Understanding How Approaches to Calibrating and Scoring Survey Item Responses Affect Results from Growth Mixture Models

Overview

Understanding how children develop on psychological and social-emotional constructs is important to a range of long-term academic and economic outcomes (e.g., Heckman et al., 2006; Soland et al., 2013), and is therefore an emphasis in fields like education and psychology. In particular, knowing how to group children on the basis of that growth—that is, understanding if there are common growth profiles they fall into—is valuable in clinical, educational, and applied research contexts (Van Horn et al., 2009). For example, teachers or psychologists might want to know if particular developmental patterns of negative affect are more or less associated with substance abuse in adolescence (Curran & Hussong, 2003), or if there are patterns in how self-control develops during early childhood (Pan & Zhu, 2018). Such profiles are especially useful in an intervention context. For instance, an intervention to reduce antisocial behavior might look very different for students who show a sharp spike in such behaviors that tapers off compared to children who show steady increases in such behaviors (Walters & Ruscio, 2013).

This interest in understanding growth profiles and how to support the psychological/social-emotional development of students has translated methodologically into the widespread use of a class of models called growth mixture models (GMMs; Grimm & Ram, 2009; Muthén, 2003). Such models estimate growth, and use parameters from that model to identify latent classes of children based on their trajectories that might not otherwise be directly observable. The application of these models to understand social-emotional and psychological development yields many benefits. First is the ability to form meaningful groups of children on the basis of their development over time. For instance, in a study of school engagement, Zhen and colleagues (2020) identified a majority of their sample (71.24%) as persistently engaged, but also found three small groups with different levels of engagement over time. Second, inspecting the parameters of latent classes allows researchers to understand complex interactions between variables (Zhen et al., 2020). For instance, in the same study, the researchers were able to characterize the three small classes as having "climbing," "descending," and "struggling" groups on the basis of the classes' growth trajectories. Third, individual class assignments may be linked to predictors or outcomes, helping to identify vulnerable subgroups and define intervention targets.

One complication associated with estimating developmental growth, including for GMMs, is that they often model growth based on repeated administrations of surveys. These surveys can suffer from imperfections, especially related to forms of self-report bias. While use of GMMs in education and psychology has exploded over the last decade or so, little is known about how decisions related to scoring surveys used in the growth model—including using scoring approaches designed to mitigate issues related to self-report bias—affect the performance of GMMs. This omission occurs despite evidence that survey scoring decisions like whether and what type of measurement model to use can substantively affect basic inferences drawn from the growth models that underlie all GMMs. For example, research shows that approaches to scoring surveys that do not match the longitudinal nature of the data like using sum scores or unidimensional item response theory (IRT) models can, in some cases, result in estimated latent slopes and slope variances being understated by roughly half compared to the true parameters (Kuhfeld & Soland, 2020). Further, scores resulting from calibration/scoring models that better match the nature of the data, including longitudinal multidimensional IRT (MIRT) models, can recover growth parameters more effectively when used in a growth model. There is also evidence that forms of self-report bias like response style bias, which occurs when two respondents select different survey item response categories on an item despite having the same true score on the construct, can lead to improper inferences about growth (Soland & Kuhfeld, 2020). Yet, practically nothing is known about how such biases might affect the ability of GMMs to accurately identify latent classes. To address these limitations of how GMMs are currently used, we will investigate two overarching research questions:

1. How does the choice of the calibration/scoring model (sum score, unidimensional IRT, longitudinal MIRT) affect recovery of true classes from a growth mixture model?
2. Does response style bias affect recovery of true classes from a growth mixture model, and does accounting for that bias in the scoring model improve recovery?

In this proposal, we use these two questions as the backbone for comprehensive analyses examining how approaches to scoring surveys impacts recovery of true classes from GMMs. This research will rely on Monte Carlo simulations, as well as social-emotional learning (SEL) data from districts serving over 1.5 million students who took an SEL survey over the course of four years due to their district's participation in the California Office to Reform Education (CORE). To the extent possible, our research will represent a fairly exhaustive examination of how scoring decisions, including approaches to mitigating response style bias, impact the ability of GMMs to help educators and psychologists identify children in need of developmental interventions. Further, we will attempt to identify best practices for scoring surveys for use in GMMs, and give practitioners the tools they need to implement those practices. Given our desire to make the work useful to non-methodologists, we do not simply assume that the most complicated measurement models are the most appropriate. Rather, we envision and investigate three possible scenarios: (1) complex longitudinal measurement models are the generating model and are therefore most appropriate for scoring; (2) such models are the generating models, but much simpler models like sum scores prove sufficient anyway; and (3) simple measurement models like unidimensional IRT models generated the data, and using more complex models either does not improve results, or even leads to improper inferences due to overfitting. By examining all three possibilities, we will help researchers understand tradeoffs in decisions about how to score surveys for use in GMMs.

Intellectual Merit

This study will advance knowledge and understanding by examining how decisions about scoring surveys, including neglecting response style bias in scoring, affect recovery of latent classes from GMMs. Practically no research we are aware of examines this issue, which we believe is particularly concerning for two reasons. First, the two models GMM essentially combines – growth models and mixture models – have each been shown to be highly sensitive to scoring and measurement decisions. With respect to growth models, decisions in scoring (e.g., what type of IRT model to use) may lead to substantial changes in growth parameter estimates, including latent slope means and variances (e.g., Soland & Kuhfeld, 2020). With respect to mixture models, issues such as measurement non-invariance and ignored local dependence between items have been linked to over- and under-extraction of latent classes by simpler mixture models (e.g., latent profile and latent curve analysis; Oberski et al., 2013; Diallo et al., 2016; Masyn, 2017; Olivera-Aguilar & Rikoon, 2018; Cole, Bauer, & Hussong, 2019). Second, while measurement issues have not been studied specifically in the context of GMMs, GMMs have been shown to be highly sensitive to even mild departures from ideal conditions, including in ways that could be related to scoring. For instance, if the response distribution of the outcome is misspecified (e.g., assumed normal when it is skewed), the number of classes may be over- or under-estimated (Bauer & Curran, 2003; De Paoli et al., 2019). Despite these complexities that can affect results from GMMs, including the veracity of the results they produce, use of GMM techniques in education and psychology has increased substantially in recent times (e.g., Shore et al., 2018). The scoring issues we will investigate raise further questions about whether GMMs may be biased because suboptimal scoring procedures are introducing bias into underlying growth models.

To help close this gap in the literature, we will start to connect several related lines of research, including (a) GMMs (e.g., Muthén, 2003) and their sensitivity to modeling choices (e.g. De Paoli et al., 2019), (b) how scoring affects recovery of growth parameters (Bauer & Curran, 2015; Kuhfeld & Soland, 2020), and (c) how addressing biases like response style differences affect growth estimates (Soland & Kuhfeld, 2020). In short, the research we propose is designed to ensure that results from GMMs are

internally valid given how the surveys they often use are scored, which has large implications for the proliferation of clinical, educational, and applied research uses of such models. In addition to investigating questions which will bridge a gap between currently disjointed areas, the proposed work will also represent a methodologically rigorous synthesis of Monte Carlo simulation and empirical data analysis. Unlike many simulation studies, parameter values in the proposed work are drawn from actual observed values in a large study of children's social-emotional growth (Soland & Kuhfeld, 2020). Given that we plan to make annotated versions of all of the analysis code described below publicly available, our hope is that the careful methodological choices in the proposed research can serve as a template for other researchers who intend to use Monte Carlo simulations to address adjacent questions, as well as choose the most appropriate scoring approach to accompany GMMs used with empirical data in their own research.

Broader Impacts

This project is designed to support researchers and educators by improving our ability to understand how students develop, including developing profiles of growth that can be used to support interventions. For example, profiles from GMMs can be used to help understand how children develop self-control, and support students in that development (Pan & Zhu, 2018). We see two broad applications of our findings in applied contexts. First, we will supply applied researchers with best practices for scoring surveys for use in GMMs, including code to produce those scores. We plan to disseminate those best practices in a variety of ways, including trainings at conferences like the Society for Research in Educational Effectiveness (SREE), and writing a teacher's corner article (tutorial) on how to score surveys for use in mixture models. Second, we will use these results to raise awareness among educators and psychologists about how to interpret the growing number of studies using GMMs to improve their practice, and the importance of paying attention to scoring. One way to raise awareness is to use our professional networks. For example, Dr. Soland will use his University of Virginia affiliations with centers like Education Policy Works (EPW) and the Center for the Advanced Study of Teaching and Learning (CASTL) to share results with educators, psychologists, and policymakers who work with faculty at Virginia's School of Education and Human Development. Another is the creation of public-facing media about how to interpret mixture models. After results are obtained from the current studies, the Co-PIs will assemble a "Growth Mixture Model Measurement Checklist" to be hosted on a departmental web server. This resource will be free-of-charge, written to be accessible to applied researchers, with a checklist containing criteria pertaining to measurement-relevant information in the Methods and Results sections of applied GMM studies (e.g., Does the study report the frequencies for each item in a scale? Are sum scores used?). This checklist will help consumers of GMM studies determine whether and to what extent a study's results can be trusted given associated measurement decisions.

RATIONALE

In this section, we briefly review GMMs. We then discuss what is known about how scoring decisions impact estimates of students' psychological and social-emotional growth, including how accounting for self-report biases like response styles in the scoring models impact those growth estimates. We conclude by pointing out gaps in the literature that our studies will address. Note that we do not review literature on how mixture models have been used to improve IRT parameter estimation. While such literature is important and robust, we propose to use IRT-based methods to improve mixture model results, not the reverse (which is the focus of most related research).

Growth Mixture Models. GMMs are a subtype of mixture models, a broad designation of models which form categories of individuals based on some pattern of responses. Whereas some models classify individuals based on their profiles of responses at one point in time (e.g., latent class analysis or latent profile analysis), in a GMM each class is characterized by a trajectory of one or more variables over time

(Grimm & Ram, 2009; Muthen, 2003). These classes can be interpreted as meaningful categories, but they also allow researchers to form inferences about complex relationships among variables (Bauer & Shanahan, 2007; Sterba & Bauer, 2014). Class membership can also be linked to predictors and outcomes, allowing researchers to model complicated relationships between individuals' trajectories of change over time and external variables (Huang & Bandeen-Roche, 2004; Asparouhov & Muthen, 2014). Perhaps unsurprisingly, GMM's have thus enjoyed widespread use in the social sciences, with the search term "growth mixture model" returning 130 citations in Web of Science SSI in the past 5 years.

However, a large and growing body of research shows the results of mixture models generally, and GMM's specifically, to be highly sensitive to even mild departures from ideal conditions. Some of these issues are shared by mixture models in general. If the response distribution of the outcome is misspecified (e.g., assumed normal when it is skewed), the number of classes may be overestimated. Additionally, if direct effects of covariates on items are erroneously omitted, the recovery of true latent classes and the parameters characterizing each class will be compromised (Asparouhov & Muthen, 2014; Diallo, Morin, & Lu; Masyn, 2017; Cole, Hussong, & Bauer, 2020). Finally, with respect to GMM's specifically, if the nature of the growth trajectory is misspecified (e.g., assumed linear when it actually follows a curvilinear course), the number of classes may be over- or underestimated (Bauer & Curran, 2004; Bauer, 2007). Similarly, erroneously assuming residual variances to be homoscedastic across classes may lead to the overestimation of between-group differences (Enders & Tofighi, 2008). The identification of extraneous classes and the overestimation of differences between them are particular concerns because, among other dangerous possibilities, they may lead researchers to interpret classes as meaningful entities when they are in fact a product of noise. This may lead educators, psychologists, and policymakers to erroneously target groups of children for prevention and intervention efforts.

Calibration/Scoring and Recovery of Growth Parameters. Outside of mixture modeling, an ever-growing body of research indicates that how survey item responses are calibrated and scored can meaningfully impact estimates of growth. In particular, evidence continues to show that using sum scores relies on extreme assumptions that are often unjustified. For example, McNeish and Wolf (2020) showed that a sum score is equivalent to fitting an extremely constrained confirmatory factor analytic model. Such a model assumes that all the loadings across items are equal (and oftentimes set to one), and that the residual variances are also equal across items. Sum scores also require that the respondents complete all items, and make it difficult to examine measurement invariance across groups and timepoints. Perhaps unsurprisingly, sum score use can lead to severe bias in observed scores, which can in turn result in misclassification of respondents (e.g., psychological patients receiving diagnoses) (McNeish & Wolf, 2020).

There is also evidence that these assumptions grow in number and are more severe in the presence of repeated measures data. Bauer and Curran (2015) pointed out that, in addition to precluding testing for longitudinal measurement invariance, sum scores fail to account for within-person correlations of scores over time. As a result, estimates of growth parameters are likely to be biased, including downwardly biased estimates of variance in those parameters. In our own research (Kuhfeld & Soland, 2020), we conducted simulation studies, and showed that using sum scores in latent growth curve models downwardly biases estimates of latent slope means and variances, in some cases by roughly 50% of the true parameter values. Further, we found that using a longitudinal MIRT model does a superior job of recovering latent growth parameters not only relative to sum scores, but also to IRT models that calibrate item parameters based on the first timepoint then carry those parameter estimates forward when scoring later timepoints. The hypothesized reason for the superior performance of the MIRT model was that, in part by accounting for correlations of scores over time, it shrinks scores to a time-specific mean, allows one to relax longitudinal invariance assumptions, and improves reliability by pooling information over time.

Despite these problems with sum scores at a point in time and longitudinally, they are used with great frequency in psychological research. For example, Flake et al. (2017) reviewed a representative sample of articles (433 scales) published in the *Journal of Personality and Social Psychology* for validity evidence supporting their intended uses, and found that most used sum scores. Roughly half of the scales

included no citation to a prior validation study, "appearing to have been developed on the fly" (Flake et al., 2017, p. 374). Further, for half the scales, internal reliability was the only psychometric evidence provided and 19% of scales had no psychometric information whatsoever. The findings of Flake et al. (2017) have been found in other research, including work by Regier et al. (2013) and Weidman et al. (2017), who noted that "researchers frequently use scales that were not systematically developed… [T]he majority of scales used include only a single item, and had unknown reliability. Together, these tactics may create…conceptual inconsistency among measures of purportedly identical [constructs] across studies" (p. 267). Such studies indicate that, although scoring models designed to help recover latent growth parameters are available and effective, they are rarely used in practice.

IRT Models Designed to Address Response Style Bias. Scoring models can also be used to address some of the biases that affect self-reports post hoc. 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. 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 lead to misdiagnosing conditions, but also that these response styles are often related to background characteristics like education level and trait anxiety (Van Vaerenbergh & Thomas, 2013). This relationship with demographics means that biases are nonrandom by student subgroup, and can severely bias comparisons on the basis of those groups.

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 multidimensional nominal response models (MNRMs; 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, the MNRM is the only scoring approach that allows one to explicitly model different response styles *and* the construct of interest (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 in the methods section.

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 MIRT 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).

A handful of studies have examined response styles in a longitudinal context. Deng, McCarthy, Piper, Baker, and Bolt (2018) modeled responses to scales of positive and negative affect from 362 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) provided evidence that response styles have a large stable component, only a small part of which is associated with demographics (Weijters et al., 2010b). In our own research (Soland & Kuhfeld, 2020), we conducted empirical and simulation analyses in which we scored surveys using IRT models that do (MNRM) 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 IRT model used.

Gaps in the Current Research. GMMs provide a useful way to classify children on the basis of their psychological/social-emotional growth, including to better target interventions to children who are not showing normative development. However, GMMs are sensitive to a range of assumptions that could be impacted by scoring decisions (e.g., normality assumptions). Research further demonstrates that estimates from growth models upon which such mixtures are based are very sensitive to the scoring approach used (Kuhfeld & Soland, 2020), including whether such models account for biases common to self-report measures (Soland & Kuhfeld, 2020). Yet, the impact of scoring approach on recovery of true latent classes from GMMs has never been investigated. In the two studies we outline, we will begin to close this gap in the research, and help applied researchers identify best practices for scoring survey measures that will be used in GMMs. As we note below, both simulation studies will include related analyses with empirical data from a survey given to over a million students each year in California.

METHODS – GENERAL CALIBRATION AND SCORING SIMULATIONS (QUESTION 1)

In our first set of simulation studies, we will examine the effect of scoring model on recovery of true latent classes from a data-generating GMM. The general steps used in the simulations include (1) simulating longitudinal true scores using a true GMM, (2) generating observed item responses based on item parameters from empirical survey data, (3) calibrating/scoring those surveys using different approaches (sum, unidimensional IRT calibrated at Time 1, longitudinal MIRT, etc.), and (4) re-estimating GMMs to examine parameter recovery, including of true latent classes/class membership. Data-generating conditions for the GMM (step 1), item response generation (step 2), and calibration/scoring models (step 3) are shown in Table 1.

Note that, in all our simulations, we take a two-step approach of scoring item responses and then using them in GMMs. That is, we eschew a one-step approach in which item parameters in the measurement model are estimated jointly with the GMM itself. While such an approach may be feasible for simple measurement models, we do not believe it is feasible in terms of model convergence, software limitations, and computational burden when examining complex measurement models. For example, these GMMs could hypothetically be parameterized as a second-order growth model, consisting of a measurement model for each timepoint, then a latent growth curve model estimated based on those latent variables in the measurement model (e.g., Hancock et al., 2001). Theoretically, the growth and measurement parameters could be allowed to vary by class, yielding a second-order GMM with measurement, growth, and mixture components (Grimm & Ram, 2009). However, given the well-known issues with local maxima in the likelihood functions of even simple GMM's, it is likely computationally intractable to estimate class memberships in combination with both a longitudinal model and a measurement model with the complexity of those used here (Hipp & Bauer, 2006). Looking ahead to our response style studies, beyond computation time and model complexity, we are not aware of cases where models like the MNRM have been fit in SEM software like Mplus, let alone used as the basis for a GMM. All told, we are not trying to argue that a one-step approach is theoretically infeasible; however, given our intent to make results useful to non-methodologists, in our view, such models are difficult to specify and estimate within the constraints of typical software packages.

Table 1. Simulation Conditions for Both Research Questions

Scoring Simulations	**Response Style Simulations**
GMM Conditions	
Number of classes (Q = 2, 4) Class separation (entropy = .65, .8, .95) Sample size (N = 250, 500, 1000) Class proportions (equal, with 50% in each class for Q = 2 and 25% in each class for Q = 4; unequal, with 85%/15% for Q = 2 and 40%/30%/20%/10% for Q = 4)	
Measurement Conditions Related to Generating Item Responses	
Measurement model (sum, unidimensional IRT, longitudinal MIRT)	Measurement model (sum, unidimensional IRT, MNRM)
Difficulty parameters (all items have a range of thresholds representing low to high "difficulty"; all items are "easy" with most respondents using the top one or two response categories)	Difficulty parameters (all items have a range of thresholds representing low to high "difficulty"; all items are "easy" with most respondents using the top one or two response categories)
Survey length (5,10,15 items)	Survey length (5,10,15 items)
Number of response categories (2,3,5)	
Longitudinal measurement invariance (Holds for all items; does not for all items)	
	Type response style bias (ERS; SDR)
	Severity of response style bias (low/high values for the latent response style means/variances)
Measurement Conditions Related to Calibrating/Scoring Item Responses	
Calibrations/scoring model (sum, unidimensional IRT, longitudinal MIRT)	Calibrations/scoring model (sum, unidimensional IRT, MNRM)
Does/does not allows some item parameters to be invariant longitudinally (IRT models only)	
	Type response style bias modeled (ERS; SDR)
	Response style bias consistency over time (is/is not consistent within persons over time)

Step 1. Simulating Data Based on a True Growth Mixture Model. True scores will be generated for each of *N* individuals (*i* = 1, …, *N*) at *T* time points (*t* = 1, …, *T*), from a *Q*-class GMM (*q* = 1, …, *Q*). For individual *i* in class *q* at time *t*, the class-specific true score θ_{qit} will be calculated as:

$$\theta_{qit} = \eta_{0qi} + \eta_{1qi}(t-1) + \eta_{2qi}(t-1)^2 + \epsilon_{it} \quad (1)$$

Note that *t* – 1 is used to represent time because time is indexed at 0. The class-specific intercept, linear slope, and quadratic slope are denoted η_{0qi}, η_{1qi} and η_{2qi}, respectively. For each class, they can be expressed as an $N \times 3$ matrix denoted $\boldsymbol{\eta_q}$, whose columns contain intercepts, linear slopes, and quadratic slopes (in that order). The matrix follows a multivariate normal distribution with mean vector $\boldsymbol{\alpha_q}$ and covariance matrix $\boldsymbol{\psi_q}$. The person-specific, time-specific error term is denoted ϵ_{it}. For the whole sample, the errors can be gathered into an $N \times T$ matrix denoted $\boldsymbol{\epsilon}$, which follows a multivariate normal distribution with means equal to a vector of 0's and covariance matrix $\boldsymbol{\phi}$. The $\boldsymbol{\phi}$ matrix is diagonal, with diagonal elements ϕ_t representing each time-specific variance.

Denote each individual's class assignment κ_i. The probability that individual *i* will be in each class, denoted $P(\kappa_i = q)$, follows a multinomial distribution with class-specific logits γ_q:

$$P(\kappa_i = q)\frac{exp(\gamma_q)}{\sum_{p=1}^{Q} exp(\gamma_p)} \quad (2)$$

Then each individual's time-specific true score θ_{it} is the value of θ_{qit} corresponding to the class that person is in, i.e., $\theta_{it} = \theta_{qit}$, s.t. $\kappa_i = q$.

Different simulation conditions will be generated by altering the above parameter values. Values of $\boldsymbol{\psi_q}$ and $\boldsymbol{\phi}$ will be held constant across conditions, reflecting our lack of hypotheses about error and latent variable covariances. As a starting point, the number of timepoints will be held constant at *T* = 4 (though we may eventually vary timepoints dependent on what we find). As shown in Table 1, we will vary four factors across conditions: sample size *N*, number of latent classes *Q*, class separation, and proportion of the sample in each class. The values in Table 1 were determined on the basis of prior simulation literature, which provides realistic benchmarks for values of entropy (e.g., Cole, Bauer, & Hussong, 2019), sample size (e.g., Li & Hser, 2011; Diallo, Morin, & Lu, 2017), and number and size of classes (e.g., Tofighi & Enders, 2008; Shader & Beauchaine, 2021). Population values of $\boldsymbol{\alpha_q}$, $\boldsymbol{\psi_q}$, and $\boldsymbol{\phi}$ matrices will be drawn from Soland and Kuhfeld (2020). Soland and Kuhfeld (2020), which is also the source of item parameters in subsequent steps, focused on a latent curve model rather than one with multiple classes. Values of $\boldsymbol{\psi_q}$ and $\boldsymbol{\phi}$ from this latent curve model will simply be repeated across all classes (i.e., error and growth factors are homoscedastic across classes). The means of intercept, linear, and quadratic growth factors will be set to the means from Soland and Kuhfeld for the first class (i.e., $\boldsymbol{\alpha_1}$). Growth factor means for other classes (i.e., $\boldsymbol{\alpha_q}, q \neq 1$) will be set to yield varying degrees of class separation, as indexed by entropy (Celeux & Soromenho, 1996). For each combination of number of classes (*Q* = 2 or 4) and set of class proportions (50%/50% or 85%/15% for *Q* = 2; 25%/25%/25%/25% or 40%/30%/20%/10% for *Q* = 4), a different set of parameters $\boldsymbol{\alpha_q}$ will be required to produce a given value of entropy. Because we intend to simulate 3 entropy conditions (low, medium, and high), this will necessitate 12 different $\boldsymbol{\alpha_q}$ vectors. The values will be determined by grid search, so that for each condition's value of *Q*, we will estimate entropy for large-*N* models with different values of $\boldsymbol{\alpha_q}, q \neq 1$. We will then choose values which correspond most closely to the entropy benchmarks in Table 1.

Step 2. Use True Scores to Generate Observed Likert-scale Item Responses. These true scores will then be used to generate observed, Likert-style item responses. Data will be generated in several ways. First, we will assume that the true measurement model is either a sum score, or a basic/unidimensional IRT model with item parameters constrained equal across timepoints. Using these fairly simple models will allow us to see what happens to GMMs when we then (wrongly) fit more

complicated measurement models for calibration/scoring purposes in the next step. All item parameters for the simple IRT models will be based on the empirical data from the study, as well as from other related studies (e.g., Kuhfeld & Soland, 2020). Second, we will generate data using a longitudinal MIRT model. As previously discussed, research indicates that scoring surveys using such models does a better job of recovering true latent growth parameters when the scores are used in growth models (e.g., Kuhfeld & Soland, 2020). All item parameters for the longitudinal MIRT will come from Kuhfeld and Soland (2020). Whether using simple or more complex measurement models to generate observed Likert item responses, we will vary a range of relevant conditions (as shown in Table 1). For example, we will vary the number of response categories, the IRT-based "difficulty" of the items (and, in particular, include a condition in which respondents mainly use the top two response categories of a hypothetical Likert scale, which commonly occurs on surveys), and the discrimination parameters. We will also vary the length of the survey (5,10, and 15 items), and do so such that longer scales are more reliable. Finally, we will simulate measurement models for which longitudinal measurement invariance does, and does not, hold to examine effects of calibration and scoring decisions in the next step when there is noninvariance.

Step 3. Calibrate and Score the Simulated Item Response Data. We will use three approaches to calibrate/score generated item responses (sum, unidimensional IRT, and longitudinal MIRT). These three approaches mirror those used to generate the item responses. Thus, we can examine cases where the model used to calibrate/score is overly complex relative to the generating measurement model and vice-versa (as well as when the measurement model matches the data-generating model).

Sum Scores. Producing sum scores is quite straightforward and involves summing up all of the observed item responses at a given point in time. However, a problem with this approach is scale indeterminacy/comparability. That is, the sum scores and IRT-based estimates are not on the same scale (McNeish & Wolf, 2020). To avoid this issue, we will use a linking approach to place the sum scores on the same scale as the IRT-based scores. This approach was used to compare sum- and IRT-based scores in a pre/post treatment context (Gorter et al., 2015, 2016) and was detailed by Tong and Kolen (2007).

Unidimensional IRT Calibrated at Time 1. In this subsection, the unidimensional IRT model is described. For Likert-type items, item calibration and scoring can be accomplished using the graded response model or GRM (Samejima, 1969). Let there be $j = 1, \dots, n$ items and $i = 1, \dots, N$ individuals. Let the response from individual i to item j at timepoint t be y_{tij}, where y_{tij} has R response categories. It can be assumed that y_{tij} takes integer values from $(0, \dots, R-1)$. Let the cumulative category response probabilities be

$$\begin{aligned} P(y_{tij} \geq 1|\theta_i) &= \frac{1}{1+\exp[-(c_{j1}+a_j\theta_i)]} \\ &\vdots \\ P(y_{tij} \geq R-1|\theta_i) &= \frac{1}{1+\exp[-(c_{j,R-1}+a_j\theta_i)]} \end{aligned} \tag{3}$$

The category response probability is the difference between two adjacent cumulative probabilities

$$P(y_{tij} = r|\theta_i) = P(y_{tij} \geq r|\theta_i) - P(y_{tij} \geq r+1|\theta_i), \tag{4}$$

where $P(y_{tij} \geq 0|\theta_i)$ is equal to 1 and $P(y_{tij} \geq R|\theta_i)$ is zero. The item parameter a_j is the slope parameter describing the relationship between item j and the latent factor and $b_{j1}, \dots, b_{j,R-1}$ are a set of $R-1$ (strictly ordered) parameters. The thresholds denote the point on the latent variable separating category r from category $r+1$.

In the unidimensional case, the logit in Equation 1 can be re-expressed in a more convenient slope-threshold form as $c_{jr} + a_j\theta_i = a_j(\theta_i - b_{jr})$, where $b_{jr} = -c_{jr}/a_j$ is the threshold (also referred to as severity or difficulty) parameter for category r. The rth threshold denotes the point on the latent

variable separating category r from category $r+1$. However, the slope-threshold form does not generalize well to multidimensional models, so we adopt the slope–intercept parameterization here and for all remaining IRT models presented.

*Longitudinal MIRT Model***.** The longitudinal MIRT model was designed to estimate latent change across time in an IRT framework. Item response data from each timepoint are combined and calibrated simultaneously across the *T* timepoints. Items are calibrated for a cohort of study participants such that each timepoint has its own latent variable estimate. In the current study, the model is operationalized using a multidimensional extension of the GRM. Let the cumulative category response probabilities be

$$\begin{aligned} P(y_{tij} \geq 1 | \boldsymbol{\theta_i}) &= \frac{1}{1+\exp[-(c_{j1} + \boldsymbol{a_j'\theta_i})]} \\ &\vdots \\ P(y_{tij} \geq R-1 | \boldsymbol{\theta_i}) &= \frac{1}{1+\exp[-(c_{j,R-1} + \boldsymbol{a_j'\theta_i})]} \end{aligned} \quad (5)$$

As with the unidimensional model, the category response probability is the difference between two adjacent cumulative probabilities. The difference between the unidimensional and multidimensional GRM is that $\boldsymbol{\theta_i}$ is now a *T*×1 vector of latent traits and $\boldsymbol{a_j'}$ a vector of slope parameters.

Given the same items are repeated across timepoints, a set of equality constraints are included for the item parameters of the repeated items. Let $\varphi_{tj}=\{\boldsymbol{a_{tj}'}, c_{tj1}, \ldots, c_{tj,R-1}\}$ be the vector of item parameters for item *j* observed at the first time point ($t=1$). One can assume that $\varphi_{1j} = \varphi_{2j} = \cdots = \varphi_{Tj}$, where φ_{Tj} is the item parameter vector for item *j* measured at time T. The first latent dimension is often (but not always) constrained for identification purposes to follow a standard normal distribution $\theta_{T1} \sim N(0,1)$, and the mean, variance, and covariance of the other latent factors are freely estimated relative to the first time point. Thus, unlike sum score and unidimensional IRT approaches, the MIRT approach explicitly accounts for the over-time correlations in the model.

MIRT calibration/scoring will be conducted in FlexMIRT. While the unidimensional models will be estimated using maximum likelihood via the Bock-Aitken EM algorithm, the MIRT models will be estimated via the Metropolis-Hastings Robbins-Monro (MH-RM) algorithm (Cai, 2010b, 2010a). (To ensure results are not being driven by such differences, we will also try to do sensitivity analyses in which all models are estimated similarly.) Estimates of the person-level scores will be produced using EAP scoring (Bock & Mislevy, 1982).

*Empirical Histogram Sensitivity Analysis***.** One potential issue with these latent variable scoring approaches is that they tend to assume the latent variables are normally distributed. However, GMMs represent a mixture of different normal distributions from different groups. Thus, assuming normality amounts to a misspecification of the true underlying latent variable in the population. To address this issue, as a sensitivity analysis, we will also produce scores using an empirical histogram approach that does not assume normality of the underlying latent variables (to the extent possible given modeling and computational constraints).

Step 4. Re-estimating the Growth Mixture Model Using Various Scores. Finally, after producing scores using the approaches detailed in the previous step, we will re-estimate the growth mixture model to see how well true latent classes and related parameters are recovered. Specifically, after scores are generated, Mplus version 8.4 (Muthen & Muthen, 2018) will be used to estimate GMM's using these scores as indicators. We will first assess the accuracy of class enumeration. For each replication, we will fit models with different numbers of classes, with *Q* ranging from 1 to 5. The number of classes favored by the Bayesian Information Criterion (BIC; Schwarz, 1987) and Akaike Information Criterion (AIC; Akaike, 2000) will each be recorded. For these indices, which seek to both optimize fit and parsimony, lower values indicate a better model. We will also consult the Lo-Mendell-Rubin Likelihood Ratio Test (VLMR; Vuong, Lo, Mendell, and Rubin, 2001) and Bootstrap Likelihood Ratio test (BLRT; McLachlan & Peel, 2000). These likelihood ratio tests compare the likelihood of a model with *Q* classes

with that of a model with Q – 1 classes; a significant value indicates that the Q-class model fits better than the Q – 1-class model. When comparing mixture models with different numbers of classes, the favored number of classes is the last value of Q for which the test is significant. For each fit index and likelihood ratio test, we will record the percentage of times the correct number of classes is chosen.

Second, for the correct model with the correct value of Q for a given cell, we will assess the accuracy of the estimated parameters defining the latent classes. This includes logits corresponding to class membership, which dictate the overall prevalence of each class, as well as the growth parameters (i.e., means and variances of intercepts and slopes) within each class. For each simulation condition, relative bias (i.e., the difference between the estimated and true values, divided by the true value) will be calculated along with the standard deviation of parameter estimates across replications. Finally, we will assess the accuracy of individual classifications. Each individual's most likely class assignment will be compared against their true class assignment using the Adjusted Rand Index (ARI; Hubert & Arabie, 1985). The ARI, which ranges from 0 to 1, is highest if two classification schemes agree; unlike Cohen's kappa, it adjusts for random chance. The ARI is independent of class labels, but we will control label-switching in class membership using the methods outlined by Tueller, Drotar, and Luebke (2011).

METHODS – RESPONSE STYLE SIMULATIONS (QUESTION 2)

Whereas the first simulation study examined the impact of general scoring decisions on recovery of true latent classes, the second simulation study will examine how much using scoring to address self-report bias post hoc impacts recovery of true latent classes. Specifically, we will examine how accounting for response style bias using an appropriate IRT model affects growth mixture results. This study will follow the steps used in the first study almost exactly. The only exception is that, rather than use a longitudinal MIRT model to score generating item responses, we will instead score item responses using the MNRM, which can be used specifically to address response style bias. Below, we provide detail on the MRNM.

*MRNM IRT Model to Address Response Style Bias***.** 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 $r = 1, \dots, R_j$ 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

$$P(Y = k\ |\mathbf{x}, \tilde{\mathbf{a}}, \mathbf{c}) = \frac{\exp\left(\tilde{\mathbf{a}}_r'\mathbf{x} + c_k\right)}{\sum_{m=1}^{R} \exp(\tilde{\mathbf{a}}_m'\mathbf{x} + c_m)} \qquad (6)$$

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

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

$$P(Y = r\ |\mathbf{x}, \mathbf{a}, \mathbf{S}, \mathbf{c}) = \frac{\exp\left([\mathbf{a} \circ \mathbf{s}_r]'\mathbf{x} + c_k\right)}{\sum_{m=1}^{R} \exp\left([\mathbf{a} \circ \mathbf{s}_m]'\mathbf{x} + c_m\right)} \qquad (7)$$

where $[\mathbf{a} \circ \mathbf{s}_r]$ is the Schur product of the slopes, $\mathbf{a}$, and $\mathbf{s}_r$ is a scoring function. $\mathbf{s}_r$ is part of a D x R 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 parallels the model parameterization used in FlexMIRT, the software we will 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 $s_{d,r}$ corresponding to row d and column r of **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 provided evidence that the best scoring function for 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. Nonetheless, as sensitivity analyses, we will also examine cases where respondents select either the top category or top two categories.

GMM Conditions. The GMM simulation conditions will be the same as in the first simulation.

Measurement Model Conditions. As shown in Table 1, most conditions in these simulations will mirror those from the more general scoring simulations. For example, when generating observed item responses, we will vary the measurement model (sum, IRT, and MNRM), the item parameters (difficulty/discrimination), and the survey length/reliability. Also, in parallel with the general scoring simulations, we will vary the type of measurement model used to calibrate and score, as well as assumptions about longitudinal invariance.

A key difference in this study is that we will also vary parameters relevant to the response style portion of the model. When generating the item responses, we will vary the degree and severity of the response style bias in the data-generating model. As detailed above, the MNRM includes latent variables for both the construct of interest and response style. Thus, we can vary the mean and variance of the response style latent variable to examine cases where the bias and its variability differs, as well as the item parameters associated with the response style factor. Further, we will take this approach for at least two types of response styles, including ERS and SDR. Finally, we will have some conditions in which individual respondents are consistent in how their response style manifests over time, and others in which it is inconsistent (e.g., letting means and variances of the response style latent variable, shift over time; Soland & Kuhfeld, 2020).

We will also vary the model we use to calibrate/score the item responses. For example, in addition to seeing how GMMs are affected when we ignore response style bias, we will also examine cases where there is no response style bias, but we use the MNRM and assume that bias exists anyway. Further, we will consider a scenario in which response style bias is present, but we model the wrong type of bias. For instance, we will generate the item responses using an ERS model, but score them using an SDR model. All of these scenarios could easily play out in practice, and potentially impact GMM results.

METHODS – EMPIRICAL ANALYSES

Beyond simulating data, we want to examine these issues when using empirical data for which we do not know the data-generating model. Thus, we will conduct empirical analyses that parallel and extend those from both sets of simulation studies. To that end, we have data from an SEL survey administered to districts serving more than 1.5 million students in California. We describe our sample, measures, and analytic approach below.

Sample. The data we have access to come from the CORE districts (Fresno, Garden Grove, Long Beach, Los Angeles, Oakland, Sacramento, San Francisco, and Santa Ana) and include 4th- through 12th-grade students who took the survey at least once during the 2014-15 through 2017-18 school years,

resulting in a sample of 361,815 students (the sample is under one million because we want to follow cohorts of students longitudinally). Using these data, we can follow individual students over time for up to four years. Specifically, we have six non-intact cohorts (students who were in Grades 4-9 in the first year, 2015, and thus had four years of data) with sizes ranging from ~55,000 to ~65,000 students each. We will use these data to examine SEL outcomes longitudinally, including via growth mixture model. Data included represent considerable diversity: ~75% of students are Latinx, ~9% are black, and ~25% are low-income.

Measures. The CORE districts' SEL survey comprises a battery of items designed to measure four SEL constructs: self-management (5 items), social awareness (4 items), growth mindset (4 items), and self-efficacy (4 items). Students in Grades 4 through 12 rate themselves on the same items using a 5-point Likert scale. These measures are supported by a bevy of evidence for their intended use with their intended population (e.g., Soland, 2018; West et al., 2018). For our own study, we will focus on three of the four constructs that correspond to intrapersonal outcomes. Those three SEL constructs are defined as follows:

- ***Self-management***, also referred to as self-control or self-regulation, is the ability to regulate one's emotions, thoughts, and behaviors effectively in different situations. This includes managing stress, delaying gratification, motivating oneself, and setting and working toward personal and academic goals (CASEL, 2005).
- ***Growth mindset*** is the belief that one's abilities can grow with effort. Students with a growth mindset believe that they can develop their skills through effort, practice, and perseverance. These students embrace challenges, see mistakes as opportunities to learn, and persist in the face of setbacks (Dweck, 2006).
- ***Self-efficacy*** is the belief in one's ability to succeed in achieving an outcome or reaching a goal. Self-efficacy reflects confidence in the ability to exert control over one's own motivation, behavior, and environment and allows students to become effective advocates for themselves (Bandura, 1997).

Analyses. We will use these data for several purposes. As a starting point, we will replicate the analyses in our simulations. That is, we will score surveys using the models described in Table 1, fit GMMs using those scores, then see how sensitive latent classes are to those decisions. While we cannot know the data-generating mechanism here, we can nonetheless see if findings are consummate with those from the simulations.

However, given the extraordinary size and representativeness of this sample (at least for the state of California), we will try to leverage the data to go beyond what was already done in the simulations. First, we will provide basic descriptive statistics that are relevant to helping researchers understand the issues we attempt to address in our study. For example, we will document how often students exhibit certain response patterns germane to response style bias, such as selecting the top and bottom response categories. Similarly, we can examine whether these response patterns are consistent for a given student over time. While such descriptives are not necessary indicative of response style bias, they nonetheless shed light on the pervasiveness of such patterns and do not typically exist on a large scale for SEL.

Second, we will document whether there are differences in the sensitivity of GMMs by construct measured. There is already some preliminary evidence that response styles differ for the same student by construct, including how those styles manifest over time for a given student (Soland & Kuhfeld, 2020). Thus, we can see if there is a single, consistent story to be told, or if it is dependent on whether one is measuring, say, self-efficacy versus social awareness. Such differences are more difficult to simulate given subtleties in how various constructs manifest themselves and relate to one another.

Third, we can examine what practical issues emerge when trying to fit some of the more complicated models. For example, there may be convergence issues or identification problems associated with models like the longitudinal MIRT and MNRM, especially when the latter is estimated for several timepoints. For example, an MNRM with four timepoints would result in an eight-dimensional IRT

model given each timepoint has a separate SEL and response style factor. To the extent we encounter such issues, we will document them and try to help other researchers avoid related pitfalls.

SIGNIFICANCE & IMPLEMENTATION

Significance. Profiles of how students develop psychologically and socio-emotionally are fundamentally important not only to understanding that development, but also to intervening when students are not showing appropriate development, including when such patterns may lead to negative outcomes like substance abuse or dropping out of school. GMMs offer a uniquely powerful way to detect these profiles. Yet, little is known about how decisions related to scoring survey responses affect results from GMMs, including the internal validity of the results they produce. This gap in the literature occurs despite two related fields of research suggesting that one should worry about scoring issues in a GMM context. First, there is evidence that GMMs can produce results that misrepresent latent class structures under even mild violations of assumptions, including assumptions that could be affected by scoring decisions. For example, GMMs can be biased due to faulty assumptions related to normality and growth model parameters, both of which could be impacted by scoring decisions. Second, there is expanding evidence that scoring decisions can have extreme effects on estimates of parameters from the growth models that underly GMMs (e.g., Kuhfeld & Soland, 2020), which could in turn impact recovery of true latent classes.

These biases risk leading researchers in child development astray. For instance, in GMM results, sometimes classes characterized by ascending or descending levels of a given behavior over development are of particular interest to researchers. In such instances, high or low classes may inform interventions given their members may be considered vulnerable (e.g., children who start off with high levels of self-efficacy, but show decreases across childhood) or susceptible to positive change (e.g., children who start with low levels of self-efficacy that increase). However, a heretofore unexamined possibility is that these apparent changes over time are actually an artifact of response styles and other sources of measurement bias. A child with an extreme response style may, for instance, give uniformly low responses in the first wave of measurement and high responses in subsequent waves, giving the erroneous impression of a high level of change.

In the two studies proposed, we will bridge these methodological literatures to investigate how scoring decisions affect results from GMMs. The first study will examine whether fitting a model designed to help recover latent growth parameters affects recovery of latent classes compared to using cruder scoring approaches like unidimensional IRT models and sum scores. The second study will examine whether response style bias might affect latent class recovery from GMMs, and if IRT scoring approaches designed to address such biases improve recovery. Both simulation studies will also incorporate analyses with empirical data using a very large sample of students who took frequently used surveys of constructs like self-efficacy. While these studies are methodological in nature, our overarching purpose is to ensure that applied researchers employing GMMs are using them in a way that produces results that are internally valid and, more importantly, can benefit educators and clinicians who use such studies to help improve outcomes for children.

Evaluation Plan and Advisory Panel. Four nationally prominent researchers will serve on our advisory panel, providing feedback on our research design and analytic strategies: Dr. Daniel Bauer (UNC), Dr. Karen Schmidt (UVA), Dr. Megan Kuhfeld (NWEA), and Dr. Andrea Hussong (UNC). Bauer is a Professor and Director of the Quantitative Psychology Program and L.L. Thurstone Psychometric Laboratory of the Department of Psychology and Neuroscience at UNC Chapel Hill. He is an expert both in GMMs and how scoring decisions can affect results from statistical models, including growth curve models. Schmidt is an Associate Professor in the Quantitative Psychology program at UVA. Her areas of expertise include modeling the latent structure of complex abilities across the lifespan and methodological investigations around item response theory models for understanding change. Kuhfeld is a Senior Research Scientist at NWEA. Her work covers a range of topics, including IRT, longitudinal growth modeling, and summer learning loss, all of which are used to provide insights that support the

academic development of students in grades K-12. Hussong is a Professor in the Department of Psychology and Neuroscience at UNC Chapel Hill, as well as a developmental scientist and licensed Clinical Psychologist dedicated to promoting health and well-being in children, youth, and families.

In fall of 2022, all four panelists will provide feedback on the simulation conditions, study design, and empirical analyses for the first study. During spring of 2023, feedback will be provided on results from that study and help provided on addressing any unforeseen challenges that arise. In fall of 2023, panelists will provide feedback on the same factors for the second study, and on results for the second study in spring of 2024. Finally, at the beginning of the third year (fall of 2024), all four scholars will provide suggestions on how best to disseminate results to the field. While all three panelists will provide input from a methodological perspective, we will additionally rely on Dr. Hussong's judgment about how such models are used to inform clinical practice, what the implications of our results might be, and how best to share them with the clinical field.

Anticipated Products/Communication Plan. Results will be disseminated in several ways. First, we plan to publish results in peer reviewed journals with audiences in psychology (e.g., *Psychological Methods*) and education (e.g., *The Journal for Research in Educational Effectiveness*). As part of that publication strategy, we will also write a tutorial for publication on how to implement the methods we employ in the studies to score surveys for use in GMMs. Second, we will present at conferences, including those run by the American Psychological Association, Society for Research on Child Development, and Society for Research in Educational Effectiveness. If feasible, we will present our findings as part of a training session at one or more of the aforementioned conferences. Finally, we will write research briefs intended for consumption by educators and psychological clinicians. Those briefs will be disseminated using extensive networks of practitioners maintained at UVA, UNC, and NWEA.

Personnel. Dr. James Soland will act as a PI. He is an Assistant Professor of Research, Statistics, and Evaluation at the University of Virginia, and is an Affiliated Research Fellow at NWEA, a non-profit assessment organization. As a psychometrician and applied statistician, he will serve in a lead role for all analyses alongside Dr. Veronica Cole, including scoring various measures used as outcomes/covariates and estimating all models.

Dr. Veronica Cole will act as a Co-PI. She is an Assistant Professor of Psychology at Wake Forest University. As a quantitative and developmental psychologist with expertise in mixture models, Dr. Cole will generate data from mixture models for all Monte Carlo simulations, help fit scoring models, and manage simulation and empirical data alongside Dr. Soland.

Managerial Arrangements. The University of Virginia will be the primary sponsor for this project and it will be led by Dr. James Soland. In particular, Soland will plan and facilitate project team meetings, ensure that deadlines are met, oversee the project budget. These managerial arrangements will be supported by considerable infrastructure for such purposes at the School of Education and Human Development at UVA, as well as the Center for Advanced Study of Teaching and Learning (CASTL) at UVA. In terms of analyses, all work will be shared jointly by Soland and Cole, both of whom have considerable experience conducting related studies and managing work on methodological grants.

Prior NSF Support. Soland does not have prior NSF support. Cole does not have prior NSF support.