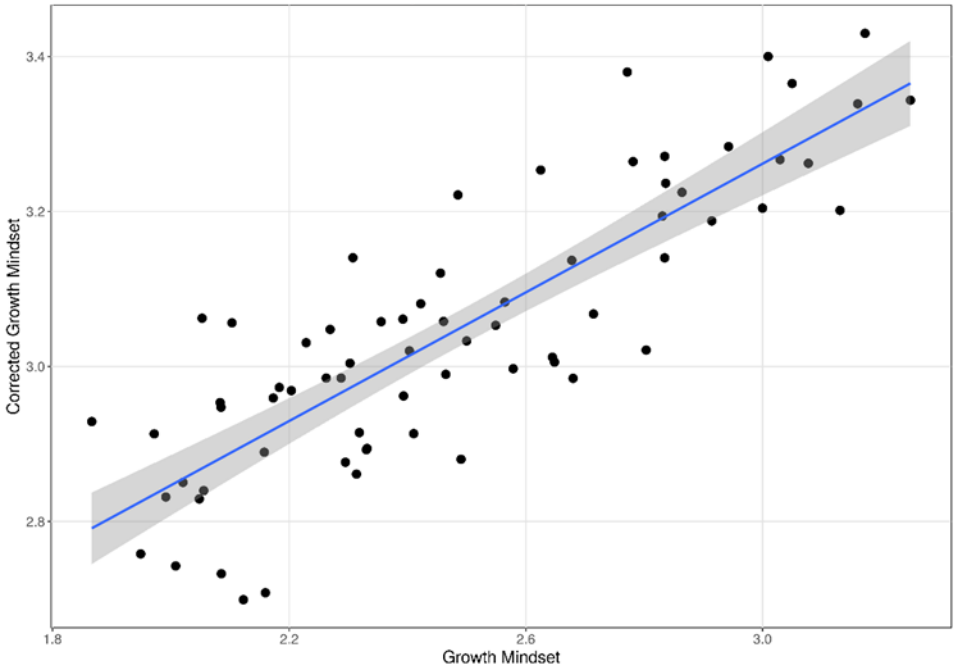
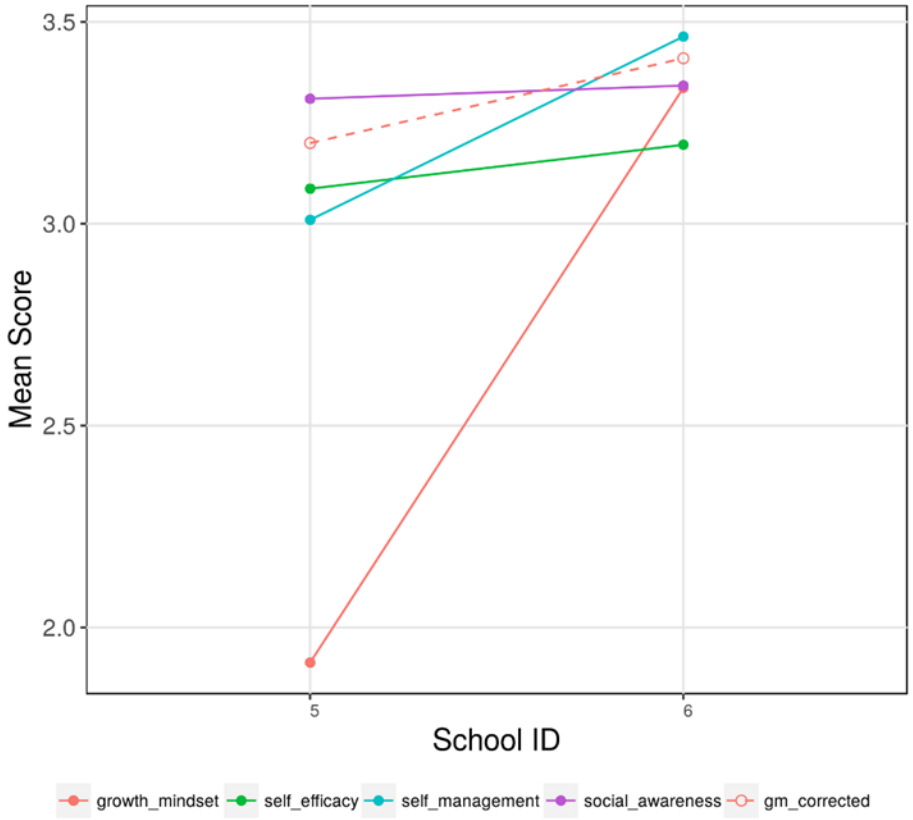


Figure 4. Illustration of Effects of School Level Correction of GM Scores Due to Confusion, Two Schools (5&6), CORE SEL Data, Grade 3

School ID	Grade	# of students	% ELL	% Confusion	Mean (SD) of SBAC ELA/Literacy	Mean of SEL Constructs				
						SM	GM	Corrected GM	SE	SA
5	3	23	43.5%	35%	2358 (65)	3.01	1.91	3.20	3.09	3.31
6	3	23	4.3%	2%	2492 (65)	3.46	3.34	3.41	3.20	3.34



Practical Implications and Conclusion

Our study has several practical implications. The first is that it lends support to a theory that the poorer psychometric performance of the GM items at lower grade levels can be attributed to confusion associated with the use of negative wording of GM items. We find in both the full and partial confusion models that the proportions in the confused class are higher at lower grade levels, with as many as 13% showing full confusion and as many as 53% showing partial confusion at the 3rd grade level, as compared to 2% and 5%, respectively, at the 12th grade level.

A second implication concerns the implications of rating scale confusion on our understanding of the construct of GM and its relation to student characteristics. In particular, while prior work has suggested considerably lower GM among ELL students, we show that such differences are significantly exaggerated due to the higher degree of rating scale confusion among ELL students. Indeed, based on the CORE SEL data, such differences are not present at all at earlier grades, and only seem to emerge to a modest extent at higher grade levels.

A third implication is that the method allows for analysis of existing GM data by conditioning psychometric evaluation on the latent classes of students who appear not to have been confused. These later results not only permit a better assessment of the amount of bias in student/school level estimates due to rating scale confusion, but could also provide a basis for corrections (i.e., focusing on the results for the unconfused class), if the GM items were to be subsequently reworded in future iterations but measurement continuity is desired.

The current psychometric study also allows us to evaluate whether the lower correlations between the GM scale and other SEL scales appears to be entirely a consequence of confusion over the negatively worded items. In this respect, it appears that while the correlations of the GM scale are suppressed in part by confusion over the negative phrasing, the correlations remain low even after applying the mixture model to address bias due to confusion. Such findings suggest that the GM construct may, relatively speaking, reflect a rather unique aspect of SEL. Of course, it is also possible that this uniqueness in part reflects the fact that the construct itself remains oriented in a reverse direction to that of the other constructs. Specifically, use of the negative phrasing suggests the measure as one of “fixed mindset”, in contrast to a measure of positive wording that would reflect “growth mindset” (Dweck, 2006). In this regard, we note that since we shared our research findings on “fixed mindset” items with the CORE districts, CORE has piloted positively phrased GM items via an innovation zone initiative and decided to replace the original GM items with positively phrased GM items and make continuous improvement on its SEL survey. In the future, we will report research findings from examining students’ responses on both positively and negatively phrased GM items which allow us to better evaluate such effects.

One other practical implication of this research involves practitioners’ uses of these “fixed mindset” items. With CORE districts recognizing the issues with negatively phrased GM

items and replacing them with positively phrased GM items in its operational SEL survey, other practitioners are still using the same negatively phrased GM items for young students, in particular. It is quite common to see statements in published papers and reports such as “ELLs reported a significantly lower level of GM compared to non-ELLs” and “lower achieving students reported a significantly lower level of GM than their peers”. We feel the need to share our findings with policy makers, educators, parents, and researchers in the field of SEL measurement so that student SEL scores are properly interpreted and the most appropriate practices and interventions are used to help students grow.

There remain several limitations to the methodology applied in this paper, some inherent to the methodology and others specific to the design in which methodology was applied. First, our method assumes measurement invariance across grades. Specifically, we assume both that the relationships between the latent traits and item scores, as well as the correlations between latent traits, remain consistent across grades. Analyses conducted in parallel to those reported in this paper suggest such measurement invariance assumptions may be reasonable, although naturally don’t convincingly rule out the possibility of grade-level measurement differences beyond those attributed to the negative phrasing of GM items. Second, as noted earlier, our analyses make certain assumptions about the nature of confusion (i.e., full versus partial) that are difficult to convincingly confirm. There are unfortunately statistical limitations to what can be done in this regard. Specifically, trying to define classes that conform to all possible forms of confusion that may emerge (e.g., different classes specific to the particular items on which students were confused) is not practically feasible. Third, the design of the current SEL survey, with only four GM items, provides limited information by which to evaluate the presence of confusion. Fortunately, the results are not fully dependent on reliability at the individual student level, which is naturally low. However, it does raise the possibility that peculiarities in the functioning of individual items could interfere with the performance of the model.

While the proposed methodology offers a way of attempting to rectify a likely source of interference in measurement, methods such as the mixture model proposed are not a panacea for measurement problems such as that observed in the CORE SEL data. As a result, the primary practical recommendation to follow from this work is that the GM items be rewritten with a positive orientation. Our mixture analysis confirms the presence of student confusion, especially at earlier grade levels, and demonstrates the presence of student-level and school-level biases in GM measurement. Such bias is systematically confounded with student level variables found to correlate with confusion, specifically, ELL status and ELA achievement. Bias corrections using an IRT mixture model are possible, and although imprecise, may nevertheless provide a mechanism to preserve continuity despite a transition to positively worded GM items.

Finally, to the extent that mixture modeling provides an increasingly easy-to-implement tool in psychometric analysis, there will likely be value in exploring additional applications beyond the effects of the negative wording considered in this paper. While traditional measurement applications have often focused on measurement differences in relation to

manifest student characteristics, mixture models emphasize latent student variables that may impact how measurement instruments function (Cohen & Bolt, 2005). Thus, it can become an easily adapted and exploratory tool for studying other measurement artifacts that disproportionately affect different student subpopulations.

References

Anastasi, A. (1982). *Psychological Testing*. 5th ed. New York, NY: Macmillan.

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.

Bolt, D., Wang, Y. C., Meyer, R. H., & Pier, L. (2019). *Evaluating the differentiation of social-emotional learning (SEL) constructs using multilevel factor analysis*. Paper presented at the NCME annual conference, Toronto, Canada.

Cohen, A. S., & Bolt, D. M. (2005). A mixture model analysis of differential item functioning. *Journal of Educational Measurement*, *42*(2), 133–148.

Durlak, J. A., Domitrovich, C. E., Weissberg, R. P., & Gullotta, T. P. (Eds.). (2015). *Handbook of social and emotional learning: Research and practice*. New York: The Guilford Press.

Dweck, C. S. (2000). *Self-theories: Their role in motivation, personality, and development*. Philadelphia, PA: Taylor and Francis.

Dweck, C. S. (2006). *Mindset: The new psychology of success*. New York, NY, US: Random House.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science, 11*, 457–472. https://doi.org/10.1214/ss/1177011136

Hughes, D. (2009). The impact of incorrect responses to reverse-coded survey items. *Research in the Schools, 16*(2), 76–88.

Jin, K-Y., Chen, H-F., Wang, W-C. (2017). Mixture item response models for inattentive responding behavior**.** *Organizational Research Methods, 21*(1), 197–225. https://doi.org/10.1177/1094428117725792

Józsa, K., & Morgan, G. A. (2017). Reversed items in Likert scales: Filtering out invalid responders. *Journal of Psychological and Educational Research, 25*(1), 7–25.

Kam, C. C. S., & Meyer, J. P. (2015). How careless responding and acquiescence response bias can influence construct dimensionality: The case of job satisfaction. *Organizational Research Methods, 18*(3), 512–541. doi:10.1177/1094428115571894

Lin, C. Y., Strong, C., Tsai, M. C., & Lee, C. T. (2017). Raters interpret positively and negatively worded items similarly in a quality of life instrument for children: Kid-KINDL. *INQUIRY: The Journal of Health Care Organization, Provision, and Financing*. https://doi.org/10.1177/0046958017696724.

Magazine, S. L., Williams, L. J., & Williams, M. L. (1996). A confirmatory factor analysis examination of reverse coding effects in Meyer and Allen's Affective and Continuance Commitment Scales. *Educational and Psychological Measurement*, *56*(2), 241–250.

Marsh, H. W. (1996). Positive and negative global self-esteem: A substantively meaningful distinction or artifactors? *Journal of Personality and Social Psychology, 70*(4), 810.

Marsh, J. A., McKibben, S., Hough, H. J., Hall, M., Allbright, T. N., Matewos, A. M., & Siqueira, C. (2018). *Enacting social-emotional learning: Practices and supports employed in CORE districts and schools*. Stanford, CA: Policy Analysis for California Education. Available at: https://edpolicyinca.org/publications/sel-practices

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*(3), 437.

Meyer, R. H., Wang, Y.C., & Rice, A. B. (2018). *Measuring students' social-emotional learning among California's CORE districts: An IRT modeling approach*. CORE-PACE research partnership working paper. Available at https://edpolicyinca.org/sites/default/files/Measuring_SEL_May-2018.pdf

Nunnally, J. (1978). *Psychometric theory (2nd ed)*. New York, NY: McGraw-Hill.

Policy Analysis for California Education. (2018). *CORE-PACE research partnership: Social-emotional learning*. Available at: https://www.edpolicyinca.org/projects/core-pace-research-partnership/sel

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.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika monograph supplement*.

Schmitt, N., & Stults, D. M. (1985). Factors defined by negatively coded items: The result of careless respondents?" *Applied Psychological Measurement, 9*(4), 367–373.

Schriesheim, C. A., Eisenbach, R. J. & Hill, K. D. (1991). The effect of negation and polar opposite item reversals on questionnaire reliability and validity: An experimental investigation. *Educational and Psychological Measurement, 51(1)*, 67–78.

Snipes, J., & Tran, L. (2017). Growth mindset, performance avoidance, and academic behaviors in Clark County School District. *Regional Educational Laboratory at West Ed*. REL 2017-226.

Sonderen E., Sanderman R., & Coyne J. C. (2013). Ineffectiveness of reverse wording of questionnaire items: Let's learn from cows in the rain. *PLOS ONE 8(7)*: e68967. https://doi.org/10.1371/journal.pone.0068967

Spiegelhalter, D., Thomas, A., & Best, N. (2003). *WinBUGS version 1.4 user manual.* Cambridge, England: MRC Biostatistics Unit.

Weijters, B., & Baumgartner, H. (2012). Misresponse to reversed and negated items in surveys: A review. *Journal of Marketing Research, 49*, 737–747.

Weijters, B., Baumgartner, H., & Schillewaert, N. (2013). Reverse item bias: An integrative model. *Psychological Methods, 18*, 320–334.

West, M. R., Buckley, K., Krachman, S. B., & Bookman, N. (2018). Development and implementation of student social-emotional surveys in the CORE districts. *Journal of Applied Developmental Psychology, 55*, 119–129.

Woods, C. M. (2006). Careless responding to reverse-worded items: Implications for confirmatory factor analysis. *Journal of Psychopathology and Behavioral Assessment*, *28*(3), 186.

Appendix A. SEL Items

I. Self-Management

Please answer how often you did the following during the past 30 days. During the past 30 days…

1. I came to class prepared.
2. I remembered and followed directions.
3. I got my work done right away instead of waiting until the last minute.
4. I paid attention, even when there were distractions.
5. I worked independently with focus.
6. I stayed calm even when others bothered or criticized me.
7. I allowed others to speak without interruption.
8. I was polite to adults and peers.
9. I kept my temper in check.

(Almost Never, Once in a While, Sometimes, Often, Almost All the Time)

II. Growth Mindset

In this section, please think about your learning in general.
Please indicate how true each of the following statements is for you:

10. My intelligence is something that I can't change very much.
11. Challenging myself won't make me any smarter.
12. There are some things I am not capable of learning.
13. If I am not naturally smart in a subject, I will never do well in it.

(Not At All True, A Little True, Somewhat True, Mostly True, Completely True)

III. Self-Efficacy

How confident are you about the following at school?

14. I can earn an A in my classes.
15. I can do well on all my tests, even when they're difficult.
16. I can master the hardest topics in my classes.
17. I can meet all the learning goals my teachers set.

(Not At All Confident, A Little Confident, Somewhat Confident, Mostly Confident, Completely Confident)

IV. Social Awareness

In this section, please help us better understand your thoughts and actions when you are with other people. Please answer how often you did the following during the past 30 days. During the past 30 days…

18. How carefully did you listen to other people's points of view?
(Not Carefully At All, Slightly Carefully, Somewhat Carefully, Quite Carefully, Extremely Carefully)

19. How much did you care about other people's feelings?
(Did Not Care At All, Cared A Little Bit, Cared Somewhat, Cared Quite A Bit, Cared A Tremendous Amount)

20. How often did you compliment others' accomplishments?
(Almost Never, Once in a while, Sometimes, Often, Almost all the time)

21. How well did you get along with students who are different from you?
(Did Not Get Along At All, Got Along A Little Bit, Got Along Somewhat, Got Along Pretty Well, Got Along Extremely Well)

22. How clearly were you able to describe your feelings?
(Not At All Clearly, Slightly Clearly, Somewhat Clearly, Quite Clearly, Extremely Clearly)

23. When others disagreed with you, how respectful were you of their views?
(Not At All Respectful, Slightly Respectful, Somewhat Respectful, Quite Respectful, Extremely Respectful)

24. To what extent were you able to stand up for yourself without putting others down?
(Not At All, A Little Bit, Somewhat, Quite A Bit, A Tremendous Amount)

25. To what extent were you able to disagree with others without starting an argument?
(Not At All, A Little Bit, Somewhat, Quite A Bit, A Tremendous Amount)

Appendix B. Interscale Correlations by Grade

(Lower Triangle: Raw Correlations; Upper Triangle: Disattenuated Correlations)

Grade 3					Grade 4				
	SM	GM	SE	SA		SM	GM	SE	SA
Self-Management	1	0.246	0.627	0.657	Self-Management	1	0.258	0.605	0.664
Growth Mindset	0.172	1	0.245	0.174	Growth Mindset	0.186	1	0.279	0.192
Self-Efficacy	0.497	0.170	1	0.655	Self-Efficacy	0.501	0.200	1	0.613
Social Awareness	0.515	0.120	0. 509	1	Social Awareness	0.534	0.134	0.491	1
Grade 5					Grade 6				
	SM	GM	SE	SA		SM	GM	SE	SA
Self-Management	1	0.298	0.571	0.675	Self-Management	1	0.313	0.566	0.596
Growth Mindset	0.222	1	0.357	0.240	Growth Mindset	0.238	1	0.401	0.246
Self-Efficacy	0.482	0.267	1	0.599	Self-Efficacy	0.487	0.305	1	0.537
Social Awareness	0.551	0.173	0.490	1	Social Awareness	0.493	0.180	0.446	1
Grade 7					Grade 8				
	SM	GM	SE	SA		SM	GM	SE	SA
Self-Management	1	0.299	0.561	0.586	Self-Management	1	0.292	0.511	0.572
Growth Mindset	0.234	1	0.393	0.254	Growth Mindset	0.232	1	0.408	0.258
Self-Efficacy	0.492	0.308	1	0.530	Self-Efficacy	0.449	0.327	1	0.479
Social Awareness	0.495	0.192	0.448	1	Social Awareness	0.482	0.199	0.406	1

Grade 9					Grade 10				
	SM	GM	SE	SA		SM	GM	SE	SA
Self-Management	1	0.288	0.492	0.563	Self-Management	1	0.307	0.436	0.518
Growth Mindset	0.233	1	0.377	0.244	Growth Mindset	0.248	1	0.395	0.261
Self-Efficacy	0.434	0.308	1	0.464	Self-Efficacy	0.382	0.323	1	0.409
Social Awareness	0.475	0.191	0.394	1	Social Awareness	0.435	0.205	0.348	1
Grade 11					Grade 12				
	SM	GM	SE	SA		SM	GM	SE	SA
Self-Management	1	0.274	0.414	0.523	Self-Management	1	0.273	0.421	0.516
Growth Mindset	0.224	1	0.371	0.234	Growth Mindset	0.225	1	0.334	0.254
Self-Efficacy	0.363	0.307	1	0.384	Self-Efficacy	0.369	0.278	1	0.398
Social Awareness	0.440	0.185	0.326	1	Social Awareness	0.436	0.203	0.340	1