ANNENBERG
BROWN UNIVERSITY

EdWorkingPaper No. 20-194

Do Response Styles Affect Estimates of Growth on Social-emotional Constructs? Evidence from Four Years of Longitudinal Survey Scores

James Soland
University of Virginia

Megan Kuhfeld
NWEA

Survey respondents use different response styles when they use the categories of the Likert scale differently despite having the same true score on the construct of interest. For example, respondents may be more likely to use the extremes of the response scale independent of their true score. Research already shows that differing response styles can create a construct-irrelevant source of bias that distorts fundamental inferences made based on survey data. While some initial studies examine the effect of response styles on survey scores in longitudinal analyses, the issue of how response styles affect estimates of growth is underexamined. In this study, we conducted empirical and simulation analyses in which we scored surveys using item response theory (IRT) models that do and do not account for response styles, and then used those different scores in growth models and compared results. Generally, we found that response styles can affect estimates of growth parameters including the slope, but that the effects vary by psychological construct, response style, and model used.

VERSION: January 2020

Suggested citation: Soland, James, and Megan Kuhfeld. (2020). Do Response Styles Affect Estimates of Growth on Social-emotional Constructs? Evidence from Four Years of Longitudinal Survey Scores. (EdWorkingPaper: 20-194). Retrieved from Annenberg Institute at Brown University: https://www.edworkingpapers.com/ai20-194

Do Response Styles Affect Estimates of Growth on Social-emotional Constructs? Evidence from Four Years of Longitudinal Survey Scores

James Soland
Assistant Professor, University of Virginia
Affiliated Research Fellow, NWEA
(Corresponding Author)

Megan Kuhfeld
Research Scientist, NWEA

NWEA
121 N.W. Everett Street
Portland, OR 97209
Ph. (503) 444-6449
jim.soland@nwea.org

Abstract

Survey respondents use different response styles when they use the categories of the Likert scale differently despite having the same true score on the construct of interest. For example, respondents may be more likely to use the extremes of the response scale independent of their true score. Research already shows that differing response styles can create a construct-irrelevant source of bias that distorts fundamental inferences made based on survey data. While some initial studies examine the effect of response styles on survey scores in longitudinal analyses, the issue of how response styles affect estimates of growth is underexamined. In this study, we conducted empirical and simulation analyses in which we scored surveys using item response theory (IRT) models that do and do not account for response styles, and then used those different scores in growth models and compared results. Generally, we found that response styles can affect estimates of growth parameters including the slope, but that the effects vary by psychological construct, response style, and model used.

Keywords: self-report bias, response style, multidimensional item response theory, growth modeling, developmental psychology

Do Response Styles Affect Estimates of Growth on Socio-emotional Constructs? Evidence from Four Years of Longitudinal Survey Scores

Most evidence that social scientists have produced on how psychological constructs develop over time relies on survey responses. When growth on the construct is the parameter of interest, respondents are frequently asked to rate their levels of a given construct on a set of survey items at multiple timepoints (Weijters, Geuens, & Schillewaert, 2010a). For example, in the field of children's socio-emotional development, much of what is known about academic self-efficacy among children (one of the most studied constructs in the educational psychology literature) involves fitting growth models using repeated self-efficacy survey measures, oftentimes relying on a sum score from the Likert scale (Bauer & Curran, 2015). Self-efficacy aside, much of what we know about how individuals develop and grow psychologically and socio-emotionally depends on data from self-report measures.

While there are many well-documented issues with self-report measures, one that has gained increased attention is differences in how individuals translate their responses to the survey items onto the Likert scale. Specifically, there is an increased concern about the potentially biasing effects of response styles, which refer to idiosyncrasies in how individual respondents use rating scales in ways that are construct-irrelevant (Baumgartner & Steenkamp, 2001). For example, respondents may be more or less likely to endorse response categories at opposite ends of the Likert scale (extreme response style or ERS) or to select higher categories on the scale that make the respondent appear in the best light (socially desirable responding or SDR) despite having the same true score on the construct (Bolt & Johnson, 2009; Deng, McCarthy, Piper, Baker, & Bolt, 2018). Research demonstrates not only that differing response styles can bias individual and aggregate inferences based on surveys, but also that these response

styles are often correlated with demographics like education level and trait anxiety (Van Vaerenbergh & Thomas, 2013). Thus, varying response styles are a threat to basic inferences that social scientists wish to make on the basis of survey scores.

Despite this growing body of research, little evidence exists on how much bias ERS and SDR induce when examining growth. This omission is troubling given just how much we know about psychological and socio-emotional development is based on growth models that use survey scores as outcomes (Bauer & Curran, 2015; Guay, Marsh, & Boivin, 2003). A handful of studies have looked at this issue with data from more than one timepoint, including the effect of response style on intra-individual variance (Deng et al., 2018). However, no studies we are aware of look at the effect of ERS and SDR on the parameters from a growth model using the associated survey scores. This omission means that we have little information on how response styles may have affected the inferences drawn from the thousands of papers that have examined growth on psychological constructs.

Further, while there is an emergent consensus that response style is a fairly stable within-person construct over time (Weijters, Geuens, & Schillewaert, 2010b; Weijters et al., 2010a; Wetzel, Lüdtke, Zettler, & Böhnke, 2016), that evidence base remains fairly small. Understanding the nature of response styles over time has implications for how they may affect growth estimates on the main psychological construct of interest. For example, models have been developed to account for differing response styles at a single point in time. If response styles are stable over time, then fitting those point-in-time models across all years in the data may be an appropriate way to correct for SDR and ERS longitudinally. If, however, response styles shift over time, then corrective submodels may need to be fit in each year to adjust for the

time-specific effect of response style for a given person. More evidence is needed to understand which approach is most justifiable.

There are also gaps in our understanding of how response styles might affect inferences drawn from survey scores produced by children, including inferences about growth. Almost all of our understanding of response styles comes from surveys administered to adult populations (e.g., Bolt & Johnson, 2009; Deng et al., 2018) or high school students (Lu & Bolt, 2015). Therefore, we can say little about the prevalence and developmental nature of response styles in adolescent populations, even though students are increasingly being evaluated on their self-report responses of their socio-emotional competencies (West, Buckley, Krachman, & Bookman, 2017). To our knowledge, our study is the first to examine developmental patterns of response styles in middle school students.

In this study, we conduct empirical and simulation studies to examine the effects of response style on estimates of students' socio-emotional development. Our empirical data consist of four years of longitudinal survey scores on three socio-emotional constructs, which we use to examine the effect of response style (ERS and SDR in particular) on estimates of growth. We do so by estimating growth using scores that do not correct for response style in a growth model, then comparing results to models using scores that do. Specifically, we use a multidimensional nominal response model (MNRM) first developed by Bolt and Johnson (2009) and further developed by Falk and Cai (2016). We fit these models in two ways: (1) assuming response style is stable over time and (2) assuming response style varies for a given person over time. In so doing, we not only examine how correcting for response style in different ways affects growth parameter estimates; we also provide evidence on how stable response styles are over time, which can provide insight into how best to correct for it. Finally, we use results from

our empirical analyses to generate data with known true scores and response style artifacts, then see how well different models recover the true score-based growth parameters.

Through these analyses, we address three broad research questions:

1. Do patterns of response styles differ over time for middle school students?
2. Does accounting for response styles affect our understanding of students' SEL development when scores are used in latent growth models?
3. Do growth models that use scores from IRT models designed to account for response style effectively recover true growth parameters?

We should note that we diverge from the common practice of beginning with a simulation study and following it up with an empirical study (here, that order is reversed). We begin with the empirical study for a couple of reasons. First, while there is some limited evidence that response style is stable over time (Weijters, Geuens, Schillewaert, 2008), we wanted to determine whether that finding held up in our own data given most of the response style research conducted to date has been on adults, not students. That is, we wanted to have some empirical basis for deciding whether or not to assume that response style is time-invariant. Second, there is not much current research that documents just how extreme the MNRM parameters capturing response style can be in practice, especially among children and young adults. We therefore use estimates of those parameters from the empirical study as the basis for our simulation, including defining response style bias that is especially large in relation to those actual empirical values. In short, the limited research on response style bias among students rather than adults meant we wanted the empirical results as a basis for our simulation.

Background

In this section, we describe methods for detecting and correcting for differing response styles, then review evidence on the effect that response styles have on inferences drawn from surveys of psychological constructs (and socio-emotional constructs in particular). Generally, several approaches have been used to address differing response styles. For example, some surveys attempt to address the issue during measurement by including anchoring vignettes designed to give respondents a common frame of reference that reduces idiosyncrasies in how the Likert scale is used (Kyllonen & Bertling, 2013). Meanwhile, other researchers have administered separate surveys designed to understand how much a particular respondent might have succumbed to social desirability bias (Fisher, 1993). The respondents' social desirability bias can then be controlled for in subsequent models. By contrast, other approaches to addressing response style bias involve fitting models that simultaneously estimate factors for the construct of interest and response style (Bolt & Johnson, 2009; Falk & Cai, 2016). Given that many (if not most) surveys typically used in modeling students' socio-emotional development do not include anchoring vignettes or separate social desirability surveys (Fisher, 1993; Kyllonen & Bertling, 2013), our study focuses on post-hoc model-based approaches to addressing response style.

Post-hoc Methods for Addressing Response Style Bias

A range of post-hoc methods have been proposed to detect and account for differing response styles. These approaches include factor analytic models (Ferrando & Lorenzo-Seva, 2007), structural equation models (SEMs) (Cheung & Rensvold, 2000), multinomial processing tree (MPT) models used in cognitive psychology (Park & Wu, 2019; Plieninger & Heck, 2018), proportional threshold models (Thissen-Roe & Thissen, 2013), IRT-based MNRM models (e.g., Bolt & Johnson, 2009; Deng et al., 2018), and still others (Weijters et al., 2010a). Extensions of

these models have also been developed, such as those proposed to the MNRM by Falk and Cai (2016).

Despite the range of models available, we focus on the MNRM because it allows one to explicitly model different response styles and jointly estimate response style and ability factors (Bolt & Johnson, 2009; Falk & Cai, 2016). That is, the MNRM not only provides an option for detecting response style bias, but also for producing scores that correct for it in an IRT framework. Further, MNRM studies suggest that, relative to other approaches, the modified generalized partial credit model results in the lowest item mean squared error (MSE) across various simulation conditions (Leventhal, 2019). We detail the MNRM below.

MNRM. The MNRM is a multivariate generalization of the standard nominal response model (NRM) developed by Bock (1972). We will generally follow the notation used by Falk and Cai (2016). Let $i = 1, \dots, N$ persons responding to $j = 1, \dots, n$ items with Y_{ij} being a random variable for the corresponding item responses and y_{ij} its realization. There are $k = 1, \dots, K_j$ possible ordered response options for item j. $\mathbf{x}_\mathrm{i}$ is then a $D\ x\ 1$ vector of person i's factor scores for $d = 1, \dots, D$ latent dimensions assumed multivariate normal with covariance structure Σ. As matrices, $\mathbf{X}$ is an $N\ x\ D$ matrix containing all factor scores and $\mathbf{Y}$ is an $N\ x\ n$ matrix of item responses. Removing subscripts for item and person, one could express the MNRM as

$$P(Y = k\ |\mathbf{x}, \tilde{\mathbf{a}}, \mathbf{c}) = \frac{\exp(\tilde{\mathbf{a}}_k'\mathbf{x} + c_k)}{\sum_{m=1}^{K} \exp(\tilde{\mathbf{a}}_m'\mathbf{x} + c_m)}$$

Where $\tilde{\mathbf{a}}_k'$ is a D x 1 vector of slopes that represents loadings of category k on the D latent variables and c_k is an intercept.

Thissen and Cai (2018) presented the following re-parameterization of the MNRM:

$$P(Y = k\ |\mathbf{x}, \mathbf{a}, \mathbf{S}, \mathbf{c}) = \frac{\exp([\mathbf{a}\ \circ \mathbf{s}_k]'\mathbf{x} + c_k)}{\sum_{m=1}^{K} \exp([\mathbf{a}\ \circ \mathbf{s}_m]'\mathbf{x} + c_m)}$$

where $[\mathbf{a} \circ \mathbf{s}_k]$ is the Schur product of the slopes, $\mathbf{a}$, and $\mathbf{s}_k$ is a scoring function. $\mathbf{s}_k$ is part of a D x K scoring function matrix $\mathbf{S}$ where each column represents a particular category for the item and each row represents a given factor. For identification, estimation of the intercept parameters is done by estimating $\boldsymbol{\gamma}$, where $\boldsymbol{c} = \boldsymbol{T_c}\boldsymbol{\gamma}$ (see Thissen & Cai [2018] for more details). We follow this parameterization rather than that proposed by Bolt and Johnson (2009) because it follows the model parameterization used in flexMIRT, the software we employ.

The scoring functions are key to understanding the applicability of the MNRM to response style issues. Though these are nominal models, a scoring function equivalent to $\{s_{d,1}, s_{d,2}, s_{d,3}, s_{d,4}, s_{d,5}\} = \{0,1,2,3,4\}$ with $s_{d,k}$ corresponding to row d and column k of $\mathbf{S}$ is equivalent to the generalized partial credit model (GPCM). By contrast, consider the case of wanting to address ERS. For an ERS factor, the scoring function would be $\{s_{d,1}, s_{d,2}, s_{d,3}, s_{d,4}, s_{d,5}\} = \{1,0,0,0,1\}$. This scoring function means that, in addition to the slopes generated by the GPCM portion of the model, the slopes on the factor can shift additionally when the respondent selected one of the extreme response categories. Thus, as illustrated by Falk and Cai (2016), the item response function resembles that of the GPCM, but the functions for the first and last response categories would look different dependent on the level of ERS detected.

Research has similarly established that the appropriate scoring function for socially desirable responding (SDR) is $\{s_{d,1}, s_{d,2}, s_{d,3}, s_{d,4}, s_{d,5}\} = \{0,0,0,1,0\}$ in the case of five response categories (e.g., Falk & Cai, 2016). The logic to this scoring function is that a student with a low true score on the trait might wish to respond to all items in a socially desirable way (Kuncel & Tellegen, 2009; Paulhus,1991) to make himself or herself look good to others. However,

selecting the top category might appear suspicious; therefore, the category below the top one is chosen.

Research on Response Styles for Socio-emotional Constructs

Research already demonstrates that, at a single point in time, differing response styles can affect fundamental inferences based on survey scores (e.g., Billiet & McClendon, 2000), though not all studies show a practically significant effect of response style bias (Plieninger, 2017). Findings of bias due to response styles have been replicated across a wide range of methodological approaches to detecting and correcting for response style, as well as a wide range of empirical datasets. Several studies have used the MNRM. For example, Bolt and Johnson (2009) used the MNRM with the Wisconsin Inventory of Smoking Dependence Motives, a self-report measure of tobacco dependence. Using those data, they identified a secondary trait related to ERS (Bolt & Johnson, 2009). Research using the MNRM has also shown that response styles can affect estimated treatment effects, and effect sizes in particular. Dowling, Bolt, Deng, and Li (2016) found that effect sizes produced by simple sum scores were small compared to those produced by the MNRM. These results suggest that accounting for ERS behavior using multidimensional IRT approaches may substantially increase the value of psychological measures as evidence to support decision-making in clinical and health policy development (Dowling et al., 2016).

Meanwhile, other studies have documented the effects of response styles in an SEM framework, many of which examine the issue in an international context. In one such study, the results of an empirical example demonstrated that American and Italian samples were invariant with respect to factor form and ERS (factor loadings), but that the construct of job content was noninvariant with respect to acquiescent response style (intercepts), meaning the job content

construct could not be compared across cultures (Cheung & Rensvold, 2000). Another international study used SEM with a national crime victimization survey in Belgium (Heerwegh & Loosveldt, 2011). Results showed that, consistent with the social desirability hypothesis, responses were significantly more positive in the telephone survey, but no evidence was found for differences in response styles across the survey modes (Heerwegh & Loosveldt, 2011). Other related work applied a response style model to survey items from TALIS 2013; results indicated that self-efficacy items were more likely to trigger ERS compared to need for professional development, and the between-country relationships among constructs changed thanks to ERS (Ju & Falk, 2019). Finally, Maydeu-Olivares and Coffman (2006) demonstrated with an empirical dataset on optimism that idiosyncratic use of the Likert scale can affect estimated scores and related inferences.

A pair of studies have examined response styles in a longitudinal context. Perhaps most relevant to our own study, Deng, McCarthy, Piper, Baker, and Bolt (2018) modeled responses to scales of positive and negative affect from smokers at clinic visits following a smoking cessation program. Those analyses revealed considerable ERS bias in the intra-individual sum score variances (Deng, McCarthy, Piper, Baker, & Bolt, 2018). A related study looked more directly at whether response styles themselves are variable over time and within persons. Weijters, Geuens, and Schillewaert (2010b) modeled four response styles: acquiescence, disacquiescence, midpoint, and ERS. Their results provide evidence that response styles have a large stable component, only a small part of which is associated with demographics (Weijters et al., 2010b).

Summary

In sum, response style bias can affect fundamental inferences that researchers draw from survey scores. Models like the MNRM have been developed not only to detect response style

bias post hoc, but also to produce scores for the construct of interest that correct for response style bias. Yet, very little of this research examines the effect of response styles on repeated measures uses of survey scores, and no studies we are aware of examine the issue in a growth modeling framework. This omission is a problem given most of what we know about students' socio-emotional growth is based on survey scores used in growth models (or closely related models).

Empirical Study

Sample

Our sample consists of a cohort of students who began in 5th grade in the 2014-15 school year and finished in 8th grade in 2017-18. Details on the analytic sample are provided in Table 1. As the table shows, the cohorts are not intact: students could move in and out of the sample as long as they had at least one survey score. Sample size in a given year ranged from 2,319 students to 3,466 students with survey scores. The district we studied is in California and has a high proportion of Latinx students, the majority of whom are low-income. On average, students at the median for math and reading achievement in the district are well below the 50th percentile for the nation per norms developed by Thum and Hauser (2015).

Measures

The district we studied offers a yearly survey of three socio-emotional constructs: self-efficacy, growth mindset, and self-management. These surveys use Likert scales with five response categories. Specific questions by construct are provided in the Appendix (Table A1) and pertinent details (reliability, mean, variance, etc.) about the scales are in Table 2. While previous researchers have analyzed the psychometric properties of the CORE district SEL surveys (West et al., 2017), we nonetheless did analyses to ensure basic properties were met. For

example, exploratory factor analysis suggests that each set of items loads on only a single factor and confirmatory factor analyses suggest sufficient model fit (RMSEA<.05).

One should note that the CORE districts included these survey measures as part of an accountability system they developed when obtaining a waiver of provisions of *The No Child Left Behind Act of 2001*. Thus, scores from these surveys had stakes attached to them for schools (if quite modest). In addition, researchers affiliated with CORE have examined teacher and school contributions to these survey scores over time, studies motivated in part by a desire to include measures other than achievement in accountability systems (West, 2016). Given these current and possible future uses of the surveys, understanding the effect of response style bias on growth is important.

Analytic Strategy

Scoring Approach. The general approach to answering our research questions involved calibrating and scoring student item responses using IRT models that did and did not account for response styles, then using those scores in growth models and comparing the parameters. We used several models to estimate student scores for each latent SEL factor (e.g., self-efficacy, growth mindset, and self-management). First, we began with a multidimensional generalized partial credit (MGPC) model that ignored response style and calibrates the item responses for all four timepoints simultaneously. As Thissen, Cai, and Bock (2010) have pointed out, the MGPC is a constrained version of the MNRM model. The path diagram for that MGPC model is depicted in Figure 1. We chose this approach to calibrate the items given research showing that such a MIRT model does a better job of recovering true scores in a longitudinal context than IRT models that are calibrated based on a single timepoint (Authors, 2019; Bauer & Curran, 2015).

Measurement invariance constraints were placed on the slopes and intercepts to fix the item parameters for each repeated item to be equal across the four timepoints.

We then supplemented these MIRT models by jointly estimating response style models using MNRM submodels (for both ERS and SDR separately). In one set of models, the response style factor was treated as stable over time. For example, the MNRM model accounting for ERS assumed that a single ERS factor affected observed item responses across all time periods. Such models thus had five dimensions, one per socio-emotional construct and timepoint plus one overall response style dimension across timepoints. In another set of models, we assumed that the effect of response styles varied over time and by item. Those models were eight dimensional, with time-specific factors for the socio-emotional and response style constructs. For those models, we allowed the covariance structure of the response style factors to be freely estimated while constraining the SEL and response style factors to be uncorrelated. Altogether, our models included: (1) 4-D MGPC with no response style factor, (2) 5-D MNRM for ERS assuming response style is time invariant, (3) 5-D MNRM for SDR assuming response style is time invariant, (4) 8-D MNRM for ERS assuming response style is time varying, and (5) 8-D MNRM for SDR assuming response style is time varying. All models were estimated in flexMIRT version 3.5 using the Metropolis-Hastings Robbins-Monro (MH-RM) algorithm (Cai, 2010a, 2010b). After obtaining estimated item parameters for each model, we scored all of the student responses using the expected a posteriori (EAP) method (Bock & Mislevy, 1982).

***Question 1. Do patterns of response styles differ by grade and over time*?** We examined this question using two approaches. The first involved descriptive statistics from our sample of roughly 3,000 students. Specifically, for each of the three SEL constructs used by the district, we began by determining on what proportion of items a given student used particular response

categories corresponding to MNRM scoring functions. For example, in terms of ERS, this approach followed the scoring function in the MNRM and meant assigning a 1 to any item response that used the highest or lowest response category, 0 otherwise. Then, the proportion of all items on which the extreme response categories were used for a given student on the given construct was calculated. Finally, the mean of those proportions across all students was estimated. For socially desirable responding, the same process was used, but coding responses to 1 if the fourth response category out of five was used, 0 otherwise. To see if the mean proportion of extreme or socially desirable item selection changed by grade or over time, we presented those means by year (2015-2018) for our cohort beginning in 5th grade.

The second involved examining the covariance structure of the response style factors in the MNRM models that treated the response style factors as time-varying (eight-dimensional models). Specifically, we examined how correlated the factors were between time one and time two versus time one with times three and four. If one assumes that response style is consistent over time, then correlations between adjacent timepoints should not be much larger than those between more distal timepoints (Prenoveau, 2016).

***Question 2. Does accounting for response styles affect our understanding of students' SEL development*?** We began by comparing the estimated latent mean and covariances for each SEL construct by time period and across models to see if general trends in scores and magnitude of their change by timepoint was comparable. In addition, we compared the correlations among the SEL latent variables from the various MNRM models to see if accounting for response style appeared to make those constructs more or less stable over time.

Response Style and Growth Estimates

After examining results from our MIRT scoring models, we used those scores as the dependent variable in a growth model. Specifically, we estimated a two-level model (multiple timepoints nested within students) for student i at time t such that

$$\hat{x}_{it} = \beta_{0i} + \beta_{1i} time_{it} + e_{ti}$$

$$\beta_{0i} = \gamma_{00} + u_{0i}$$

$$\beta_{1i} = \gamma_{01} + u_{1i,}$$

where $time_{it}$ is coded {0,1,2,3} so that the intercept represents the average student's score on the latent construct in 5th grade. In this model:

$$e_{it} \sim \mathrm{N}(0, \sigma_{it}^2)$$

$$\boldsymbol{u_i} \sim \mathrm{MVN}(\mathbf{0}, \boldsymbol{\tau}).$$

Several parameters from these models were compared within construct and across models. First, we examined $\hat{\gamma}_{00}$ and $\hat{\gamma}_{01}$ to see how fixed effect estimates of the SEL construct in 5th grade and growth on that construct through 8th grade compared. Given our aim of recovering growth estimates, the consistency of $\hat{\gamma}_{01}$ across models was of particular importance. Second, we used $\boldsymbol{\tau}$ to compare the variance and covariance of the random effects across growth models.

Results

***Question 1. Do Patterns of Response Styles Differ by Grade and over Time*?** Table 2 presents the mean proportion of items on which each student chose response categories associated with extreme responses and socially desirable response styles by time period for the cohort. Thus, on average, students who were in 5th grade in 2015 selected either 1 or 5 on 44% of growth mindset items, and those same students selected either 1 or 5 on 37% of those items when they reached 8th grade in 2018. In general, these rates are high primarily because students

were quite likely to use the top response category when responding to items. ERS tended to decline as students moved through school, dropping by anywhere from 7 to 10 percentage points between 5th and 8th grade. While rates of selecting the socially desirable response category were lower than the rates of selecting extreme response categories, there was no clear pattern to how the rates of category selection changed over time.

Table 3 shows correlation matrices for latent response time factors by construct and model for each of our eight-dimensional MIRT models. In general, correlations among response style factor scores are low to moderate. Further, they tend to decrease over time, though the diminution is not always consistent by construct and response style. For example, for growth mindset and ERS, correlations of the time one and time four response style factors decreased by .29 compared to the time one and time two correlation. For self-efficacy, that same decrease was .09. Whereas self-management tends to produce decreased correlations over time regardless of model, self-efficacy shows relatively small decreases in correlations across all models. In short, the latent response style covariance matrices indicate that assuming response styles are stable traits may not always be justified, and the tenability of that assumption likely depends on the SEL construct and type of response style.

***Question 2. Does accounting for response styles affect our understanding of students' SEL development*?** As discussed in the methods section, several approaches were taken to examine this question. First, Table 4 shows means and variances of scores on the three SEL constructs by timepoint and model used to produce the scores. In some cases, accounting for response style appears to have little effect on means and variances. For example, there are few differences in means and variances across timepoints and models for self-efficacy. For self-management, there are practically no differences except for the time-varying ERS model. By

contrast, there are statistically and practically significant differences by model for growth mindset. For most growth mindset models that account for response style, the estimated SEL factor latent means are lower by the third and fourth timepoints relative to the model that does not account for response style. However, the means are actually higher at those timepoints for the MNRM ERS model that assumes response style is time-varying.

Tables 5(a)-5(c) present correlation matrices for the latent SEL factors over time for growth mindset, self-efficacy, and self-management (respectively). Research already shows that these three SEL constructs tend to show lower rank-order stability than achievement, with correlations decreasing steadily over time (Authors, 2019). However, one might hypothesize that some of the observed rank-order stability of the SEL constructs could be due to consistent use of the same response style across repeated survey administrations, meaning that the true stability of SEL scores could be lower than we think. For all three constructs, the stability of the latent factors does not appear to shift much by model. For growth mindset, the correlation between times one and four using the model with no response style correction is .08 units lower than the correlation between times one and two. That decrease in the correlation tends to be quite similar for models that do account for response style, though the drop for the time-varying MNRM with extreme response styles is larger (.21). Results are comparable for the other constructs.

Finally, Table 6 presents growth parameter estimates by construct and model. Once again, results are variable by SEL construct and model. For example, on growth mindset, changes in the MNRM slope relative to the slope from the model that does not account for response style ranges from .02 logits (MNRM ERS time-invariant) to .06 logits (MNRM ERS time-varying), the latter representing a slope that is more than half that of the original estimate. By contrast, slope parameters differ little across all models for both self-efficacy and self-

management. Meanwhile, though most of the variance components do not appear to be heavily affected by accounting for response style, there are noteworthy exceptions. In particular for self-efficacy and growth mindset, the variances of the constant and slope change by more than .4 units for both MNRM models that account for SDR (time-invariant and time-varying).

Simulation Study

Two small simulation studies were conducted to examine the ability of the MNRM model to recover true factor scores and the growth estimates based on those true scores under different degrees of extreme response styles. Our empirical study results allowed us to (a) test assumptions about whether response styles are time-varying and (b) obtain generating parameters that would be expected in real-world data. We used those findings to set up two simulation studies (one mirroring observed conditions, one that assumes a stronger influence of ERS) where the true scores are known. In so doing, we could determine whether growth parameters are better recovered when we (a) explicitly account for the influence of response styles or (b) ignore response styles in the calibration/scoring models.

Methods

The data-generating models for all conditions were based in part on models fit to actual data in the empirical analyses. Specifically, we used the estimated item parameters and factor correlations from the time-varying MNRM for growth mindset (Table A2 in the appendix) in the empirical study as population values for the simulation. We chose those model results because they produced the largest difference in the estimated growth parameters in the empirical study. For simplicity, we refer to the simulation based exactly on the empirical results as “Simulation 1.”

Response Style and Growth Estimates

To further test how much differing response styles might affect growth estimates, we once again generated data based on the empirical MNRM parameters for growth mindset, but with one key change: we doubled the slope parameters on the response style factors. While these response style parameters are large relative to the empirical results, they are not out of line with similar parameters reported in other studies (e.g., Deng et al., 2018). We refer to these conditions as "Simulation 2."

All data were simulated using flexMIRT. We used a sample size of 3,000 simulees and data from four timepoints. While the timepoints matched the empirical study, we increased the sample size relative to the empirical data in order to understand the effects of response style with less concern for sampling error. All simulations were replicated 100 times. Details common to estimated scores in simulations included a burn-in of 10 for draws from the posterior predictive distribution, followed by 3000 Stage I iterations, 1000 Stage II iterations with constant gain constants of 0.1, and 2000 Stage III iterations with an initial gain constant of 1. All but one model converged in fewer than 300 Stage III iterations.

Through these two simulations, we were able to investigate how different estimated growth parameters were when using true scores, scores from the MIRT model in Figure 1 that ignored response style, and scores from the time-varying MNRM that matched the data generating process. As before, the scoring function for the MNRM was [1,0,0,0,1]. Each simulation proceeded by (1) simulating observed and true scores, (2) scoring simulated observed scores, and (3) fitting the growth models from Question 2 in the empirical study to the IRT-based scores, as well as the true scores from the simulation. Then, as for Question 2 in the empirical study, we compared the fixed effects and variance components from the growth models across different sets of scores.

Results

In general, results from the simulation study are mixed. We began by examining parameter recovery for the MNRM model. The generating and estimated item parameters (averaged across the 100 replications) are shown in Appendix Tables A2 and A3 for Simulation Study 1 and Appendix Tables A4 and A5 for Simulation Study 2. Overall, the generating **a** and $\boldsymbol{\gamma}$ parameters from the generating MNRM were well recovered in both simulation studies. For the examination of growth, the recovery of the latent means and covariances is critical. Table 7 presents the generating and estimated latent means and covariances under the GPCM and MNRM for both simulations. The upper three rows are for the simulation based exactly on the empirical example (Simulation 1) and the bottom three rows are for the simulation with the response style slope parameters in the generating model doubled (Simulation 2). As the table shows, scores in timepoints two through four are better recovered for Simulation 2 using the MNRM, but neither model does appreciably better for Simulation 1. The MNRM also does a better job of recovering correlations of scores by timepoint, especially for Simulation 2. However, the MNRM also leads to variances that are overstated relative to the GPCM regardless of simulation and timepoint.

A somewhat similar pattern emerges for the growth parameter estimates displayed in Table 8. For fixed effects estimates of the slope, the MNRM outperforms the GPCM for Simulation 2. Similarly, for Simulation 2, the covariance of the slope and intercept based on the MNRM better matches estimates based on true scores. However, the variances of the intercept and slope produced by the MNRM are understated relative to estimates based on the true score. (In fact, the GPCM produces variance estimates that better match those from the model that used true scores.)

Finally, in terms of practical significance, one should note that the slope estimates are only slightly different for Simulation 1 when using the GPCM versus the MNRM. For Simulation 2, the slope is understated by .033 units when using the GPCM relative to the true score model compared to -.014 for the MNRM compared to the true score model. Thus, bias in the point estimates of the slope relative to true score slopes is not drastically different between the GPCM and MNRM, even in the case of extreme response style bias.

Discussion

Research has clearly documented that response style bias can affect basic inferences researchers and other educational stakeholders might wish to make based on surveys (e.g., Deng et al., 2018). However, relatively little is known about how differing response styles might affect growth estimates. This gap in the literature is problematic given the vast majority of what is known about how children and students develop psychologically and socio-emotionally is based on self-report survey scores. In our study, we used SEL survey scores from a large school district to conduct empirical and simulation studies that investigate the effect of response styles on growth parameters of interest. Our results make several contributions to the field.

First, while some studies suggest that response style is fairly trait-like (stable over time), we provide some evidence that its stability is dependent on the grade level, construct, and type of response style. For example, in the empirical data, use of the extreme response categories on the Likert scale declined for a cohort of students that began in 5^{th} grade, but changed little for SDR. We also showed that correlations among latent response style factors in our eight-dimensional models declined substantively as the time between the measurements increased for some but not all response styles and constructs. Ultimately, correcting for response style bias in repeated measures designs requires answering a first-order question: should the bias be addressed at each

timepoint or across all timepoints? Our results indicate that the answer to this question may depend on the type of response style and the construct being measured.

Second, we show that the effect of response style bias on growth estimates also depends on the construct and response style. For example, the slope on growth mindset was 1.5 times larger when assuming ERS is time-varying. For self-efficacy, the slope and intercept variances were 10 times smaller for an IRT model that did not account for response style versus those that did (in some cases). However, for other construct-response style-model combinations, estimated constants, slopes, and their variances differed little if at all between models that accounted for response style bias and those that did not. Changes in mean scores and correlations of estimated scores over time were also sensitive to response style, but not universally so by construct and type of response style.

Third, to better understand the situations in which ERS could bias estimates of growth, we conducted a small set of simulation studies based on our empirical data. We produced this simulation by taking empirical results that showed strong bias (relative to our other empirical analyses) and doubling the slopes on the ERS parameters. In general, for this extreme case, the MNRM did a better job of recovering true scores and covariances of the scores over time. The MNRM also better matched point estimates of the slope, and slope-intercept covariances. However, while point estimates of the slope differed by about .03 units between the true score and GPCM models, the MNRM point estimate of the slope was off by .014 units in the opposite direction compared to the true score model. Further, the MNRM tended to understate the variances of the slope and intercept parameters. Thus, one cannot definitely say that the MNRM outperformed the GPCM in recovering growth parameters even when the generating model involved especially strong response style bias. In general, differences between the GPCM and

MNRM growth parameters were modest in practical terms despite the generating model assuming extreme time-varying response style bias.

Limitations

A few limitations of this study bear mention. First, our empirical study was limited to only a single district that serves a high proportion of low-income and English learner students. Further, that district only administered a single survey of each construct per year. Thus, issues of generalizability remain. In terms of the measures, results might shift if different surveys of growth mindset, self-efficacy, and self-management were used. In terms of the sample, one cannot be sure results would hold with a different or more representative set of students.

Second, like in any simulation study, we were limited in terms of the range of conditions, assumptions, and models we could use to test our hypotheses. For example, results could differ dependent on the specific data generating model. We also did not compare all possible variations on the available MNRM models. Thus, results should be replicated using different data generating assumptions and scoring models.

Conclusion

Response style bias has been shown to affect scores from surveys in ways that undermine their desired uses. In this study, we examine the effect of this form of bias on estimates of students' socio-emotional growth. We find that the stability of within-student response style factors tends to differ by type of response style and construct. Thus, researchers attempting to estimate growth who are worried about this issue may initially want to investigate how much response styles change over time in their sample and for their construct in order to determine whether to use an IRT model that treats response style as time-varying. We also show that, while response style bias does not uniformly affect growth parameter estimates, for certain

constructs and response style types, estimates of slopes and their variances can change substantively dependent on whether the IRT model used to score the survey accounts for response style bias.

References

Authors (2019).

Bauer, D., & Curran, P. (2015). The discrepancy between measurement and modeling in longitudinal data analysis. *Advances in Multilevel Modeling for Educational Research: Addressing Practical Issues Found in Real-World Applications*, 3–38.

Baumgartner, H., & Steenkamp, J.-B. E. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research*, *38*(2), 143–156.

Billiet, J. B., & McClendon, M. J. (2000). Modeling acquiescence in measurement models for two balanced sets of items. *Structural Equation Modeling*, *7*(4), 608–628.

Bolt, D. M., & Johnson, T. R. (2009). Addressing score bias and differential item functioning due to individual differences in response style. *Applied Psychological Measurement*, *33*(5), 335–352.

Cheung, G. W., & Rensvold, R. B. (2000). Assessing extreme and acquiescence response sets in cross-cultural research using structural equations modeling. *Journal of Cross-Cultural Psychology*, *31*(2), 187–212.

Deng, S., E. McCarthy, D., E. Piper, M., B. Baker, T., & Bolt, D. M. (2018). Extreme response style and the measurement of intra-individual variability in affect. *Multivariate Behavioral Research*, *53*(2), 199–218.

Dowling, N. M., Bolt, D. M., Deng, S., & Li, C. (2016). Measurement and control of bias in patient reported outcomes using multidimensional item response theory. *BMC Medical Research Methodology*, *16*(1), 63.

Falk, C. F., & Cai, L. (2016). A flexible full-information approach to the modeling of response styles. *Psychological Methods*, *21*(3), 328.

Ferrando, P. J., & Lorenzo-Seva, U. (2007). An item response theory model for incorporating response time data in binary personality items. *Applied Psychological Measurement*, *31*(6), 525–543.

Fisher, R. J. (1993). Social desirability bias and the validity of indirect questioning. *Journal of Consumer Research*, *20*(2), 303–315.

Guay, F., Marsh, H. W., & Boivin, M. (2003). Academic self-concept and academic achievement: Developmental perspectives on their causal ordering. *Journal of Educational Psychology*, *95*(1), 124.

Heerwegh, D., & Loosveldt, G. (2011). Assessing mode effects in a national crime victimization survey using structural equation models: Social desirability bias and acquiescence. *Journal of Official Statistics*, *27*(1), 49.

Kyllonen, P. C., & Bertling, J. P. (2013). Innovative questionnaire assessment methods to increase cross-country comparability. *Handbook of International Large-Scale Assessment: Background, Technical Issues, and Methods of Data Analysis*, 277–285.

Leventhal, B. C. (2019). Extreme Response Style: A Simulation Study Comparison of Three Multidimensional Item Response Models. *Applied Psychological Measurement*, *43*(4), 322–335.

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

Plieninger, H. (2017). Mountain or molehill? A simulation study on the impact of response styles. *Educational and Psychological Measurement*, *77*(1), 32–53.

Plieninger, H., & Heck, D. W. (2018). A new model for acquiescence at the interface of psychometrics and cognitive psychology. *Multivariate Behavioral Research*, *53*(5), 633–654.

Prenoveau, J. M. (2016). Specifying and interpreting latent state–trait models with autoregression: An illustration. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(5), 731–749.

Thissen, D., & Cai, L. (2018). Nominal categories models. In W. J. van der Linden & R. K. Hambleton (Eds.), Handbook of modern item response theory (2nd ed.). New York, NY: Chapman & Hall.

Thissen-Roe, A., & Thissen, D. (2013). A two-decision model for responses to Likert-type items. *Journal of Educational and Behavioral Statistics*, *38*(5), 522–547.

Weijters, B., Geuens, M., & Schillewaert, N. (2010a). The individual consistency of acquiescence and extreme response style in self-report questionnaires. *Applied Psychological Measurement*, *34*(2), 105–121.

Weijters, B., Geuens, M., & Schillewaert, N. (2010b). The stability of individual response styles. *Psychological Methods*, *15*(1), 96.

West, M. R. (2016). Should non-cognitive skills be included in school accountability systems? Preliminary evidence from California's CORE districts. *Evidence Speaks Reports*, *1*(13).

West, M. R., Buckley, K., Krachman, S. B., & Bookman, N. (2017). Development and implementation of student social-emotional surveys in the CORE Districts. *Journal of Applied Developmental Psychology*.

Wetzel, E., Lüdtke, O., Zettler, I., & Böhnke, J. R. (2016). The stability of extreme response style and acquiescence over 8 years. *Assessment*, *23*(3), 279–291.

Tables/Figures

Table 1
Statistics on Analytic Sample

	2014-15	2015-16	2016-17	2017-18
Prop. Male	0.487	0.492	0.500	0.478
Prop. Special Ed.	0.094	0.110	0.102	0.066
Prop. Latinx	0.959	0.969	0.957	0.965
Number of students with survey scores	2,319	3,266	2,842	2,694
Math percentile (median) - MAP Growth	29	31	33	30
Reading percentile (median) - MAP Growth	29	25	29	27

Table 2

Proportion of All Responses that Used Item Response Categories Corresponding to Response Style Scoring Functions

		Extreme Response Style				Socially Desirable Response Style			
	construct	2015	2016	2017	2018	2015	2016	2017	2018
Grade 5									
	Growth Mindset	0.442	0.428	0.386	0.369	0.158	0.147	0.151	0.136
	s.e.	(0.005)	(0.005)	(0.006)	(0.005)	(0.004)	(0.004)	(0.004)	(0.004)
	N	2774	2383	1909	2202	2726	2381	1904	2202
	Self-efficacy	0.414	0.398	0.355	0.334	0.305	0.313	0.294	0.294
	s.e.	(0.004)	(0.005)	(0.005)	(0.005)	(0.005)	(0.006)	(0.007)	(0.006)
	N	2774	2383	1909	2202	2726	2380	1905	2201
	Self-management	0.450	0.429	0.380	0.348	0.274	0.308	0.317	0.333
	s.e.	(0.005)	(0.006)	(0.006)	(0.006)	(0.004)	(0.004)	(0.005)	(0.005)
	N	2773	2383	1909	2202	2773	2383	1909	2202

Table 3

Correlation Matrix for Latent Response Time Factors by Construct and Model

Growth Mindset					Self-efficacy					Self-management				
MNRM - extreme					MNRM - extreme					MNRM - extreme				
Time 1	1.00				Time 1	1.00				Time 1	1.00			
Time 2	0.59	1.00			Time 2	0.42	1.00			Time 2	0.35	1.00		
Time 3	0.43	0.56	1.00		Time 3	0.35	0.46	1.00		Time 3	0.15	0.55	1.00	
Time 4	0.36	0.50	0.66	1.00	Time 4	0.33	0.39	0.50	1.00	Time 4	0.17	0.50	0.61	1.00
MNRM - socially desirable					MNRM - socially desirable					MNRM - socially desirable				
Time 1	1.00				Time 1	1.00	0.27	0.33	0.29	Time 1	1.00	0.42	0.36	0.26
Time 2	0.43	1.00			Time 2	0.27	1.00	0.49	0.40	Time 2	0.42	1.00	0.59	0.54
Time 3	0.36	0.45	1.00		Time 3	0.33	0.49	1.00	0.56	Time 3	0.36	0.59	1.00	0.56
Time 4	0.16	0.35	0.54	1.00	Time 4	0.29	0.40	0.56	1.00	Time 4	0.26	0.54	0.56	1.00

Table 4

Means and Variances of Scores by Timepoint

Construct/Model	Score				Variance			
Growth Mindset	Time 1	Time 2	Time 3	Time 4	Time 1	Time 2	Time 3	Time 4
No response style	0	0.106	0.270	0.409	1	1.454	1.599	1.520
MNRM time invariant (extreme)	0	0.100	0.250	0.364	1	1.519	1.793	1.798
MNRM time invariant (socially desirable)	0	0.111	0.270	0.405	1	1.435	1.744	1.720
MNRM time-varying (extreme)	0	0.122	0.340	0.505	1	1.523	1.858	1.826
MNRM time-varying (socially desirable)	0	0.108	0.276	0.396	1	1.437	1.719	1.688
Self-efficacy								
No response style	0	-0.039	-0.242	-0.314	1	1.002	1.033	0.978
MNRM time invariant (extreme)	0	-0.042	-0.245	-0.326	1	1.000	1.010	0.954
MNRM time invariant (socially desirable)	0	-0.042	-0.248	-0.332	1	1.005	1.037	0.956
MNRM time-varying (extreme)	0	-0.026	-0.229	-0.309	1	1.015	1.048	0.962
MNRM time-varying (socially desirable)	0	-0.041	-0.246	-0.324	1	1.001	1.028	0.940
Self-management								
No response style	0	-0.026	-0.208	-0.248	1	0.999	1.127	1.011
MNRM time invariant (extreme)	0	-0.006	-0.184	-0.233	1	0.993	1.087	0.928
MNRM time invariant (socially desirable)	0	-0.003	-0.184	-0.229	1	1.026	1.126	0.992
MNRM time-varying (extreme)	0	0.055	-0.120	-0.172	1	1.116	1.247	1.054
MNRM time-varying (socially desirable)	0	-0.008	-0.192	-0.233	1	1.017	1.113	0.993

Table 5(a)

Growth Mindset: Correlation Matrix for Latent SEL Factors by Model

No Correction				
Time 1	1.00			
Time 2	0.50	1.00		
Time 3	0.48	0.62	1.00	
Time 4	0.42	0.54	0.67	1.00

MNRM Time Invariant - extreme				
Time 1	1.00			
Time 2	0.46	1.00		
Time 3	0.44	0.59	1.00	
Time 4	0.37	0.48	0.59	1.00

MNRM Time Varying - extreme				
Time 1	1.00			
Time 2	0.57	1.00		
Time 3	0.44	0.57	1.00	
Time 4	0.36	0.50	0.66	1.00

MNRM Time Invariant – socially desirable				
Time 1	1.00			
Time 2	0.51	1.00		
Time 3	0.49	0.63	1.00	
Time 4	0.43	0.55	0.67	1.00

MNRM Time Invariant – socially desirable				
Time 1	1.00			
Time 2	0.51	1.00		
Time 3	0.49	0.65	1.00	
Time 4	0.43	0.55	0.69	1.00

Table 5(b)

Self-management: Correlation Matrix for Latent SEL Factors by Model

No Correction				
Time 1	1.00			
Time 2	0.46	1.00		
Time 3	0.38	0.54	1.00	
Time 4	0.35	0.48	0.62	1.00

MNRM Time Invariant - extreme				
Time 1	1.00			
Time 2	0.48	1.00		
Time 3	0.40	0.56	1.00	
Time 4	0.37	0.49	0.64	1.00

MNRM Time Varying - extreme				
Time 1	1.00			
Time 2	0.51	1.00		
Time 3	0.43	0.60	1.00	
Time 4	0.39	0.51	0.67	1.00

MNRM Time Invariant – socially desirable				
Time 1	1.00			
Time 2	0.46	1.00		
Time 3	0.38	0.56	1.00	
Time 4	0.36	0.49	0.63	1.00

MNRM Time Invariant – socially desirable				
Time 1	1.00			
Time 2	0.47	1.00		
Time 3	0.39	0.56	1.00	
Time 4	0.36	0.50	0.64	1.00

Table 5(c)

Self-management: Correlation Matrix for Latent SEL Factors by Model

No Correction				
Time 1	1.00			
Time 2	0.55	1.00		
Time 3	0.38	0.64	1.00	
Time 4	0.38	0.52	0.65	1.00

MNRM Time Invariant - extreme				
Time 1	1.00			
Time 2	0.56	1.00		
Time 3	0.39	0.64	1.00	
Time 4	0.38	0.52	0.67	1.00

MNRM Time Varying - extreme				
Time 1	1.00			
Time 2	0.61	1.00		
Time 3	0.44	0.67	1.00	
Time 4	0.43	0.56	0.70	1.00

MNRM Time Invariant – socially desirable				
Time 1	1.00			
Time 2	0.55	1.00		
Time 3	0.39	0.64	1.00	
Time 4	0.39	0.53	0.67	1.00

MNRM Time Invariant – socially desirable				
Time 1	1.00			
Time 2	0.56	1.00		
Time 3	0.39	0.64	1.00	
Time 4	0.39	0.53	0.67	1.00

Table 6

Growth Parameter Estimates by Construct and Model

	Fixed Effects				Variance Components		
Model/Construct	Int.	Int. S.E.	Slope	Slope S.E.	Var. Int.	Var. Slope	Cov. Int./Slp.
Growth Mindset							
No response style	-0.006	0.012	0.112	0.004	0.464	0.028	0.032
MNRM time invariant (extreme)	-0.011	0.011	0.138	0.004	0.405	0.022	0.026
MNRM time invariant (socially desirable)	-0.008	0.012	0.136	0.004	0.466	0.033	0.049
MNRM time-varying (extreme)	-0.018	0.011	0.173	0.004	0.400	0.027	0.058
MNRM time-varying (socially desirable)	-0.009	0.012	0.137	0.004	0.457	0.032	0.053
Self-efficacy							
No response style	0.024	0.012	-0.115	0.004	0.444	0.031	-0.023
MNRM time invariant (extreme)	0.023	0.012	-0.118	0.004	0.449	0.030	-0.027
MNRM time invariant (socially desirable)	0.023	0.012	-0.118	0.004	0.030	0.454	-0.025
MNRM time-varying (extreme)	0.029	0.012	-0.113	0.004	0.472	0.029	-0.024
MNRM time-varying (socially desirable)	0.024	0.012	-0.120	0.004	0.030	0.451	-0.025
Self-management							
No response style	0.019	0.012	-0.093	0.004	0.501	0.037	-0.030
MNRM time invariant (extreme)	0.028	0.012	-0.088	0.004	0.482	0.035	-0.038
MNRM time invariant (socially desirable)	0.022	0.012	-0.088	0.004	0.035	0.505	-0.030
MNRM time-varying (extreme)	0.042	0.012	-0.069	0.004	0.526	0.032	-0.022
MNRM time-varying (socially desirable)	0.026	0.012	-0.087	0.004	0.036	0.505	-0.031

Table 7

Means and Variances of Scores by Timepoint from Simulations

Construct/Model	Score				Variance				Correlations					
Sim. 1 - Original Empirical	Time 1	Time 2	Time 3	Time 4	Time 1	Time 2	Time 3	Time 4	T1,T2	T1,T3	T1,T4	T2,T3	T2,T4	T3,T4
True Scores	0	0.122	0.340	0.505	1	1.523	1.858	1.826	0.491	0.466	0.423	0.646	0.547	0.651
GPCM (4D)	0	0.118	0.319	0.471	1	1.601	1.915	1.876	0.470	0.431	0.394	0.611	0.515	0.639
MNRM extreme (8D)	0	0.137	0.370	0.546	1	1.702	2.080	2.033	0.487	0.457	0.424	0.634	0.536	0.651
Sim. 2 - Extreme Empirical														
True Scores	0	0.122	0.340	0.505	1	1.523	1.858	1.826	0.491	0.466	0.423	0.646	0.547	0.651
GPCM (4D)	0	0.106	0.282	0.415	1	1.537	1.816	1.779	0.452	0.407	0.369	0.587	0.492	0.626
MNRM extreme (8D)	0	0.138	0.366	0.547	1	1.729	2.098	2.065	0.485	0.457	0.423	0.632	0.532	0.646

Table 8

Growth Parameters from the Simulation Studies

	Fixed Effects				Variance Components		
Simulation and Model	Int.	Int. S.E.	Slope	Slop.S.E.	Var. Int.	Var. Slope	Cov. Int./Slp.
Sim. 1 - Original Empirical							
True Score	-0.019	0.002	0.174	0.001	0.545	0.053	0.065
GPCM (4D)	-0.014	0.001	0.161	0.001	0.494	0.046	0.059
MNRM time-varying (8D)	-0.020	0.002	0.187	0.002	0.469	0.039	0.078
Sim. 2 - Extreme Empirical							
True Score	-0.019	0.002	0.174	0.001	0.545	0.053	0.065
GPCM (4D)	-0.009	0.001	0.141	0.001	0.488	0.049	0.043
MNRM time-varying (8D)	-0.023	0.002	0.188	0.002	0.464	0.038	0.078

Note. GPCM and MNRM estimates are averaged across the 100 replications.

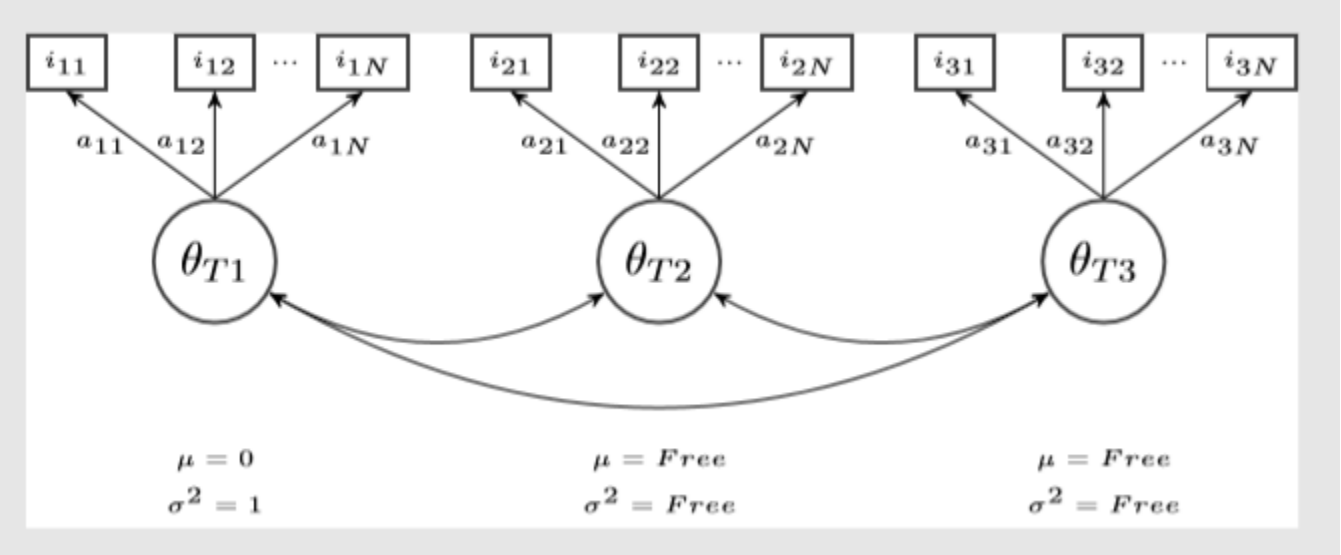

Figure 1. Path diagram for GPCM model used to score SEL item responses.

Appendix

Table A1

Items from District Surveys

Agree or disagree with the following (5 point Likert scale)
Growth Mindset
My intelligence is something that I can't change very much.
Challenging myself won't make me any smarter.
There are some things I am not capable of learning.
If I am not naturally smart in a subject, I will never do well in it.
Self-efficacy
I can earn an A in my classes.
I can do well on all my tests, even when they're difficult.
I can master the hardest topics in my classes.
I can meet all the learning goals my teachers set.
Self-management
I came to class prepared.
I remembered and followed directions.
I got my work done right away instead of waiting until the last minute.
I paid attention and resisted distractions.
I worked independently with focus.

Table A2

Generating and Average Estimated Item Slopes for MNRM and GPCM Conditions in Simulation 1

Condition	Item	a1	a2	a3	a4	a5	a6	a7	a8
True slopes	v1	0.32	0	0	0	1.561	0	0	0
	v2	0.446	0	0	0	1.497	0	0	0
	v3	0.719	0	0	0	1.468	0	0	0
	v4	0.648	0	0	0	1.484	0	0	0
	v5	0	0.32	0	0	0	1.561	0	0
	v6	0	0.446	0	0	0	1.497	0	0
	v7	0	0.719	0	0	0	1.468	0	0
	v8	0	0.648	0	0	0	1.484	0	0
	v9	0	0	0.32	0	0	0	1.561	0
	v10	0	0	0.446	0	0	0	1.497	0
	v11	0	0	0.719	0	0	0	1.468	0
	v12	0	0	0.648	0	0	0	1.484	0
	v13	0	0	0	0.32	0	0	0	1.561
	v14	0	0	0	0.446	0	0	0	1.497
	v15	0	0	0	0.719	0	0	0	1.468
	v16	0	0	0	0.648	0	0	0	1.484
GPCM slopes	v1	0.336	0	0	0	—	—	—	—
	v2	0.506	0	0	0	—	—	—	—
	v3	0.76	0	0	0	—	—	—	—
	v4	0.731	0	0	0	—	—	—	—
	v5	0	0.336	0	0	—	—	—	—
	v6	0	0.506	0	0	—	—	—	—
	v7	0	0.76	0	0	—	—	—	—
	v8	0	0.731	0	0	—	—	—	—
	v9	0	0	0.336	0	—	—	—	—
	v10	0	0	0.506	0	—	—	—	—
	v11	0	0	0.76	0	—	—	—	—
	v12	0	0	0.731	0	—	—	—	—
	v13	0	0	0	0.336	—	—	—	—
	v14	0	0	0	0.506	—	—	—	—
	v15	0	0	0	0.76	—	—	—	—
	v16	0	0	0	0.731	—	—	—	—
MNRM slopes	v1	0.307	0	0	0	1.572	0	0	0
	v2	0.426	0	0	0	1.502	0	0	0
	v3	0.688	0	0	0	1.469	0	0	0
	v4	0.62	0	0	0	1.487	0	0	0
	v5	0	0.307	0	0	0	1.572	0	0
	v6	0	0.426	0	0	0	1.502	0	0
	v7	0	0.688	0	0	0	1.469	0	0
	v8	0	0.62	0	0	0	1.487	0	0
	v9	0	0	0.307	0	0	0	1.572	0
	v10	0	0	0.426	0	0	0	1.502	0
	v11	0	0	0.688	0	0	0	1.469	0
	v12	0	0	0.62	0	0	0	1.487	0
	v13	0	0	0	0.307	0	0	0	1.572
	v14	0	0	0	0.426	0	0	0	1.502
	v15	0	0	0	0.688	0	0	0	1.469
	v16	0	0	0	0.62	0	0	0	1.487

Response Style and Growth Estimates

Note. Reported GPCM and MNRM estimates are averaged across 100 replications.

Table A3

Generating and Average Estimated Gamma Parameters for MNRM and GPCM Conditions in Simulation 1

	True (8D MNRM)				4D GPC				8D MNRM			
Item	g1	g2	g3	g4	g1	g2	g3	g4	g1	g2	g3	g4
v1	-0.025	1.246	0.337	-0.028	-0.031	0.836	0.321	-0.1	-0.027	1.253	0.337	-0.026
v2	0.39	0.185	0.239	-0.125	0.394	0.239	0.185	-0.12	0.388	0.183	0.237	-0.124
v3	0.291	1.879	0.022	0.088	0.283	1.571	-0.031	0.039	0.286	1.88	0.022	0.087
v4	0.752	0.763	0.129	-0.033	0.768	0.794	0.042	-0.028	0.748	0.767	0.131	-0.031
v5	-0.025	1.246	0.337	-0.028	-0.031	0.836	0.321	-0.1	-0.027	1.253	0.337	-0.026
v6	0.39	0.185	0.239	-0.125	0.394	0.239	0.185	-0.12	0.388	0.183	0.237	-0.124
v7	0.291	1.879	0.022	0.088	0.283	1.571	-0.031	0.039	0.286	1.88	0.022	0.087
v8	0.752	0.763	0.129	-0.033	0.768	0.794	0.042	-0.028	0.748	0.767	0.131	-0.031
v9	-0.025	1.246	0.337	-0.028	-0.031	0.836	0.321	-0.1	-0.027	1.253	0.337	-0.026
v10	0.39	0.185	0.239	-0.125	0.394	0.239	0.185	-0.12	0.388	0.183	0.237	-0.124
v11	0.291	1.879	0.022	0.088	0.283	1.571	-0.031	0.039	0.286	1.88	0.022	0.087
v12	0.752	0.763	0.129	-0.033	0.768	0.794	0.042	-0.028	0.748	0.767	0.131	-0.031
v13	-0.025	1.246	0.337	-0.028	-0.031	0.836	0.321	-0.1	-0.027	1.253	0.337	-0.026
v14	0.39	0.185	0.239	-0.125	0.394	0.239	0.185	-0.12	0.388	0.183	0.237	-0.124
v15	0.291	1.879	0.022	0.088	0.283	1.571	-0.031	0.039	0.286	1.88	0.022	0.087
v16	0.752	0.763	0.129	-0.033	0.768	0.794	0.042	-0.028	0.748	0.767	0.131	-0.031

Note. Reported GPCM and MNRM estimates are averaged across 100 replications.

Table A4

Generating and Average Estimated Item Slopes for MNRM and GPCM Conditions in Simulation 2

Condition	Item	a1	a2	a3	a4	a5	a6	a7	a8
True slopes	v1	0.32	0	0	0	3.122	0	0	0
	v2	0.446	0	0	0	2.995	0	0	0
	v3	0.719	0	0	0	2.936	0	0	0
	v4	0.648	0	0	0	2.969	0	0	0
	v5	0	0.32	0	0	0	3.122	0	0
	v6	0	0.446	0	0	0	2.995	0	0
	v7	0	0.719	0	0	0	2.936	0	0
	v8	0	0.648	0	0	0	2.969	0	0
	v9	0	0	0.32	0	0	0	3.122	0
	v10	0	0	0.446	0	0	0	2.995	0
	v11	0	0	0.719	0	0	0	2.936	0
	v12	0	0	0.648	0	0	0	2.969	0
	v13	0	0	0	0.32	0	0	0	3.122
	v14	0	0	0	0.446	0	0	0	2.995
	v15	0	0	0	0.719	0	0	0	2.936
	v16	0	0	0	0.648	0	0	0	2.969
GPCM slopes	v1	0.359	0	0	0	—	—	—	—
	v2	0.597	0	0	0	—	—	—	—
	v3	0.844	0	0	0	—	—	—	—
	v4	0.855	0	0	0	—	—	—	—
	v5	0	0.359	0	0	—	—	—	—
	v6	0	0.597	0	0	—	—	—	—
	v7	0	0.844	0	0	—	—	—	—
	v8	0	0.855	0	0	—	—	—	—
	v9	0	0	0.359	0	—	—	—	—
	v10	0	0	0.597	0	—	—	—	—
	v11	0	0	0.844	0	—	—	—	—
	v12	0	0	0.855	0	—	—	—	—
	v13	0	0	0	0.359	—	—	—	—
	v14	0	0	0	0.597	—	—	—	—
	v15	0	0	0	0.844	—	—	—	—
	v16	0	0	0	0.855	—	—	—	—
MNRM slopes	v1	0.305	0	0	0	3.144	0	0	0
	v2	0.423	0	0	0	3.001	0	0	0
	v3	0.685	0	0	0	2.939	0	0	0
	v4	0.616	0	0	0	2.968	0	0	0
	v5	0	0.305	0	0	0	3.144	0	0
	v6	0	0.423	0	0	0	3.001	0	0
	v7	0	0.685	0	0	0	2.939	0	0
	v8	0	0.616	0	0	0	2.968	0	0
	v9	0	0	0.305	0	0	0	3.144	0
	v10	0	0	0.423	0	0	0	3.001	0
	v11	0	0	0.685	0	0	0	2.939	0
	v12	0	0	0.616	0	0	0	2.968	0
	v13	0	0	0	0.305	0	0	0	3.144
	v14	0	0	0	0.423	0	0	0	3.001
	v15	0	0	0	0.685	0	0	0	2.939
	v16	0	0	0	0.616	0	0	0	2.968

Note. Reported GPCM and MNRM estimates are averaged across 100 replications.

Table A5

Generating and Average Estimated Gamma Parameters for MNRM and GPCM Conditions in Simulation 2

	True (8D MNRM)				4D GPC				8D MNRM			
Item	g1	g2	g3	g4	g1	g2	g3	g4	g1	g2	g3	g4
v1	-0.025	1.246	0.337	-0.028	-0.031	0.478	0.307	-0.161	-0.026	1.254	0.336	-0.026
v2	0.39	0.185	0.239	-0.125	0.406	0.316	0.131	-0.111	0.388	0.185	0.235	-0.122
v3	0.291	1.879	0.022	0.088	0.287	1.264	-0.087	-0.006	0.288	1.883	0.02	0.087
v4	0.752	0.763	0.129	-0.033	0.795	0.848	-0.046	-0.018	0.75	0.767	0.131	-0.029
v5	-0.025	1.246	0.337	-0.028	-0.031	0.478	0.307	-0.161	-0.026	1.254	0.336	-0.026
v6	0.39	0.185	0.239	-0.125	0.406	0.316	0.131	-0.111	0.388	0.185	0.235	-0.122
v7	0.291	1.879	0.022	0.088	0.287	1.264	-0.087	-0.006	0.288	1.883	0.02	0.087
v8	0.752	0.763	0.129	-0.033	0.795	0.848	-0.046	-0.018	0.75	0.767	0.131	-0.029
v9	-0.025	1.246	0.337	-0.028	-0.031	0.478	0.307	-0.161	-0.026	1.254	0.336	-0.026
v10	0.39	0.185	0.239	-0.125	0.406	0.316	0.131	-0.111	0.388	0.185	0.235	-0.122
v11	0.291	1.879	0.022	0.088	0.287	1.264	-0.087	-0.006	0.288	1.883	0.02	0.087
v12	0.752	0.763	0.129	-0.033	0.795	0.848	-0.046	-0.018	0.75	0.767	0.131	-0.029
v13	-0.025	1.246	0.337	-0.028	-0.031	0.478	0.307	-0.161	-0.026	1.254	0.336	-0.026
v14	0.39	0.185	0.239	-0.125	0.406	0.316	0.131	-0.111	0.388	0.185	0.235	-0.122
v15	0.291	1.879	0.022	0.088	0.287	1.264	-0.087	-0.006	0.288	1.883	0.02	0.087
v16	0.752	0.763	0.129	-0.033	0.795	0.848	-0.046	-0.018	0.75	0.767	0.131	-0.029

Note. Reported GPCM and MNRM estimates are averaged across 100 replications.