Evaluating Anchoring Methods to Analyze Longitudinal Data with Item Response Models

Tara Lucile Valladares
Purcellville, Virginia

B.A. in Psychology, University of Virginia, 2016

A Thesis presented to the Graduate Faculty
of the University of Virginia in Candidacy for
the Degree of Master of Arts

Department of Psychology

University of Virginia
November, 2019

Evaluating Anchoring Methods to Analyze Longitudinal Data with Item Response Models

Tara L. Valladares
University of Virginia

Author Note

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. 1842490. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

Evaluating Anchoring Methods to Analyze Longitudinal Data with Item Response Models

Tara Lucile Valladares
Purcellville, Virginia

B.A. in Psychology, University of Virginia, 2016

A Thesis presented to the Graduate Faculty
of the University of Virginia in Candidacy for
the Degree of Master of Arts

Department of Psychology

University of Virginia
November, 2019

Abstract

Ordinal data is ubiquitous in psychological research, but it presents unique challenges in longitudinal analysis. Presently available longitudinal item response models (IRMs) can be computationally prohibitive for large, multiwave datasets, while lower computational alternatives may not produce useful estimates. Longitudinal anchoring is a possible solution to these issues. By anchoring separate IRMs together, person and item estimates can be obtained without limiting the number of timepoints that can be analyzed. A simulation study examining the performance of longitudinal anchoring was conducted. Six anchoring methods were evaluated: Floated, All Times, Time One, Mean, Random, and Cross-sectional. The results suggest that the Mean and the Cross-sectional anchoring methods performed the best. While the Time One, Random, and Floated methods produced similar item and person estimates, model fit was poor. The All Times method should be avoided as it cannot produce reliable change estimates. Longitudinal anchoring is an easily implemented solution when analyzing large, complex longitudinal datasets and shows promise as a low-computation method of producing latent trait estimates.

Contents

Abstract 2

Introduction 5

The Use of Questionnaires in Longitudinal Studies 5

Analysis of Ordinal Data 8

Item Response Models 10

Scaling Ordinal Data to Interval 10

The Rasch Model 13

Anchoring 15

Uses of the Rasch Model 16

Software and Estimation 18

Change Over Time and IRMs 18

Virtual Persons and Items 21

Longitudinal Anchoring 22

Anchoring Methods 25

Floated analyses 26

All Times Anchored Models 26

Mean Anchored Models 28

Time One Anchored Models 29

Random Anchoring 29

Cross-sectional Anchoring 30

General Comments on Anchoring Methods 32

Simulation 34

Conditions 34

Process 35

Evaluation 36

Accuracy of the Person Parameters 36

Accuracy of the Item Parameters 37
Slope and Variance of the Slope 38
Model Fit 39

Simulation Results **40**
Person Ability Estimates 40
Standard Error of the Estimate 40
Item Parameter Recovery 44
Slope and Slope Variance 48
Overall Model Fit 50

Discussion **57**
Future Directions 60
Conclusions 61

References **63**

Evaluating Anchoring Methods to Analyze Longitudinal Data with Item Response Models

It is nearly impossible to find a substantial longitudinal dataset the does not use scales or questionnaires. The use of scales and questionnaires is ubiquitous in psychological research, yet proper analysis and interpretation the data that produced is not always conducted. In many contexts, item response modeling is the method of choice for ordinal data. Sum-scored questionnaire data can produce biased model parameters or facetious results, especially when working with longitudinal data (Bereiter, 1963; Embretson, 1991, 1992; Glas, Geerlings, van de Laar, & Taal, 2009; Gorter, Fox, & Twisk, 2015; Pastor & Beretvas, 2006). For example, ordinal data can produce spurious interaction terms when analyzed using ANOVA techniques (Davison & Sharma, 1990; Embretson, 1996). In addition, change score reliability calculations are fraught with problems such as reliability coefficients with negative signs (see Bereiter, 1963, for a discussion).

The currently available methods for analyzing longitudinal scale data are more limited and less feasible than their single timepoint counterparts when considering issues such as parameter estimation, flexibility, and computational ease. With respect to item response models (IRMs), there are even fewer available longitudinal IRMs appropriate for large, complex analyses. This issue becomes more poignant as new, immensely large longitudinal research studies come to fruition. In the present project, I use simulation to examine several methods of analyzing longitudinal data by anchoring separate item response models together. Longitudinal anchoring shows promise as a low-computation method of producing latent trait estimates from longitudinal data.

The Use of Questionnaires in Longitudinal Studies

The use of surveys and questionnaires is widespread in longitudinal research. Abstract traits such as anxiety, personality, and social functioning are not easily assessed without self-report or rating scales. A researcher can count the number of alcoholic drinks a participant consumes each week, but they cannot use a ruler to

measure psychological dependence on alcohol. The lack of specific, objective measurement such as with thermometers, spectrometers, or rulers creates statistical and measurement quandaries not found in other fields. The use of surveys and questionnaires in longitudinal studies is therefore a reflection of the nature of psychological research. Below, some examples of large longitudinal studies and their use of questionnaires are detailed.

The Minnesota Twin Family Study (MTFS) is a longitudinal study focusing on the etiology of substance abuse disorders within twin pairs and their parents (Iacono, Carlson, Taylor, Elkins, & McGue, 1999; Iacono & McGue, 2002). Beginning in 1991, the study now includes over 2,700 twin pairs. Various scales on academic achievement, psychopathology, personality, and social functioning are used, as well as genetic and physiological information. While the original focus of the study was substance use, there are numerous publications based on the MTFS that examine other facets such as religiosity, personality, and psychopathology development (e.g., Blonigen, Hicks, Krueger, Patrick, & Iacono, 2005; Ehringer, Rhee, Young, Corley, & Hewitt, 2006; Hopwood et al., 2011; Koenig, McGue, & Iacono, 2008; Waldron, Malone, McGue, & Iacono, 2018).

The Longitudinal Studies of Child Abuse and Neglect (LONGSCAN) is consortium of studies concerned with the causes and effects of childhood mistreatment (Runyan et al., 1998). Low and high risk children and their guardians were recruited at multiple sites across the United States. Running from 1991 to 2011, approximately 1,400 families were included. The study included scales of cognitive ability, social functioning, psychopathology and distress, family dysfunction, and externalizing behavior. The LONGSCAN data has been used in various publications focusing on effects of childhood abuse on physical health, substance abuse, aggression, psychopathology, and suicidality (e.g., Flaherty et al., 2006; R. M. Johnson et al., 2002; Kotch et al., 2008; Lewis et al., 2019; Litrownik, Newton, Hunter, English, & Everson, 2003; Thompson et al., 2005).

The Fragile Families and Child Wellbeing Study (FFCWS) focuses on the trajectories of children of unwed parents (Reichman, Teitler, Garfinkel, & McLanahan,

2001; Waldfogel, Craigie, & Brooks-Gunn, 2010). Nearly 5,000 families have participated in the FFCWS since 1998 starting at the birth of their first child. The study is still ongoing today. The FFCWS includes many scales measuring children's cognitive ability, language skills, and social behaviors, as well as parental disciplinary strategies. Research using the FFCWS has focused on various aspects of fragile families and child development, including the effects of corporal punishment and parental incarceration on child well-being and factors that encourage stable parental relationships (e.g., Carlson, McLanahan, & England, 2004; MacKenzie, Nicklas, Brooks-Gunn, & Waldfogel, 2011; Schneider, MacKenzie, Waldfogel, & Brooks-Gunn, 2015; Turney & Wildeman, 2015; Wildeman, 2010).

The Notre Dame Study of Health & Well-being (NDHWB) is a presently running longitudinal study funded by the National Institutes of Health (Bergeman & Deboeck, 2014). The purpose of the NDHWB is to understand the pathways that lead to healthy aging with a focus on stress and resilience. The NDHWB administers scales on topics such as stress, physical and mental health, sleep quality, cognitive functioning, and mood. Uniquely, the NDHWB has a multiple-time scale design, where data is collected at yearly intervals and for bursts of 10 and 56 consecutive days. In particular, the 56 day bursts of data allow researchers to examine cycles that would otherwise be impossible to uncover at yearly or monthly intervals. Yet, 56 timepoints is far past the limit that many longitudinal models can realistically handle. One 56 day dataset from the NDHWB and the study by Erbacher, Schmidt, Boker, and Bergeman (2012) that utilized it is the impetus of the present study.

Scale data forms the foundation of most major longitudinal studies in psychology, and use of ordinal data in psychology and related fields is extensive (Cliff, 1989; Harwell & Gatti, 2001). Yet, not all data is created equal nor can it be treated equally. In particular, scales and questionnaires often produce *ordinal* data, which possesses unique qualities separate from standard interval data. These peculiarities can cause issues in data analysis; many of the common traditional statistical techniques rely on assumptions that are often only satisfied by interval data.

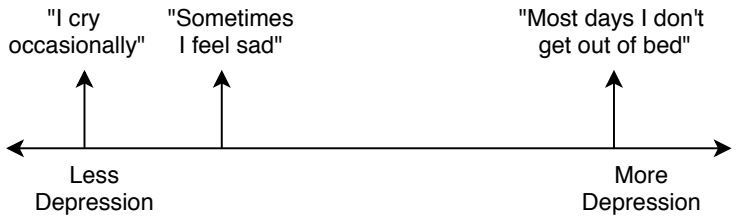

Figure 1. Relative amounts of "depression" required to endorse each item. Note the similar amounts of depression necessary to endorse the items "*I cry...*" and "*Sometimes...*", as compared to "*Most days...*"

Analysis of Ordinal Data

Ordinal data possesses rank but lacks equal intervals (Stevens, 1946). While the categories of ordinal data can be ordered, the intervals between the categories do not possess equal intervals. Ordinal data is commonly used to measure an underlying latent trait. Latent traits can be defined as dimensions or qualities that are not directly observable but are believed to influence observable actions or behaviors. For example, math ability can be considered as a latent trait that can measured through a math aptitude test. The math aptitude test produces ordinal data using the math questions on the test. A student's answers to the test are ordinal data that are believed to be associated with the latent trait of math ability.

Likert-style questions, educational assessments, and many dichotomous "yes-no" items typically produce ordinal data to measure latent traits. While ordinal data is often represented with numbers, ordinal data does not possess the same properties as natural numbers. Of most importance, ordinal data does not possess equal intervals. Suppose the scenario as depicted by Figure 1, is as follows. A researcher uses dichotomous item scale with three questions that measures the latent trait of depression: "I cry occasionally," "Sometimes I feel sad," and "Most days I don't get out of bed." The minimum score on this assessment is 0, and the maximum score is 3. The difference in actual "depression that may be needed to move from a score of 1 to 2 may

be much smaller than the amount needed to move from a score of 2 to 3 (see Figure 1). That is, it is relatively "easy" for a given participant to endorse "I cry occasionally" and "Sometimes I feel sad." Much more "depression" is needed to endorse "Most days I don't get out of bed." This discrepancy is not reflected if the total scores on the questionnaire are analyzed as if they are natural numbers with equal intervals between them. For natural numbers, there is not more space between the numbers 1 and 2 as there is 2 and 3. Thus, this scale, and many like it, possesses rank ordered scores, but not equal interval measurement of the underlying latent trait.

Despite these unique characteristics, ordinal data is often treated as interval or ratio data. Whether it is due to lack of knowledge about ordinal data or its disregard, this can cause issues in the results of many common statistical tests. Since the magnitude between levels of ordinal data can be unequal, linearity is not guaranteed. As a result, while ordinal data can be used to measure a latent trait, the population mean and variances produced by ordinal data are not guaranteed to be equivalent to mean and variances of the underlying latent trait (Davison & Sharma, 1988). Further, in most circumstances ordinal data cannot be truly normally distributed because normal distributions require continuous, interval scales (Harwell & Gatti, 2001; Lord & Novick, 1968). Ordinal data may approximate a normal distribution, but only if the scale possesses a sufficient number of categories.

Comparing means derived from ordinal data, such as in *t*-tests and in the main effects tests in ANOVA, is generally not an issue when standard assumptions hold or are approximated (Davison & Sharma, 1988, 1990). However, if a test is too hard or too easy for the respondents, the true difference in means between two groups or two timepoints can be masked (Embretson, 1994). The detection of interactions and multiplicative relationships (e.g., quadratic trends) can be impacted when ordinal data is used, as well. In a factorial ANOVA, ordinal data can falsely produce significant interaction effects when none are actually present (Davison & Sharma, 1990; Embretson, 1996). Multiple regression suffers from the same issue: non-linear variables can produce spurious interactions, and quadratic trends can appear as interactions

between variables (Busemeyer & Jones, 1983; Kang & Waller, 2005; Morse, Johanson, & Griffeth, 2012). Similarly for latent growth curve models, using ordinal data without modeling the underlying factors can produce biased estimates of slope (Yang, Olsen, Coyne, & Yu, 2017). The use of ordinal data as compared to continuous interval data is also associated with a loss of statistical power (Russell, Pinto, & Bobko, 1991).

It should be noted that ordinal data is not a prescription for statistical failure, nor is the division between ordinal and interval data always obvious (Harwell & Gatti, 2001; Velleman & Wilkinson, 1993). While a variable such as "years of education," may seem to possess interval scaling due to the even passage of time, four years in high school, four years of college, or four years of graduate school do not necessarily represent the equal increases in education or knowledge. Depending on the statistical question being asked, the same variable may be appropriately considered interval or ordinal.

Item Response Models

Scaling Ordinal Data to Interval

One method of overcoming unequal intervals is by transforming the ordinal data into interval data on a log odds unit (logit) scale through item response models (IRMs). When IRMs fit the data properly, they produce person and item parameters on the same interval-level scale (G. H. Fischer, 1995; Perline, Wright, & Wainer, 1979). Because many of the issues with ordinal data occur due to the lack of an interval scale, IRM derived trait estimates can mitigate some of the common issues described above.

In general, to allow comparison between measurements two things must be established: the standard that a score is compared to and the numerical basis of that comparison (Embretson & Reise, 2000). In Classical Test Theory (CTT), the standard is often the norm group, and the basis of comparison is order. That is, what is considered a high or low score is based on responses from a standardizing sample of participants. Even if the scores are normalized, interval data is not truly achieved; scores are only relatively higher or lower than each other, but the intervals between the data do not have meaning. Two participants may have math ability scores of 5 and 10,

but it is unknown how much more capable the higher scorer is than the lower scorer.

In Item Response Theory, persons and items are placed on a common scale to give meaning to the trait level. The standard is the latent trait scale and the numerical basis of comparison is interval (Embretson & Reise, 2000). As both persons and items receive placement on the latent trait scale, there can be meaning and predictive power in knowing their respective abilities and difficulties. For example, in the Rasch model when a person and an item have equal location of the item's threshold and the person's trait score, there is a 50% chance that the item will be endorsed. If the participant ability and item difficulties are known, the odds of a participant endorsing each item can be predicted.

IRMs possess two broad categories of parameters: those for persons and those for items. *Person abilities*, or trait levels, refer to where a respondent is on the latent trait scale. Higher values indicate higher levels of latent trait and are associated with increased likelihood of passing or endorsing an item. *Item difficulty* similarly refers to the location of the item on the latent trait scale. In Rasch models, dichotomous items are represented using logistic curves. The item difficulty is equal to the inflection point of the curve. This is where the probability of passing the item is 50% for a person with a trait level equal to the item difficulty. The term *threshold* refers to locations where the likelihood of picking one option over another changes. In the dichotomous Rasch model, the thresholds are at the inflection point. Before the threshold, a respondent is most likely to fail the item. After the threshold, a respondent is most likely to pass the item. Polytomous IRM models have multiple thresholds within each item. These thresholds represent where the highest likelihood of responding moves from one category to the next.

The approximation of a interval scale in IRMs is based upon the theory of conjoint measurement (Luce & Tukey, 1964). In the theory of conjoint measurement, interval-scale measurement can be obtained if the outcome variable is an additive function of two other variables. All three variables must also possess order by magnitude. The additive function between the variables can be inherently present, or

created through the use of a monotonic function. In the Rasch model, the additive combination of trait level and item difficulty can be established as

$$log(\text{Item Odds}) = log(\text{Trait Level}) - log(\text{Item Difficulty})$$

where Item Odds is likelihood of endorsing or passing an item (Embretson & Reise, 2000; Rasch, 1960). Performance on a questionnaire can be measured through the log odds of passing the item based on the additive relationship between the logs of the trait level and item difficulty.

According to Luce and Tukey (1964), three conditions must be met to support additivity. The first two, solvability and the Archmidean condition, establish that the unit of comparison is bounded by the laws of the natural number system (Embretson & Reise, 2000). The third, double cancellation, is the condition that is most relevant to establishing additivity in IRMs. Double cancellation refers to the consistent ordering of the three additive variables. For IRMs, these are the likelihood of passing an item (item performance), respondent trait level, and item difficulty. First, single cancellation establishes ordering within the variables. It must be true that the probability of passing a specific item always increases as participant ability increases. It must also be true that the probability of passing for a specific person always decreases as item difficulty increases. Double cancellation is the combination of these two facts. It must be true that as trait level increases and item difficulty decreases, the probability of passing an item always increase. There cannot be items where a decrease in difficulty does not correspond with an increased rate of passing. Models that allow for this, like the Three Parameter model, do not achieve conjoint measurement. As item difficulty, trait level, and performance can all be ordered via magnitude and are additive through their logs, the Rasch model fulfills the requirements of conjoint measurement (G. H. Fischer, 1995; Perline et al., 1979; Rasch, 1960).

The use of IRM scaled scores instead of raw scores has been shown to mitigate the issues of unequal interval measurement. In factorial ANOVA and multiple regression, the false detection of interaction effects is reduced when IRM scaled scores are used instead of raw sum scores (Embretson, 1996; Kang & Waller, 2005; Morse et al., 2012).

Morse et al. (2012) found that the likelihood of a Type 1 error was 8 times greater when using sum scores in multiple regression as compared to IRM derived scores. Spurious interactions in ANOVA and multiple regression can be caused by tests that are poorly matched in difficulty to the target sample or by tests that measure only a limited range of trait levels accurately (Embretson, 1996; Morse et al., 2012). The difficulty and reliability of a test can be evaluated using IRMs, which allows interactions under these conditions to be scrutinized.

Rescaling ordinal data into interval data is not without risks. Not all ordinal data can be scaled into interval data. If the ordinal data is not representative of an underlying interval latent trait then no amount of rescaling will produce accurate interval data (Michell, 2009; Salzberger, 2010). Just because a scale attempts to measure a latent trait, doesn't mean that the latent trait exists or the items are related to it. In a best case scenario, the IRM model would show poor fit; however, IRM fit statistics do not always perform as expected (Karabatsos, 2000). Further, it has not been firmly established that IRM models outside of the Rasch family produce true interval data (Harwell & Gatti, 2001; Salzberger, 2010). Finally, sampling error present in the original data is subsumed into the rescaled data (Harwell & Gatti, 2001). When the assumptions of the IRM are not met, scaled scores may be no more reliable than sum scores.

The Rasch Model

The Rasch model is a psychometric model where the probability of endorsing an item is determined by the difference between the person's ability score and the item's difficulty. This relationship between the differences of the person and item parameters is modeled using a logistic function. Both persons and items are place on the same continuous latent θ-scale. The basic form of the Rasch model for dichotomous items, as adapted from Rasch (1960), is:

$$P(X_{ij} = 1|\theta_i, \beta_j) = \frac{exp(\theta_i - \beta_j)}{1 + exp(\theta_i - \beta_j)}$$

where θ represents the ability (or trait level) of person i and β represents the difficulty (or location) of item j. In the Rasch model, when a person and an item have the same ability and difficulty, the person has a 50% chance of passing or endorsing the item.

When there is no prior knowledge of item parameters, the mean of either the item or person parameters is commonly set to 0 for identification and interpretation. Should the item parameter mean be set to zero, the high and low values of person ability are given in respect to the average location of the items, and vice versa. In the Rasch model, the only parameter estimated for each item is the difficulty. Item discrimination is fixed at 1.0. As a result, all items are assumed to be equally discriminating with the same strength of association with the underlying factor. The lower and upper asymptotes of the item response functions are fixed to 0 and 1. In practice, this assumes that as a respondent's latent trait approaches negative infinity, the probability of endorsing any item approaches 0. Similarly, as a respondent's latent trait approaches positive infinity, the probability of endorsing any item approaches 1. Because of the assumptions and parameter restrictions of the Rasch model, the unweighted raw scores are sufficient statistics for the person and item parameters (Lord & Novick, 1968). In the unidimensional Rasch model, each raw total score is associated with one latent ability estimate, regardless of which items are endorsed.

Other IRMs allow for the estimation of different parameters, such as item discrimination and asymptotes. The Two Parameter Logistic Model (2PL) is a common relaxation of the Rasch model (Birnbaum, 1968). In the 2PL, item discrimination is allowed to vary between items so they may associate with the latent trait more or less strongly. Other models, such as the Three and the Four Parameter Logistic Models, free the lower and/or upper asymptotes of the response functions to vary between items (Barton & Lord, 1981). The lower asymptote is generally considered to be the "pseudo-guessing" parameter. It can be used, for example, to model responses to multiple choice tests where a respondent has a fixed chance to correctly pick the right answer regardless of their trait level. Similarly, the upper asymptote can be relaxed in situations where high trait levels are not associated with probability asymptotes of 1. In

all of these models, the total score is not a sufficient statistic. That is, the same raw total score can be associated with multiple latent ability estimations depending on which items were successfully endorsed and which were not. All the above models, including the Rasch model, can be expanded to accept the use of polytomous data.

Additional assumptions of the Rasch model include unidimensionality and local independence. To achieve local independence, responses to separate items must be correlated only through the underlying latent trait that is specified in the model. High local dependence can bias model parameters and inflate test reliability (W. Chen & Thissen, 1997; W. Wang & Wilson, 2005). Local dependence can be caused by a variety of issues related to test content and respondent characteristics. The standard example occurs in reading comprehension tests: if one set of items all ask about the same reading passage, these items will be more related to each other than the other items. Local dependence can also be caused by repeated measurement of the same person, tests where speed is important, by items that are too closely related beyond the latent trait, or by multidimensionality not specified by the model.

Anchoring

Anchoring, or equating, is a common practice in item response models for both cross-sectional and longitudinal research (e.g., De Ayala, 2009; Embretson & Reise, 2000; Von Davier, Xu, & Carstensen, 2011). Anchoring is an important procedure in setting the scale of measurement of IRMs. In the same way that a ruler is not useful if it can stretch or compress, the scale of Rasch models must be stable to enable comparison of person abilities or item difficulties. The *scale* of a Rasch model can be defined as the meaning of its units, specifically the interval between units and the placement of the center point (i.e., the location of zero). Because the means of either the persons or items must be fixed (typically, to zero) prior to the estimation of a Rasch model, there is no guarantee that any two models attempting to measure the same latent trait are estimated onto the same scale. While the underlying relationship between persons and items may be the same for two different models, the location of

zero and the standard deviation of the latent trait may not match. For example, if one Rasch model centers the scale on persons (i.e., the mean of the person abilities is fixed to be zero) and another Rasch model centers the scale on items, then the relationship between items and persons would be identical between the models but the estimated latent traits and item difficulties would be different. This also implies that the same score from two unconnected models is not necessarily reflective of the same magnitude of latent trait. Thus, the lack of a predetermined scale prevents parameters generated by two models from being reliably compared to each other.

Equating is the general practice of connecting two or more separate IRMs to share the same latent trait scale. There are many methods of equating. Some methods involve the mathematical adjustment of the model parameters, while others involve administering similar items on different test forms (Cook & Eignor, 1991). Not all item parameters need to be known to place two tests together on the same trait scale. Large-scale tests frequently use sets of similar items that are common across multiple test forms. If two different tests have enough overlapping items, then two tests can be equated (Angoff, 1984). This means that the abilities of participants can be compared even if the participants did not respond to all of the same items. Because of this, equating is useful in a variety of measurement contexts. *Anchoring* is a method of equating that is typically accomplished by using the same item parameters in each model. Anchoring is the subject of interest in the present study.

Uses of the Rasch Model

The Rasch model was originally designed to measure intelligence and achievement among school children (Rasch, 1960). Since then, the Rasch family models are commonly used to create and evaluate academic and achievement tests. For example, the Programme for International Student Assessment (PISA), a worldwide study of academic achievement, was built using conditional Rasch models (OECD, 2006) After the original calibration, data from subsequent years are placed on the same scale using tradition IRM equating techniques (OECD, 2006). The PISA has been administered

nearly 2 million times in over 90 countries between 2000 and 2012, and is still on-going today. Other tests that use Rasch measurement include the Wide Range Achievement Test 4, the Kaufman Assessment Battery for Children, and the Woodcock-Johnson Psycho-Educational Battery (Kaufman, 2004; Wilkinson & Robertson, 2006; Woodcock, Johnson, & Mather, 1990). In addition, the Lexile Measures is an example of a large scale reading comprehension test based on the Rasch model (Stenner, 1996).

The Rasch model has been extensively used to create and validate scales in the fields of psychology, education, and health. The "person measure" of the Rasch model can easily stand in for latent dimensions such as psychological traits and disease severity. In psychology, the Rasch model has been used to measure thoughts and behaviors related to mental health (Chiang, Green, & Cox, 2009; Kendel et al., 2010), personal well-being (Misajon, Pallant, & Bliuc, 2016; Tomyn, Stokes, Cummins, & Dias, 2019), and working memory, among many other latent traits (Bowles & Salthouse, 2003; Vock & Holling, 2008). In education, the Rasch model has been used to measure a variety of in-classroom phenomenon such as teacher effectiveness (Jones & Bergin, 2019), the learning of quantum mechanics (Testa et al., 2018), and student maladaptive behaviors (Garcia, Lambert, Epstein, & Cullinan, 2018). In health care, the Rasch model is commonly used to measure patient quality of life and symptom severity in a variety of fields (Janssen, Phillipson, O'Connor, & Johns, 2017; Pasternak et al., 2016; Pillas, Selai, & Schrag, 2017; Prieto et al., 2003). The Rasch model is also a popular tool to measure symptom change over time (Glas et al., 2009; Norquist, Fitzpatrick, Dawson, & Jenkinson, 2004).

Fundamentally, the Rasch model calculates the likelihood of a set of responses that are associated with an entity. Thus, the person and item measures do not have to be from traditional "persons" or "items". This flexibility allows the Rasch model to be used in a variety of fields outside of psychology and education. For example, the Rasch model has been successfully used to model gene expression, business centralization, and the epidemiological distribution of diseases (Chao, Tsay, Lin, Shau, & Chao, 2001; H. Li & Hong, 2001; Martin, McKelvie, & Lumpkin, 2016).

Software and Estimation

Item response models, including the Rasch model, can be estimated with a wide range of software. Popular IRM specific software includes WINSTEPS, PARSCALE, IRTPRO, BILOG, and ConQuest. (Adams, Wu, & Wilson, 2015; Cai, Thissen, & du Toit, 2019; du Toit, 2003; Linacre, 2019). IRMs can also be run in general purpose statistical software such as R, Mplus, and STATA (Muthén & Muthén, 1998-2017; R Core Team, 2013; StataCorp, 2017). Platform choice is largely driven by the models and estimation methods that are available, as well as the number of participants and items the software can handle. For example, the mirt package in R allows for flexible estimation of a variety of item response models, including exploratory and confirmatory models, multidimensional models, and latent class models (Chalmers et al., 2012).

A variety of estimation methods are available for IRMs. Common IRM estimation methods include full information maximum likelihood, joint maximum likelihood, marginal maximum likelihood, and conditional maximum likelihood (see M. S. Johnson et al., 2007, for a discussion of common estimation methods). Other methods, such as the Metropolis-Hastings Robbins-Monro (MHRM) algorithm and the Expectation-Maximization algorithm are also available (Bock & Aitkin, 1981; Cai, 2010a). The choice of an estimation method is largely driven by model complexity; models with many latent factors may be better served by MHRM than by marginal maximum likelihood, for example (Cai, 2010a).

Change Over Time and IRMs

It is well known that basic change scores can be unreliable in estimating change (Bereiter, 1963; Cronbach & Furby, 1970). Specific statistical models for longitudinal data exist because longitudinal data possesses peculiar qualities that are distinct from cross-sectional or single timepoint data. These qualities include changes over time, auto-correlations between measures, and attrition. All these factors and more can bias the results of statistical models when they are uncontrolled. Therefore, there is a great need for item response models that are appropriate for longitudinal data. However,

despite the benefits of IRMs in analyzing ordinal data, the use of longitudinal IRMs is not common.

The dearth of longitudinal item response analyses in the literature is not solely due to the non-existence of longitudinal IRMs. Longitudinal IRMs were of interest to researchers soon after the conception of the original Rasch latent trait model (Andersen, 1985). Longitudinal IRMs include the multidimensional latent trait model for measuring learning and change, bifactor models, and autoregressive models, among others (Embretson, 1991; Gibbons & Hedeker, 1992; Jeon & Rabe-Hesketh, 2016; Liu & Hedeker, 2006). Second-order longitudinal models are presently on the rise, an area that shows much promise in expanding the flexibility and use of longitudinal IRMs (Cai, 2010b; Paek, Park, Cai, & Chi, 2014; C. Wang, Kohli, & Henn, 2016).

Longitudinal IRMs have several significant drawbacks, though. First, they are generally computationally demanding, which creates barriers to their applications in large, multiwave datasets. Second, IRMs in general require a significant amount of data for proper estimation. The amount of data necessary increases as model complexity increases; as a result, multidimensional longitudinal models require very large amounts of data to fit. Finally, the maximum number of timepoints that can be examined is heavily censured by the estimation method. The multiple integral in the marginal maximum log-likelihood functions used in many IRMs caps the number of timepoints to 15, 10, or even 5 (Bacci, 2012; Wood et al., 2002). Rasch models can be structured as multilevel models, which reduces the impact of timepoints on convergence. But, the failure to converge rate for these models was found to range from 30% to 60% regardless of the simulation condition (Bacci, 2012). That is, there may be a significant non-zero baseline rate of failure that is separate from the model's context. All together these drawbacks present a sort of catch-22: the researcher must have enough data to estimate a multidimensional model but not so much that estimation becomes intractable.

Low-computation longitudinal IRM alternatives are available but they may not produce the information that the researcher desires. For example, while the linear logistic model with relaxed assumptions (LLRA) can be used to detect change over

time, neither the person or item parameters can be directly examined (G. H. Fischer, 1989; Rusch & Hatzinger, 2009). In the LLRA, person parameters are conditioned out (i.e., not estimated) and time is expressed as changes in the item parameters (G. H. Fischer, 1983). Thus, only overall or group differences can be examined.

Finally, many longitudinal IRMs are only appropriate in specific circumstances. Some models require that the person or the item parameters are already known (Andrade & Tavares, 2005; Tavares & Andrade, 2001). While obviously not ideal, the barrier to utilizing longitudinal IRMs may be, or simply seem, unsurpassable to applied users. It is not unusual to see researchers simply ignoring the variable of time and using traditional, single timepoint IRMs for analyzing longitudinal data (Chang & Chan, 1995; Wright, 1996, 2003).

Similar issues arise when analyzing longitudinal ordinal data outside of IRMs. For example, structural equation modeling (SEM) is the tool of choice for many longitudinal analyses, but lengthy ordinal scales can cause significant problems. If every item of a scale is included separately in the model, the number of estimated parameters increases. Unlike interval data, additional parameters for dichotomous ordinal items should be included to estimate the item thresholds. These parameters are in addition to the factor loadings of the items, additional variances, and parameters to estimate violations to unidimensionality, any of which may not be stable across different timepoints. Thus, lengthy ordinal scales can greatly increase model complexity, the necessary sample size, and estimation time required for a model, which can all lead to intractable problems in estimation. Parceling (using one or more sums or averages of two or more items) is a common method around this, but in many circumstances can cause model mispecification, inflated model fit, and biased parameters (for reviews, see Bandalos & Finney, 2001; Little, Rhemtulla, Gibson, & Schoemann, 2013; Yang, Nay, & Hoyle, 2010). There is also limited evidence that parceling can falsely produce significant estimates of linear growth when there is none (Yang et al., 2010).

The use of estimated factor scores from IRMs has been proposed as one method around simultaneously estimating IRMs and growth parameters (McArdle, Grimm,

Hamagami, Bowles, & Meredith, 2009; Yang et al., 2010). In this process, IRMs are used to estimate factor scores and the factor scores are then used as input variables in longitudinal models. This can allow the researcher to evaluate the function of a scale over time and produce interval level data without increasing their target model's complexity. One method of producing longitudinal IRM factor scores is longitudinal anchoring.

Virtual Persons and Items. The use of virtual persons and items is common in IRMs and in longitudinal anchoring, specifically. The term *virtual* refers to items or persons that exist only in a formal or mathematical sense, as in contrast to real items on a test or the real participants who responded to it. Virtual items are established not by their physical existence, but by the different conditions under which they occur. For example, a single item responded to at two different timepoints may be considered as two separate, unrelated "virtual" items (Embretson, 1991; G. H. Fischer, 1989; Verguts & De Boeck, 2000). Virtual items are used in many circumstances outside of longitudinal analysis. Real items may be separated into virtual items to model how a previous correct or incorrect response influenced future responses (Hoskens & De Boeck, 2001). They can also be used to measure local item dependence or test fatigue (Ackerman & Spray, 1986; Kubinger, 2008). As there are always more virtual items than real items, the number of item parameters to estimate will increase when virtual items are used. This can cause issues in estimation in certain situations.

Virtual persons are similar to virtual items. Virtual persons occur when the responses of a real person are split as if they came from two or more virtual persons. Generally, these splits are based on the different conditions that a person took the test under. In longitudinal analysis, this is often the different occasions of measurement. For example, the responses of one person at three timepoints may be considered as the responses of three independent, virtual persons. This approach is used in the linear logistic model to assess change over time (G. H. Fischer, 1989; Rusch & Hatzinger, 2009).

One concern with virtual persons and items is that the local dependencies among

the real persons' responses are not modeled. Most unidimensional IRMs are built on the assumption that the specific latent dimension of interest is the sole influence on responses to items (Lord & Novick, 1968). Once the influence of trait level is removed, the responses of any two individuals or the relationship between any two items should be relatively uncorrelated. This may be violated when virtual persons or items are used. Sets of responses from one person are likely to be additionally correlated to each other because they came from the same person. A specific person may have patterns of responding that are not associated with the latent trait. Similarly, individual questions may elicit specific patterns of response, creating relationships between repetitions of the same item. If these local dependencies are unaccounted for they can bias trait estimation (Marais, 2009; Olsbjerg & Christensen, 2015). Moreover, local dependencies within persons can either inflate or mask the true change over time (Marais, 2009; Olsbjerg & Christensen, 2015).

It should be noted that, in general, it is not feasible to include both virtual persons *and* virtual items in the same standard Rasch model. If virtual persons and items are defined by the timepoint that each occurred in, each virtual person would only have response data for each set of virtual items (see Figure 2). That is, virtual person A_1 would only have responses for all virtual items at time 1, and virtual person A_2, would only have responses for for all virtual items at time 2. Because no data overlaps between items or people, each timepoint would be estimated separately from each other. In essence, this method produces the same outcome running each timepoint in a separate model.

Longitudinal Anchoring

Linking separate IRMs together through a specific method of longitudinal anchoring was proposed by Erbacher et al. (2012) as a low-computation alternative to longitudinal Rasch models. Standard anchoring occurs when separate models are placed on the same scale by possessing the same item parameter estimations. In longitudinal anchoring, separate models are run for each timepoint with the models linked through a

Items		Time 1			Time 2			Time 3		
		1_1	2_1	3_1	1_2	2_2	3_2	1_3	2_3	3_3
	A_1	1	1	1						
	B_1	0	0	1						
	C_1	0	1	0						
	A_2				1	1	0			
Persons	B_2				1	0	0			
	C_2				0	0	0			
	A_3							1	1	1
	B_3							1	0	0
	C_3							0	0	1

Figure 2. An example data matrix where the same three persons (A, B, C) responded to the same three items at three timepoints. The use of both virtual persons and items creates no overlap between timepoints.

common set of item parameters. Longitudinal anchoring is a two step procedure: first, the item parameters are created, and then the single timepoint models are run. There are several methods of estimating the shared item parameters that are unique to longitudinal anchoring; these are described in detail in a later section. As longitudinal anchoring does not involve simultaneous estimation of all timepoints, the computational burden is low. If a single timepoint IRM can be run, then longitudinal anchoring can be used.

To accurately compare multiple IRMs, item parameters must be anchored to the same logit scale. This increases in importance when comparing the same scale over time, as one important property of IRMs is measurement invariance. To achieve measurement invariance, a scale must measure the same attribute(s) equivalently across different situations and conditions (Horn & McArdle, 1992). Items within a scale must measure the underlying trait exactly the same every time they are administered regardless of

who is responding to the item (Embretson & Reise, 2000). As a result, item difficulties must not change over time to be considered invariant. If the true difficulty of an item changes across situations, then that item is not measuring the same attribute equivalently in all situations. This requirement is also logical; an unaltered item should not, itself, evolve. While an individual's latent trait may increase or decrease, the true difficulty of each item should remain constant. This may be increasingly relevant in longitudinal contexts if item are poorly matched to participants during some years but not others or from the beginning to the end of the study. Poorly estimated item parameters can produce biased or inaccurate trait estimates (Svetina et al., 2013).

Nevertheless, item drift can occur. Item drift is broadly defined as systematic changes in item parameters over time. Item drift has a variety of causes including developmental changes in the participants, repeated exposure to items, and, for academic tests, changes in curriculum (Goldstein, 1983). The size of the item drift corresponds to the size of the resulting bias in person ability estimates (Wells, Subkoviak, & Serlin, 2002). Although it is useful to examine how item difficulties change over time, it is nevertheless difficult to compare change in latent trait scores when the test is unstable. In many circumstances, there is no theoretical reason for why a given item should elicit different responses over time. Changes in a participant's true latent trait should be the only cause of change in the estimated latent trait, otherwise change would be improperly measured. Most methods of longitudinal anchoring assume that item parameters do not change over time as the item parameters are fixed to be equal across all timepoints. Models that allow for items to change over time may be useful in evaluating item function, but they are likely less useful for producing accurate latent trait estimates.

Longitudinal anchoring has added benefits for scale evaluation. The results from Erbacher et al. (2012) suggest that longitudinal anchoring can be used to identify misfitting persons and items. Many scales used in psychological research have not been psychometrically evaluated over time due in part to the lack of longitudinal item response models. Even if a researcher chooses not to use latent trait estimates from

IRMs in their analyses, it would still be useful to ensure that all items are functioning similarly throughout the duration of their study and to identify grossly misfitting respondents. Validating this low-computational alternative may encourage psychometric evaluation of longitudinal scales.

Longitudinal anchoring has not yet been thoroughly evaluated as a method of linking more than two separate IRMs together. For larger numbers of models, it was first proposed by Erbacher et al. (2012) to evaluate the performance of the PANAS scale over 56 consecutive days. When there were two timepoints, another study found that longitudinal anchoring via fixed parameters performed equivalently or better than other anchoring procedures such as mean/mean linking and concurrent calibration (L. Fischer, Gnambs, Rohm, & Carstensen, 2019). von Davier, Carstensen, and von Davier (2006) suggested that longitudinal anchoring was not appropriate when populations differed in ability, though this claim was not tested.

Because most longitudinal IRMs can easily handle two timepoints, it is critical to evaluate the performance of longitudinal anchoring when there are many more timepoints. There is a dearth of available IRM methods that allow for such kinds of analyses. In addition, when there are more than two timepoints there are more methods of creating the fixed item parameters. These different methods have not been staunchly evaluated by simulation or otherwise.

In the present study, I examined several methods of longitudinal anchoring under different change conditions. Four of these methods were drawn from Erbacher et al. (2012). Two additional methods, the Random sampling method and the Cross-sectional method, were also examined.

Anchoring Methods

Six different anchoring methods were examined: floated analyses (Floated method), all timepoint anchored models (All Times method), time one anchored models (Time One method), mean anchored models (Mean method), random sample anchored models (Random method), and cross-sectional anchored models (Cross-sectional

method). More information on the original use of the first four methods can be found in Erbacher et al. (2012), though the titles of the methods have been adjusted for generalizability.

Floated analyses. For the Floated method, each timepoint of data is used to fit its own separate model. For each model, the mean of the items is fixed to 0. This is because the Rasch model requires fixing the mean parameter of either the person or the item distribution. Centering the scale on the persons would prevent any comparison of latent trait scores between each timepoint as change would then be primarily expressed in the item difficulties.

Because the trait scores are not locked onto the same scale, they are not guaranteed to be comparable across years. Furthermore, this method allows item difficulties to change over time, which violates measurement invariance. This method is mainly used to create a baseline of comparison for the other anchoring methods both in this study and in Erbacher et al. (2012). This method, though, is sometimes used to compare data from two timepoints (e.g., Kahler, Hustad, Barnett, Strong, & Borsari, 2008).

All Times Anchored Models. In the All Times method, all participants and items across all years are first run together in one, overarching "parent" model. Repeated administrations of items are considered as separate virtual items. This method has also been referred to as "racking" (see Figure **??** for an illustration; Wright, 1996, 2003). The parent "racked" model is fit, and the resulting item difficulties are extracted. These difficulties are then used to run separate models, one for each timepoint. For example, the item difficulties from the all the timepoints are estimated in the overarching parent model. Then, the resulting first timepoint item estimates are used to anchor a separate model that only contains the first timepoint responses. This first timepoint model would then produce latent trait estimates for all participants with data for that timepoint. This process is repeated for each and every timepoint. Thus the item parameters are estimated using all available data, but the person parameters are estimated using only item parameters from one year.

Items		Time 1			Time 2			Time 3		
		1_1	2_1	3_1	1_2	2_2	3_2	1_3	2_3	3_3
	A	1	1	1	1	1	0	1	1	1
Persons	B	0	0	1	1	0	0	1	0	0
	C	0	1	1	0	0	0	0	0	1

Figure 3. A "racked" data matrix used in the All Times method. Here, virtual items are used. Each repetition of an item is considered to be a completely separate item in the model. Persons appear once, but items appear multiple times.

Erbacher et al. (2012) claimed that this method, referred to therein as the "56-day method", allows for direct comparison of latent trait estimates over time because the item difficulties are all on the same logit scale. This method use virtual items, where each presentation of an item at each timepoint is considered as a separate item. Item difficulties of each item at each time are free to change which violates measurement invariance. The All Times method may also violate local independence. Sets of virtual items from the same real item are likely to be more correlated with each other above and beyond their relationship to the latent trait. While the individual year models will not violate local independence, the latent trait estimates may still be biased because of the biased item estimates.

I hypothesize that the item parameters produced by this model will show extreme bias when there is a non-zero mean change in trait scores over time. Because the parent model holds person trait scores constant to produce item estimates, any true change in the person latent trait will be reflected as a change in the item parameters. For example, consider an item that is collected at four time points. A group of people with positive growth across these four time points responds to this item. If the people are in the data matrix once, but the items are represented as virtual items, then the positive increase in the participant's scores will be reflected in the items. Since people cannot receive more than one growth parameter in a standard Rasch model, their growth must

be expressed in the items, instead. Each successive presentation of the item will be considered easier than the last, resulting in four different difficulty estimations for each virtual presentation of the item. This purposely occurs in the linear logistic model with relaxed assumptions (LLRA) (G. H. Fischer, 1973, 1989). It is one drawback of the LLRA: because change is reflected in the items, person scores are uninterpretable. Change in the item parameters can also be purposely examined in "racked" analyses (Wright, 1996, 2003).

Thus, while the All Times method may perform adequately when there is no average change in participant scores, the model is likely to produce biased estimates when participant scores do change over time. This bias may be revealed by poor model fit, bias in the recovered parameters, or high standard errors of the latent trait. In essence, the separate, single timepoint models may not use the same centered logit scale. Because there was not a significant change in participant scores from the beginning to the end in Erbacher et al. (2012), the direct influence of change on anchoring has not been examined.

Mean Anchored Models. In the Mean method, an overarching, racked parent model is first run, identical to the All Times method. Then, the average of the virtual item locations is calculated for each item. Each timepoint of data is then run as its own separate model, using the mean item parameters. Because the mean of the item parameters is taken, each model uses the same item difficulties. Erbacher et al. (2012) used only small fraction of the timepoints for the mean (3 of 56), but we will use all available, simulated timepoints (10 of 10).

Similar to the All Times method, as persons are held stable in the parent model, change may occur in the items instead. Taking the mean of the virtual items may ameliorate some of this bias, assuming that change is evenly influencing all items at all times. Latent trait scores may be comparable across years, depending on the bias present in the item difficulties. I hypothesize that the Mean method will perform better than the All Times method, but that it will still show some biases in parameter estimation.

Time One Anchored Models. The Time One method is most similar to traditional anchoring procedures. First, a Rasch model is run on the first collection of data. Then, the item difficulties from that single model are used to anchor all other timepoints in their own separate models. Each model has the same item difficulties. As each model will be on the same logit scale, person parameters are comparable across years. Stable item difficulties allow for measurement invariance.

Presumably, any timepoint can be used to anchor the scale. The Time One method may be more generally referred to as the Single Time method. Using a single timepoint of data is a common anchoring practice in longitudinal studies (e.g., Boone & Scantlebury, 2006; Deacon, Kieffer, & Laroche, 2014; C. Li et al., 2017; Watson, Kelly, & Izard, 2006). The statement that any timepoint can be used as the scale anchor implies that no timepoint is more appropriate than another. In applications of this method, different timepoints may lead to more or less accurate item parameter estimations. The first timepoint may be preferable if the sample at the first timepoint is the most representative of the entire sample. In other circumstances, a timepoint near the middle may be preferable if significant change between the first and last timepoint is expected. Items that are poorly matched to participants may be poorly estimated. When trait change is expected across a study, the most prudent timepoint to use for anchoring will be the timepoint where persons and items are best matched in difficulty and where the participant sample is most representative.

Random Anchoring. In the Random method, a model is run using a random sample of the full dataset (as adapted from Mallinson (2011)). The random dataset contains a single set of responses from each participant at a randomly chosen timepoint. For example, the anchoring set could consist of data from participant 1 at time 3, participant 2 at time 2, participant 3 at time 1, and so on (see Figure 4 for an illustration). The anchoring model is fit using the randomly sampled data. The item difficulties from this model are used to anchor the separate timepoint models. To the author's knowledge, this method has primarily been used when only a few timepoints are considered (Curran et al., 2008; Mallinson, 2011; Wright, 2003).

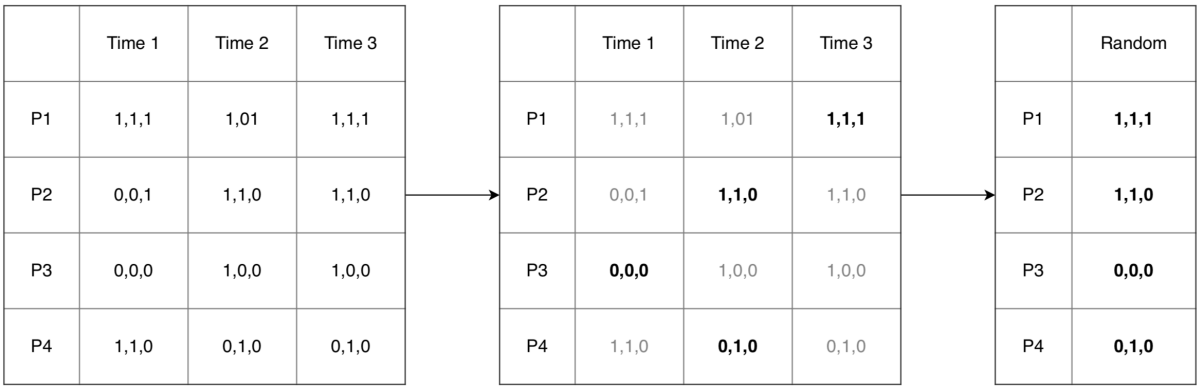

Figure 4. An example of selecting data for the random method. A random time point of data is chosen for each of the observed persons (P1 through P4). These random selections are combined to create the anchoring data.

The purpose of the Random method is to include one "slice" of data from every participant in order to avoid local dependence and equally represent each timepoint in the item estimation. The item parameters resulting from this model are influenced by data at all timepoints. The Random method upholds local independence and measurement invariance because item difficulties do not change over time between the models, and no virtual persons or items are used. Latent trait scores are comparable across years.

The Random method may perform poorly when missingness is high or nonrandom. Supplemental samples similarly may skew the random sample's representation. If the random sample includes every participant, each participant will contribute equally to the anchoring method, regardless if the participants had 1 or 100 complete waves of data. Thus, participants with less data, and timepoints with more participants, would be over-represented and contribute more to the anchoring sample. If the missing data is not at random then parameter estimation may be biased. If participant attrition is caused by confounding variables, the random sample of participants may not be representative sample of the true population. When there is little to no missing data, I hypothesize that this method will perform similarly to the Time One method.

Cross-sectional Anchoring. In the Cross-sectional method, all repeated measurements are considered as if they are from separate, virtual participants. This

Items			1	2	3
Persons	Time 1	A_1	1	1	1
		B_1	0	0	1
		C_1	0	1	1
	Time 2	A_2	1	1	0
		B_2	1	0	0
		C_2	0	0	0
	Time 3	A_3	1	1	1
		B_3	1	0	0
		C_3	0	0	1

Figure 5. A "stacked" data matrix. This form of data uses virtual persons. Each person appears multiple times and each item appears once.

method has also been referred to as "stacking" because the data is put into long format (see Figure 5 for an illustration; Wright, 1996, 2003).

This stacked data set is used to fit the parent anchoring model. Item parameters from the parent model are used to fit separate models for each timepoint. This method was not examined in Erbacher et al. (2012), but has been used in other studies to create fixed item parameters and analyze longitudinal data (e.g., Anselmi, Vidotto, Bettinardi, & Bertolotti, 2015; Jodoin, Keller, & Swaminathan, 2003; C. Li et al., 2017; Sturrock et al., 2015). In some of the previous studies, the Cross-sectional parent model itself was analyzed. This anchoring method has the advantage of including data from every participant in the anchoring estimates, but the possibility of bias exists due to the repetition of persons and violations of local independence.

General Comments on Anchoring Methods

While previous studies have used certain aspects of the anchoring methods described above, the technique of using a parent model to calibrate item difficulties and then using separate models at more than two timepoints to obtain person trait estimations is unique to this study and Erbacher et al. (2012). Many of the techniques used in the setup of the parent models (e.g., using virtual person or items, "racking" and "stacking") have been used to analyze longitudinal data with IRMs (e.g., Chang & Chan, 1995; Mallinson, 2011; Wright, 1996, 2003). The difference between the previous studies and the present method is the use of single timepoint models to estimate trait levels. The goal of the two-step procedure (parent model, then separate models) is to provide a computationally feasible method of reducing the impact of local dependence on the estimations of latent trait scores.

Anchoring method choice likely depends on the data the method is applied to. If the study has high attrition rates, the Time One anchoring method may be preferable to combat against non-missing at random attrition. Anchoring item difficulties to a wave of data where participants and items are poorly matched may result in poorly estimated parameters. It is presently unknown how the different anchoring methods compare to each other as they each use different methods of creating the item difficulties and possess different qualities (see Figure 6 for an overview of the methods).

Previously, Erbacher et al. (2012) recommended using the Time One ("day 1") and Mean ("mean-3") methods, though the All Times ("56-day") method also produced similar fit statistics. It should be noted that in Erbacher et al. (2012), it is not known if there was significant average change from baseline across the study period. Only the person and item fit was examined in Erbacher et al. (2012); change in factor scores was not examined. As the study used self-reported daily mood over 56 consecutive days, there may have not been a significant average change from day one to day fifty-six. Previous research has found that individuals typically have stable variability in mood over time (Eid & Diener, 1999; Penner, Shiffman, Paty, & Fritzsche, 1994). Thus, it is unknown if the presence and degree of average change impacts the performance of the

	Same Scale	Stable Items	Local Independence	Pools Data
Floated			✓	
Time One	✓	✓	✓	
All Times	✓		~	✓
Mean	✓	✓	~	✓
Random	✓	✓	✓	~
Cross-sectional	✓	✓		✓

Figure 6. A summary of the qualities of each anchoring method. Quality present: ✓Quality partially upheld: ~

different anchoring methods.

Longitudinal anchoring is an efficient method to produce latent trait estimates, examine changes in item and person fit across timepoints, and evaluate longitudinal scale functioning. It should be noted that while anchoring separate IRMs produces latent trait scores at each timepoint, it does not directly estimate relevant growth parameters (e.g., intercept, slope, and their covariance). Latent trait scores from anchored IRMs can be used as input in growth focused models (e.g., latent growth curves, structural equation modeling) (Gorter et al., 2015; McArdle et al., 2009).

Because person trait scores are estimates with estimation error, more precisely measured trait scores lead to less bias in secondary analyses (Glas et al., 2009; Mislevy, Johnson, & Muraki, 1992). Thus, the person scores produced by longitudinal anchoring are only as useful as they are accurate. Traditionally, this compound error is reduced by simultaneously modeling the latent factors and the growth factors by using methods such as longitudinal IRMs or structural equation modeling. While this may be ideal, directly modeling factors for multiple scales at multiple time points can quickly increase computation time and require too many parameters, as noted in the section *Change Over Time and IRMs*. Longitudinal anchoring may be an acceptable solution, especially when alternatives, such as treating ordinal data as continuous, can bias growth curve

estimates (Yang et al., 2017).

Simulation

Conditions

The performances of the six anchoring methods were evaluated using simulation. There were 100 simulations of each anchoring method under four different change conditions. In each condition, 500 simulated participants responded to a set of 50 items collected over ten timepoints.

Two parameters were manipulated: the population slope of the latent trait and the variance of the slope. Change over time was evaluating using a standard latent growth model. In latent growth models (LGM), change in a trait is calculated by estimating latent variables for the intercept and slope (**??**). An LGM fits a regression line per person, which for linear growth is,

$$y_i = b_1 + t_i b_2 + e$$

where y_i is the value of the Rasch-derived trait at time i, b_1 is the intercept, b_2 is the slope, t is the time point, and e is the time-specific prediction error. The distribution of the intercepts (b_1) and slopes (b_2) across participants are modeled as two separate latent variables. The mean of the latent slope is an estimate of the average change over time in the sample. The variance of the latent slope represents the interindividual differences in the change over time.

Each participant's rate of change was drawn from a distribution with a mean and variance determined by the condition. The four conditions are as follows: mean slope of 0 with zero slope variance, mean slope of 0 with variance, mean slope of 0.10 with variance, and mean slope of 0.20 with variance. For all conditions with slope variance, the variance of the slope was equal to 0.0025, producing a standard deviation of 0.05 around the mean slope parameter.

Thus, for the 0.10 slope condition the simulated mean population latent trait increased by approximately 1 from time one to ten. For the 0.20 slope condition, the

population mean increased by 2 from time one to time ten. Under the zero slope with variance condition, the mean change of the population trait level over the 10 timepoints is approximately zero. Each simulated participant's rate of change was drawn from a distribution with mean zero and variance equal to 0.0025. Simulated participates therefore had small stable positive or negative rates of change, but these individual rates of change canceled out to produce negligible population mean change. There was no change in participants latent trait in the Zero Slope, Zero Variance condition.

Process. The simulation was conducted using the *mirt* package (version 1.30) in R 3.5.3 using full information maximum likelihood estimation (Chalmers et al., 2012; R Core Team, 2013). In all models where the item locations were estimated, the sum of the item locations was set to zero. The mean of the person abilities was always free to vary. The general procedure of the simulation and the values of the fixed parameters are detailed below.

Simulate persons and items. For each simulation under each condition, initial abilities at time 1 for 500 persons were drawn from a normal distribution with mean of zero and standard deviation of one. Rate of change (slope) for each person was drawn from a normal distribution with mean and variance according to the respective condition. Person abilities for rest of the 9 timepoints were generated according to each person's initial ability and individual slope.

The difficulties of fifty dichotomous items were drawn from a uniform distribution from -2 to 2 for each simulation[1]. Item discrimination values were equal to one. Responses were then generated for each participant at each timepoint using the Rasch model. Even if a single participant had the same simulated trait level at all timepoints, responses at each timepoint are not guaranteed to be identical under the Rasch model. This process was repeated for 100 simulations for each change condition, for a total of 400 sets of persons and items. For each simulation under a given condition, the six anchoring methods used that same set of responses.

[1]Pilot testing showed that normal distributions would occasionally produce item difficulties that were be too extreme to be properly estimated. A uniform distribution prevented model non-convergence and therefore spurious simulation failure.

Fit the parent model. The simulated responses were used to fit the parent model according to the anchoring method. In the Floated method, there was no parent model. All parent models and the floated models were fit with the "*itemtype = 'Rasch'*" argument to fix item discrimination parameters to 1 and freely estimate the population variance. The mean of the item difficulties was set to 0 and the mean of the person abilities was freely estimated.

Fit the individual models. Individual models for each of the ten timepoints were fit using the fixed item parameters generated by the parent models. Person parameters were estimated from the individual models.

Evaluation

As the Rasch model assumes a structure on the functioning of items and the pattern of responses, anchoring methods were evaluated based on the models' fit parameters, the accuracy of the estimated item and person parameters, and the standard errors (SE) of the estimate. The evaluated outcomes include the standard error of the estimate, the RMSE of the item parameters, the recovery of slope and the variance of the slope, the M_2 statistic, RMSEA, SRMR, and AIC. Standard residual infit and outfit measures were not evaluated due to the instability in the interpretation of misfit cutoff scores (Karabatsos, 2000).

Accuracy of the Person Parameters. Latent trait scores were evaluated through the standard error of the estimate and, by proxy, slope recovery. The standard error of the estimate (SEE), also known as the standard error of measurement, is a measure of the precision of the estimated person latent trait (Embretson & Reise, 2000). It can be defined as,

$$SE(\theta) = \frac{1}{\sqrt{TI(\theta))}}$$

where TI is the test's item information function. More accurate anchoring methods will produce lower standard errors for the latent trait estimations.

Root Mean Square Error (RMSE) was not used to evaluate bias in person ability estimates. Because the scales of the models were set on the items, high values of RMSE

for person abilities could indicate poor person estimation *or* accurate person estimation with a shifted scale. As both of these cases can be determined by combining item RMSE and slope recovery, person RMSE was not evaluated.

Accuracy of the Item Parameters. The accuracy of the item difficulties was evaluated using root mean square error (RMSE). RMSE is a measure of the distance between the true (simulated) and observed (estimated) item locations and is calculated by,

$$RMSE = \sqrt{\frac{\sum(S_{\mathrm{i}} - O_{\mathrm{i}})^2}{n}}$$

where S represents the true simulated item location, O represents the estimated item location, and n represents the number of items. Lower values of RMSE indicate that the model estimated the item locations closer to their true values. RMSE is a common method of model evaluation, but there is no universally defined threshold to determine model fit. Rather, RMSE is helpful in comparing different models to each other.

It is important to note the impact of IRM scaling on RMSE. In this simulation, the mean of the item parameters is fixed to zero. This serves to minimize differences in scaling between all conditions. However, if the estimation accuracy of the items is uneven then RMSE can be inflated. If, for example, 49 items are perfectly estimated and one item is egregiously overestimated, the mean of the set will shift to compensate and thus the locations of the accurate items will also shift. That is, the good items are still perfectly estimated, but they are on a different scale than the parent model.

The impact of scaling can be seen in Figure 7. This graph shows a sample of simulated item difficulties and their respective estimations from a Rasch model. While the ordering of the items is similar between the two graphs, on average the estimated items (red) have been knocked down about .5 to compensate for the handful of poorly estimated items. Adding back in the average bias (black line) makes it clear which items were poorly estimated.

Regardless, RMSE can be a useful measure of item estimation. Small values of RMSE still indicate good item estimation. Larger values of RMSE indicate either poor estimation of many items or significantly poor estimation of a single item.

Slope and Variance of the Slope. Proper estimation of the slope, or the average rate of change across time, is necessary for longitudinal anchoring to be an effective method of data analysis. Poor estimation of the latent trait scores will create bias in the average rate of change across the ten timepoints. If the unit measure of latent trait scores is comparable across models within a single anchoring method, then the rate of change over the ten timepoints should be comparable regardless of any scaling differences. Thus, recovery of the population slope also serves as a proxy measure of the accuracy of the latent trait scores. If the latent traits are estimated properly, but on a different scale, the average rate of change should still be preserved.

To produce the slope and slope variance estimates, the estimated latent trait scores from each timepoint are used to fit a growth curve model. Model estimation is conducted using the lavaan package in R (version 0.6.4, Rosseel, 2012). The values of the linear slope, intercept, and their respective variances are freely estimated. A lower bound of zero was set for the slope variance to prevent the estimation of negative

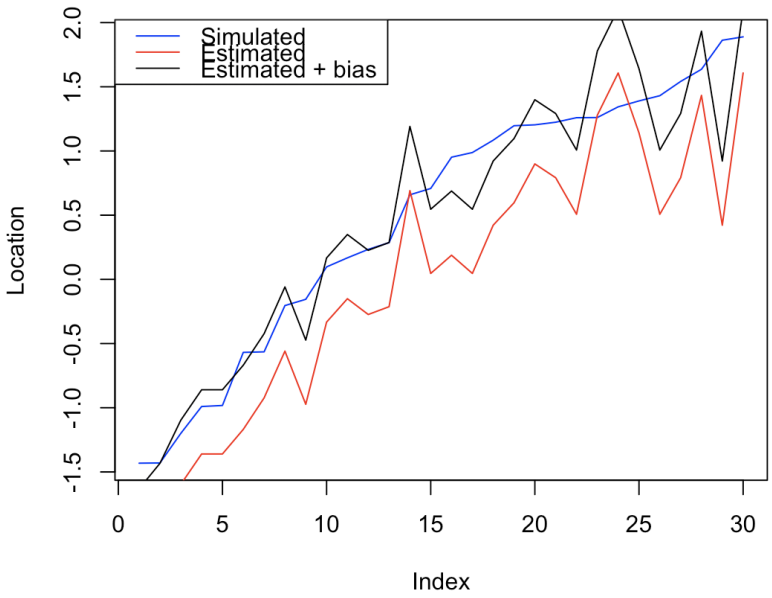

Figure 7. An example of how poorly estimated items throw off the estimation of all of other items. Items are ordered from (simulated) least to greatest locations for ease of interpretation.

variance.

Model Fit. Overall model fit is evaluated using the M_2, RMSEA, SRMR, and AIC. The M_2 statistic (Maydeu-Olivares & Joe, 2006) is a proxy for chi-squared goodness of fit testing for IRMs. M_2 evaluates fit based on marginal residuals and has more accurate Type 1 errors than chi-squared distributions for IRMs, making it an appropriate model fit index.

Root Mean Square Error of Approximation (RMSEA) was first introduced by Steiger (1989) with contributions by Browne and Cudeck (1993). It evaluates how well the model reproduces the observed covariance matrix. RMSEA can be defined as

$$RMSEA = sqrt\frac{T - df}{N - df}$$

where T is an asymptotically chi-square distributed test statistic based on residual covariances, df is the degrees of freedom, and N is the sample size (Maydeu-Olivares, Cai, & Hernández, 2011). Reasonable model fit is generally defined as RMSEA values less than 0.08, with very good model fit less than 0.05 (Browne & Cudeck, 1993).

Standardized Root Mean Square Residual (SRMR) is an absolute measure of model fit. SRMR measures the standardized difference between the residuals of the sample covariance matrix and the hypothesized model. As adapted from F. F. Chen (2007), SRMR is defined as

$$SRMR = \sqrt{\frac{2\sum\sum[(s_{ij} - \sigma_{ij})/(s_{ii}s_{jj})^2}{p(p+1)}}$$

where s_{ij} is the observed covariance, σ_{ij} is the model-implied covariance, s_{ii} and s_{jj} are the observed standard deviations, and p is the number of observed variables. An SRMR value of zero indicates perfect fit. The general cutoff for desirable SRMR fit is 0.08 (Hu & Bentler, 1999). Both RMSEA and SRMR can be impacted by sample sizes (DiStefano, 2002).

Akaike's Information Criteria (AIC) (Akaike, 1998) is a common information-based measure of model fit. AIC is defined as

$$AIC = 2k - 2ln(L)$$

where k is the model's degrees of freedom and L is the value of the likelihood. As AIC is a measure of the relative information that a model provides, there is no specific cutoff or desirable value. Rather, AIC is used comparatively between models where lower values indicate better fit. Anchoring methods that produce more accurate person and item estimations should produce lower values of AIC.

Simulation Results

Person Ability Estimates

For all conditions and methods, the simulated trait level at time one was zero. All methods estimated the mean trait level to be approximately zero at the first timepoint, except the All Times method (Table 1). For the All Times method, the average estimated mean trait level at time one was 0.43 for the 0.10 slope condition and 0.88 for 0.20 slope condition. For the 0.10 and 0.20 slope conditions, the mean population change in trait level across ten timepoints was simulated to be 1 and 2, respectively. The starting value for the All Times method under the two conditions was less than half of the total change in trait level over the ten timepoints. Sample trajectory plots from the Random anchoring method can be seen in Figures 8, 9, 10, and 11.

The standard deviation (SD) of the trait level was simulated to be 1 at the first timepoint. Under all methods and conditions, the estimated SD was approximately 0.95. This is slightly lower than the true simulated SD. This is may be due to the misestimation of extreme trait levels; trait levels can be truncated if they fall beyond the range of the items.

Standard Error of the Estimate

The standard error of the estimate (SEE) was similar across all conditions (see Figure 12). For the Zero Slope, No Variance condition, the mean SEE was approximately 0.325 for all methods. This increased slightly to 0.328 for the Zero Slope, with Variance condition. It was approximately 0.332 and 0.350 for the 0.10 and 0.20 slope conditions, respectively. Overall, the SEE increased slightly across conditions as

Estimated Population Parameters at Time One

	Zero Slope, zero variance	Zero Slope, w/ variance	0.10 Slope, w/ variance	0.20 Slope, w/ variance
	Mean of the Trait (SE)			
Floated	0.003 (0.017)	-0.015 (0.016)	-0.011 (0.018)	-0.004 (0.015)
Racked	0.003 (0.017)	-0.011 (0.018)	**0.433** (0.016)	**0.885** (0.015)
Time One	0.003 (0.017)	-0.011 (0.018)	-0.015 (0.016)	-0.004 (0.015)
Mean	0.002 (0.017)	-0.010 (0.018)	-0.014 (0.016)	-0.003 (0.015)
Random	0.002 (0.017)	-0.010 (0.018)	-0.014 (0.016)	-0.003 (0.015)
Stacked	0.002 (0.017)	-0.010 (0.018)	-0.015 (0.016)	-0.004 (0.015)
	Standard Deviation of the Trait (SE)			
Floated	0.953 (0.003)	0.947 (0.004)	0.950 (0.004)	0.949 (0.004)
Racked	0.953 (0.003)	0.950 (0.004)	0.952 (0.004)	0.951 (0.004)
Time One	0.953 (0.003)	0.947 (0.004)	0.950 (0.004)	0.949 (0.004)
Mean	0.952 (0.003)	0.944 (0.004)	0.946 (0.004)	0.945 (0.004)
Random	0.953 (0.003)	0.946 (0.004)	0.949 (0.004)	0.949 (0.004)
Stacked	0.951 (0.003)	0.945 (0.004)	0.947 (0.004)	0.946 (0.004)

Table 1

Average estimated population parameters of the latent trait at time one and the associated standard error. For all conditions, the true mean of the latent trait was 0 with a standard deviation of 1. All methods estimated the mean to be approximately zero, except for the Racked method (in bold).

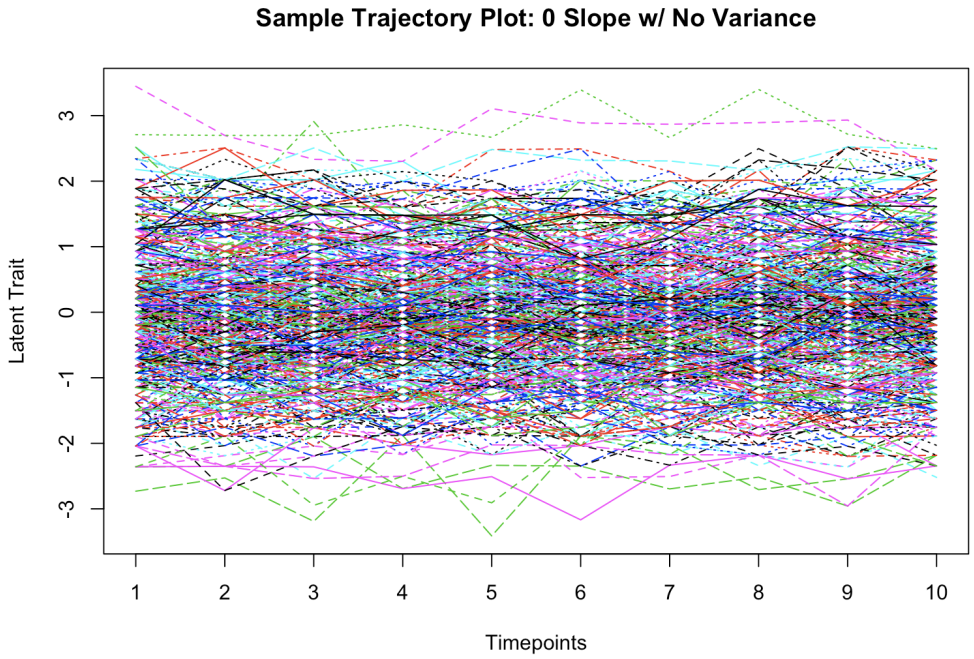

Figure 8. Participant trajectory plot under the No Slope, No Variance condition, using the Random anchoring method.

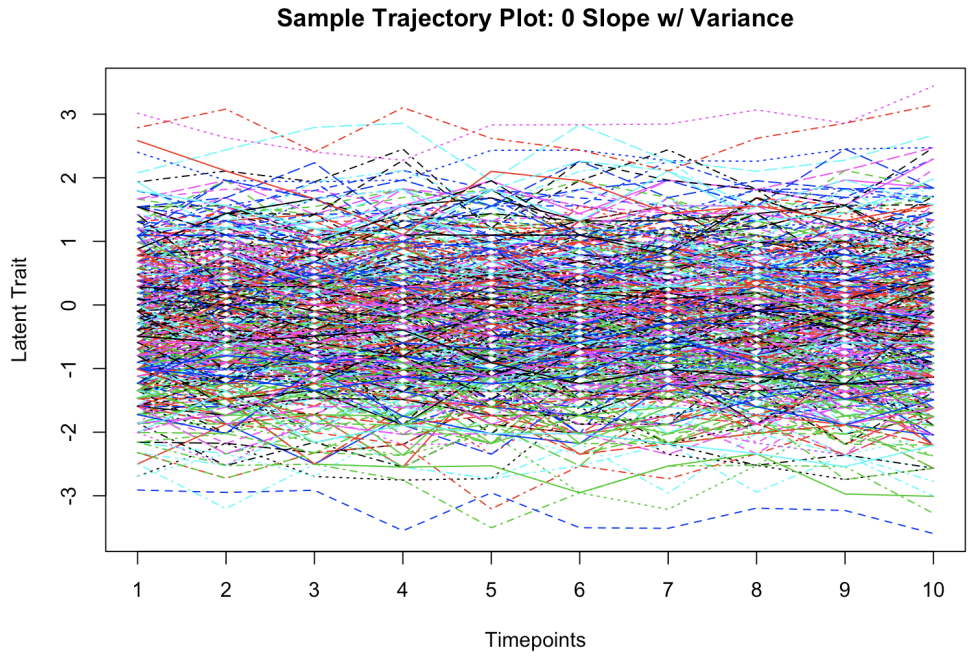

Figure 9. Participant trajectory plot under the No Slope, with Variance condition, using the Random anchoring method.

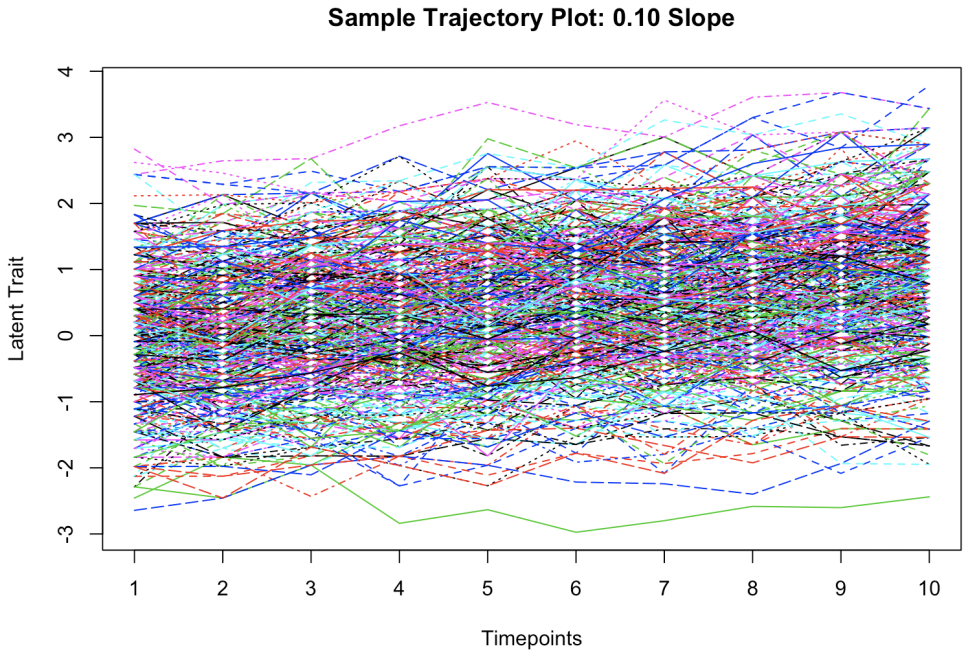

Figure 10. Participant trajectory plot under the 0.10 Slope, with Variance condition, using the Random anchoring method.

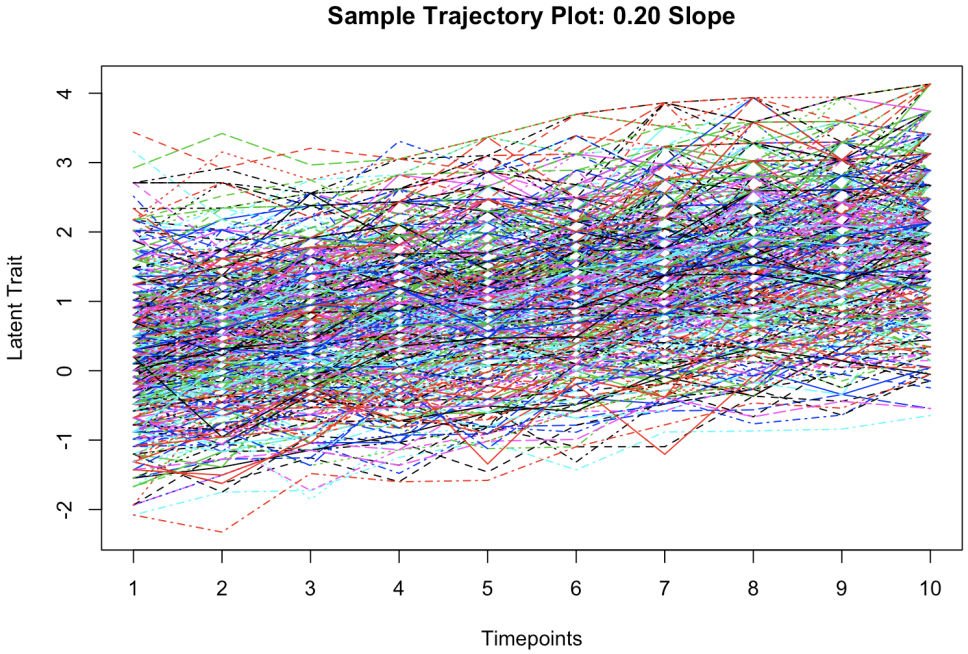

Figure 11. Participant trajectory plot under the 0.20 Slope, with Variance condition, using the Random anchoring method.

the trait levels became less well matched to the item range.

A similar pattern is seen for the minimum and maximum values of the SEE. The minimum SEE across simulations and conditions was 0.295. This varied very little across methods. The maximum SEE did vary by condition, but not by method. The maximum SEE rose from about 0.58 to 0.60 to 0.63 to 0.68 across the four change conditions. Within each condition, there was little variance between the anchoring methods. The change in maximum SEE is more related more to the change conditionals than the individual anchoring methods. There was little difference in SEE among the anchoring methods.

Item Parameter Recovery

RMSE was calculated by comparing the simulated item locations to the estimated item locations. Under the Time One, Mean, Random and Cross-sectional methods, item difficulties were not permitted to change over time. Item difficulties could change over time for the Floated and All Times methods. There is a consistent pattern of RMSE between methods across the simulation conditions (see Figure 13). The Mean and Cross-sectional methods consistently had the lowest median RMSE. The Random, Time One, and Floated methods performed less well than the Mean and Cross-sectional methods. The All Times method performed similarly to the Random, Time One, and Floated methods when there was zero mean change over time. Under the 0.1 and 0.2 conditions, RMSE values for the All Times method increased drastically. The average item RMSE was four times higher for the All Times method than the Mean method under the 0.20 slope condition. There is substantial variation of the RMSE, though the variation was generally consistent across conditions. The Mean and Cross-sectional methods had the largest variance of RMSE.

An examination of the item trajectory plots shows that the change in the person parameters was clearly expressed in the item parameters for the All Times method (Figure 14). When change occurred under the All Times method, the estimated item locations were far away from the simulated item locations. It is not the case that the

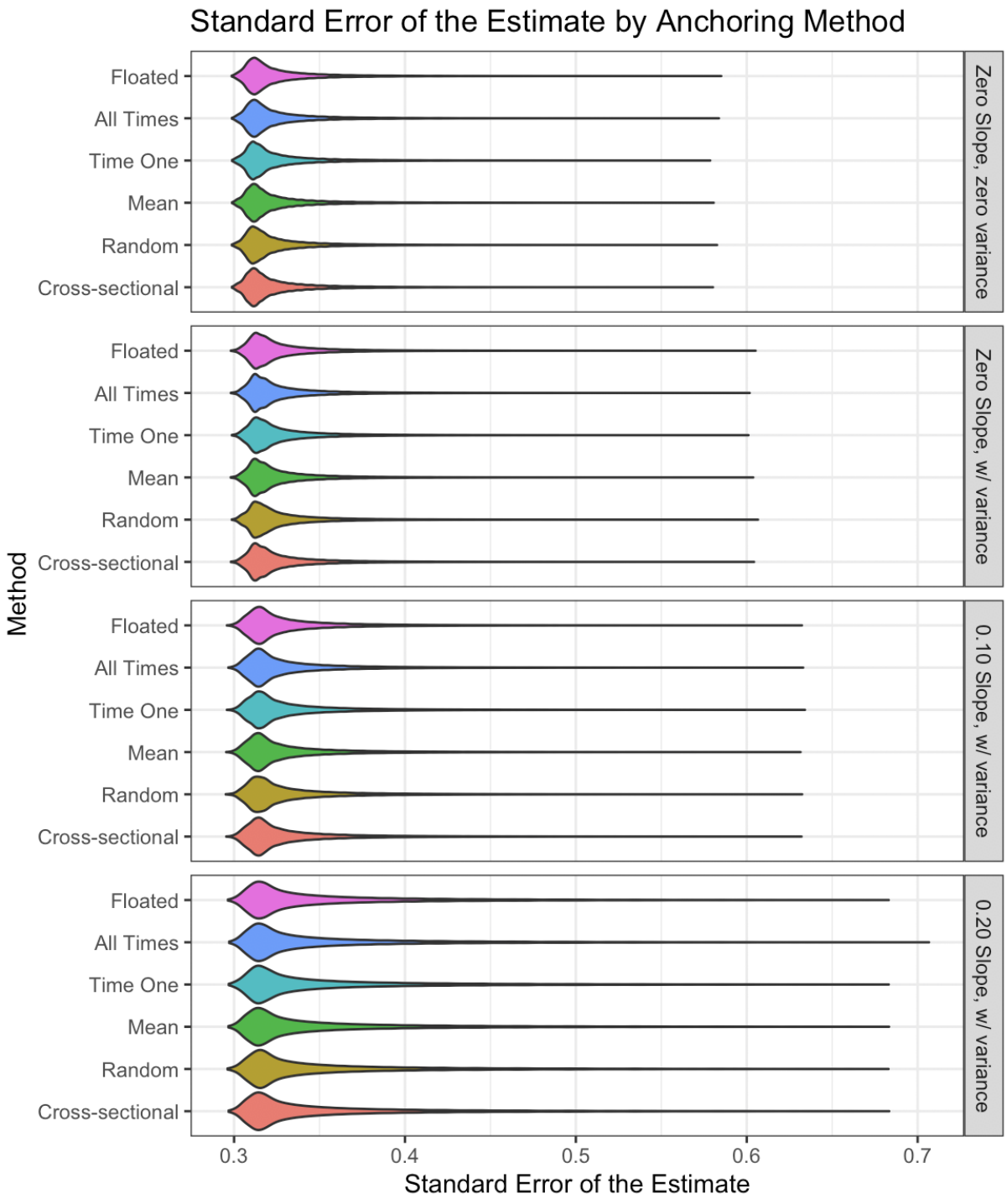

Figure 12. Distribution of the standard error of the estimates generated by condition and anchoring method. SEE was more strongly impacted by condition than anchoring method.

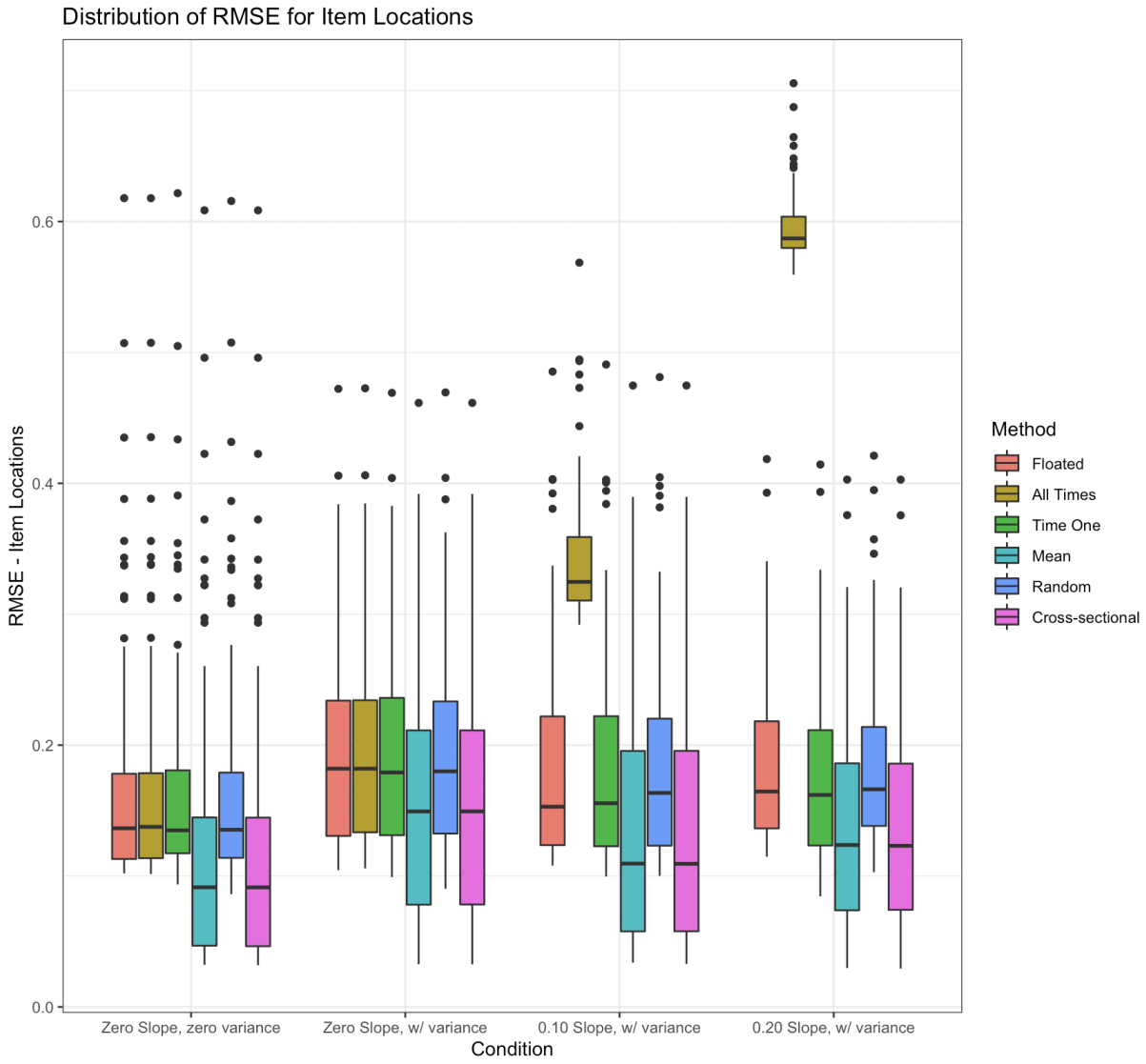

Figure 13. Shown here is the distribution of item location RMSE. The Mean and Cross-sectional methods had the lowest RMSE across all conditions, while the All Times method performed poorly when the population changed over time.

scale of the estimated items were different than the simulated items, but rather that the estimations were incorrect. Small variations in item difficulty can also be seen in the Floated method; despite these variations, item difficulty RMSE for the Floated method is comparable to the other anchoring methods.

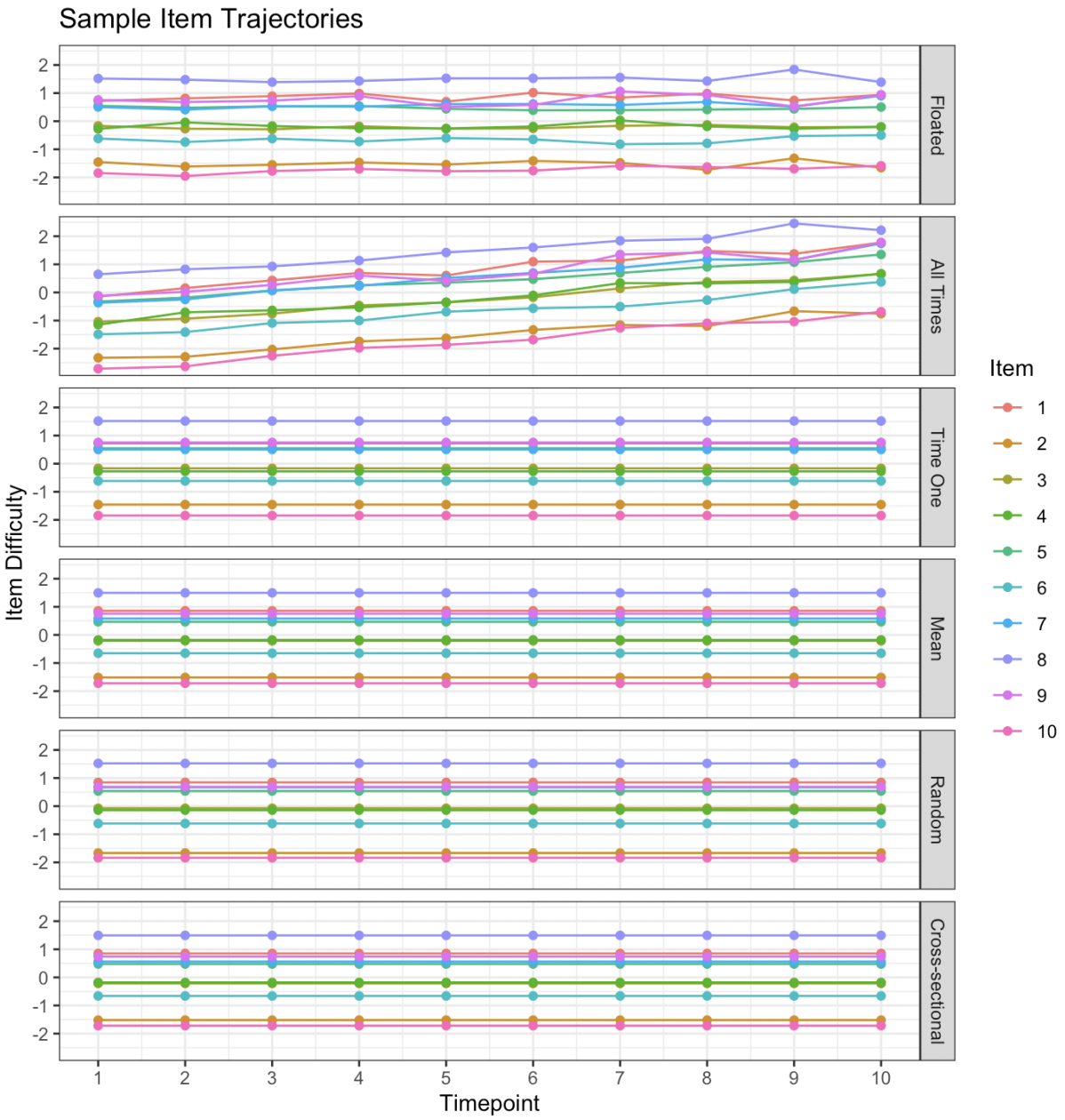

Figure 14. Sample item trajectories for the six anchoring methods under the 0.20 condition. While both the Floated and the All Times methods allow for variation in the item difficulties over time, only the All Times method expresses participant change in the item difficulties.

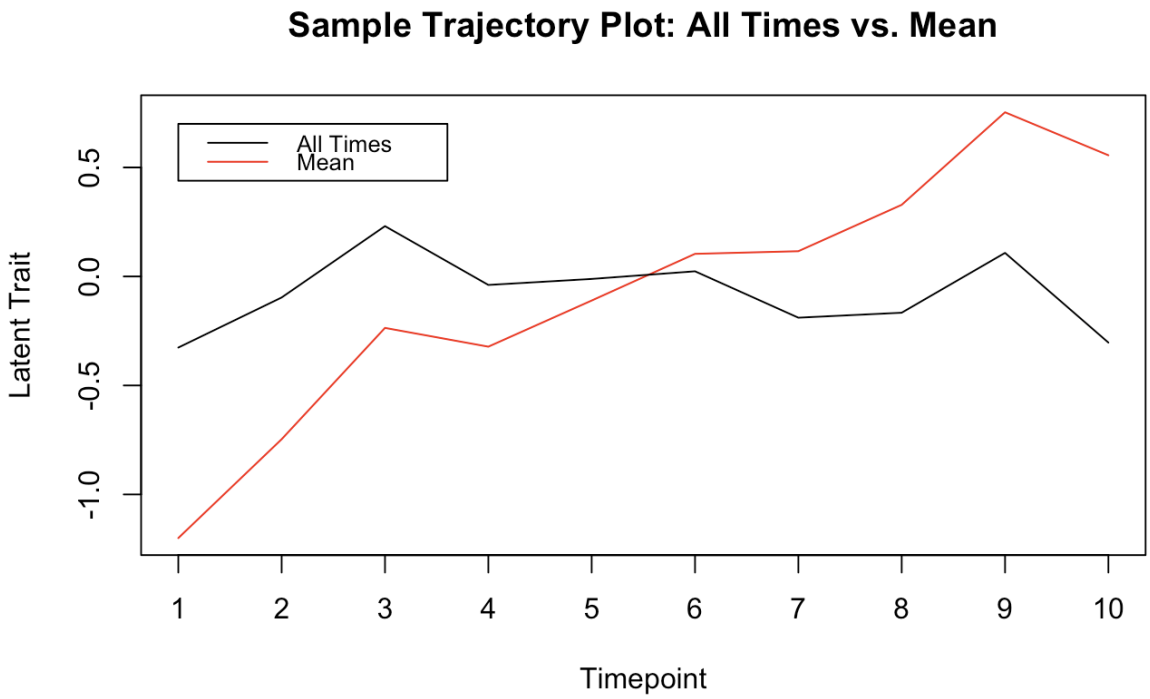

Figure 15. The trajectory of the same participant estimated by the Mean and All Waves anchoring methods. The All Wave method removes the participant's growth.

Slope and Slope Variance

A key concern in longitudinal data analysis is proper estimation of the average change over time. All six anchoring methods performed identically in recovering the mean slope when the simulated slope was zero (Table 2). When the mean slope was simulated to be 0.10 or 0.20, the Racked method failed completely. In both conditions, little to no overall change was detected using the Racked method.

Figure 15 shows the estimated latent trait of the same participant generated using two different anchoring methods. These results show that the Racked method did not truly estimate the slope to be zero during the Zero Slope conditions. Rather, no change was possible to be estimated. This is further supported by the lack of variance in the slope estimates; the variance of the estimate was approximately zero for all conditions. The other five methods successfully recovered simulated slopes of 0.10 and 0.20. The bias in slope recovery for the other methods was extremely small.

When the simulated slope variance was equal to zero, the estimated slope variance

was approximately zero under all anchoring methods. When the slope variance was simulated to be 0.0025 in the other three conditions, the variance of the slope was consistently underestimated regardless of anchoring method (Figure 16). The estimated variance for the No Slope, with Variance and the 0.10 slope conditions was closer to 0.0021 than 0.0025. The 0.20 slope condition fared the worst, with the average slope variance estimated to be less than 0.0020. Notably, there were larger differences between change conditions than anchoring methods when considering slope variance. Overall the anchoring methods themselves had little impact on the recovery of the slope variance.

Though the Racked method could not recover the true slope, the variance of the slope was preserved. That is, average change over time was expressed in the items rather than the persons, but variation around the average change was preserved. Some participants still had small positive or negative trajectories, though there was no estimated change over time at the population level.

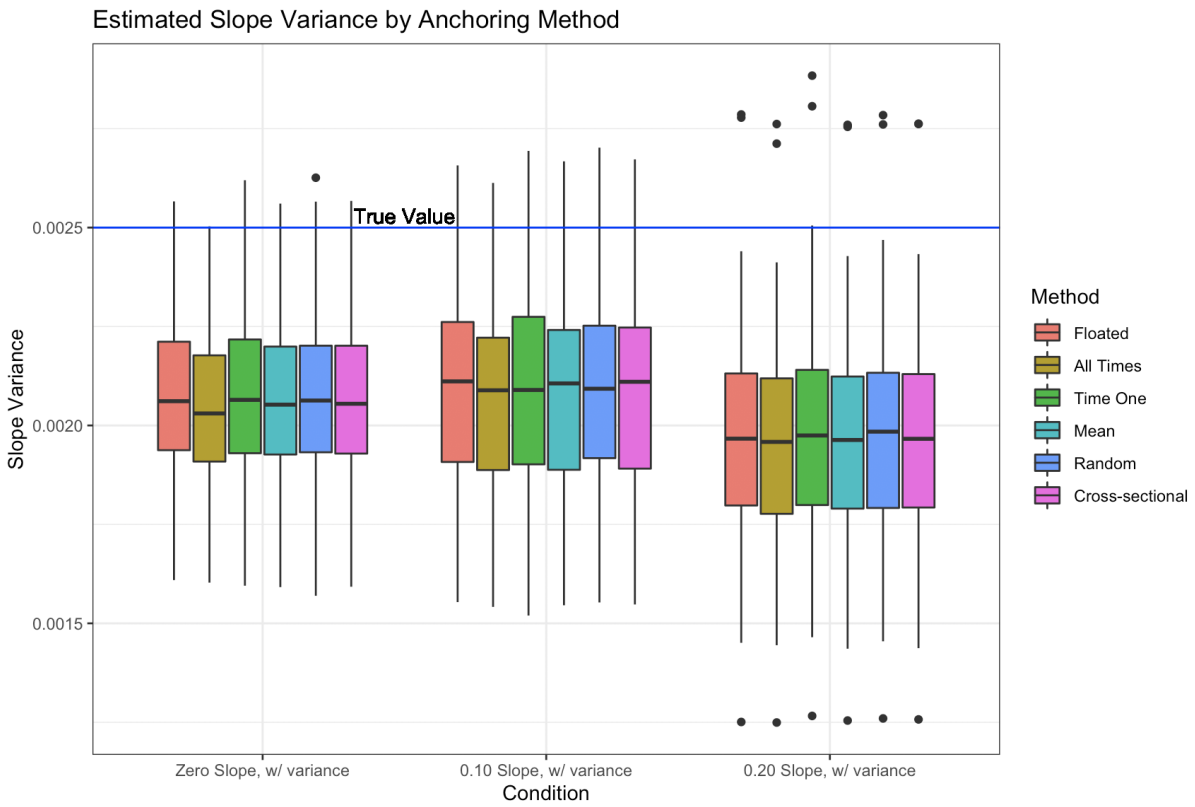

Figure 16. Estimated slope variance by condition and anchoring method. The slope variance was underestimated by all methods under all conditions. The No Slope, No Variance condition is not shown; the slope was estimated to be zero under all methods.

Overall Model Fit

The percentage of models showing significant misfit ($p<.05$) based on the M_2 statistic varied across the anchoring methods but was relatively stable across change conditions (Table 3). For the Mean, Cross-sectional, and Floated methods, model misfit was around or less than 6% for all conditions. The percentage of misfitting models for the All Times method was less than 1% for all four change conditions. In contrast, approximately one third of models misfit using the Random and Time One anchoring methods. The proportion of misfitting models was similar over the ten timepoints for all anchoring methods except for Time One (Figure 17). For the Time One method, significant model misfit was found only in timepoints 2 through 10.

RMSEA was extremely low, less than 0.01, for all anchoring methods under all conditions (Table 4). RMSEA was highest for the Time One and Random methods.

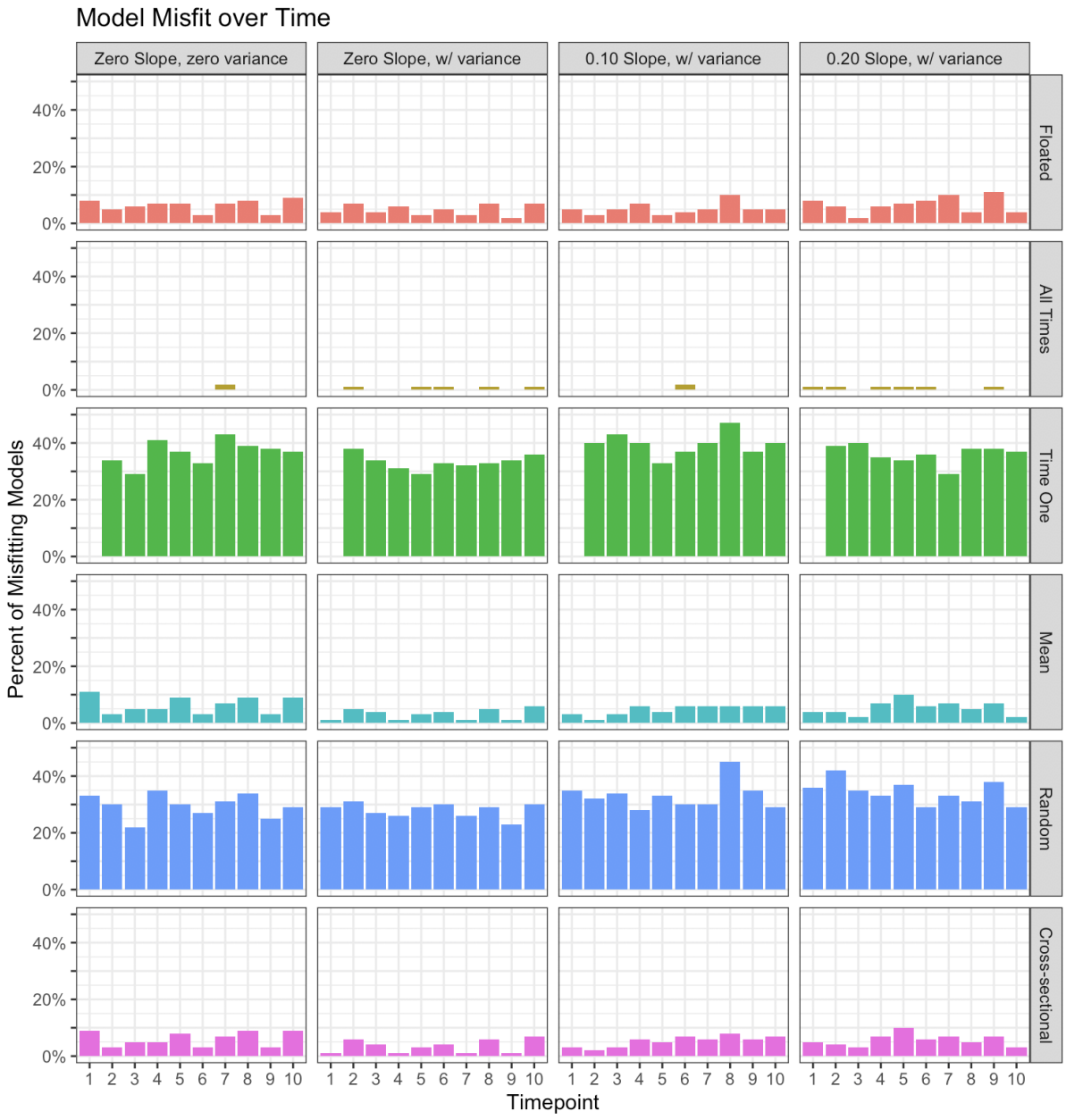

Figure 17. Model misfit measured by the M_2 statistic over the ten timepoints.

SRMR was approximately 0.04 for all anchoring methods under all conditions (Table 5). There was little variation in RMSEA or SRMR across methods and all values were below the appropriate cutoffs.

Across conditions, the overall pattern for AIC was generally consistent (Table 6). From lowest to highest mean AIC, the methods were: All Times, Cross-Sectional & Mean, Random & Time One, and Floated. The All Times method consistently had the lowest mean AIC values, and the Floated method had the highest. The Cross-sectional and Mean methods had similar mean values of AIC. The Time One and Random methods were also similar to each other, but the Time One method had lower AIC under the 0.20 slope condition.

For the no change conditions, AIC varied little over the ten timepoints. But, as can be see in Figure 18, there was a large decrease in AIC across the ten timepoints for the 0.10 and 0.20 conditions. This pattern can be understood in the specific context of IRMs. The more closely that persons and items match in latent trait, the more variability there is among person responses. Under the Rasch model, when a person has the same trait level as an item, they have a 50% chance of passing or failing the item. If a person's trait level is much higher than the item, they will almost always pass the item. The more that items and persons mismatch in skill level, the less variability there is in responses. Thus, AIC for models where persons and items are greatly mismatched will be lower than for models where the trait level of persons and items is very similar. This mismatch grew as person abilities grew over the ten timepoints.

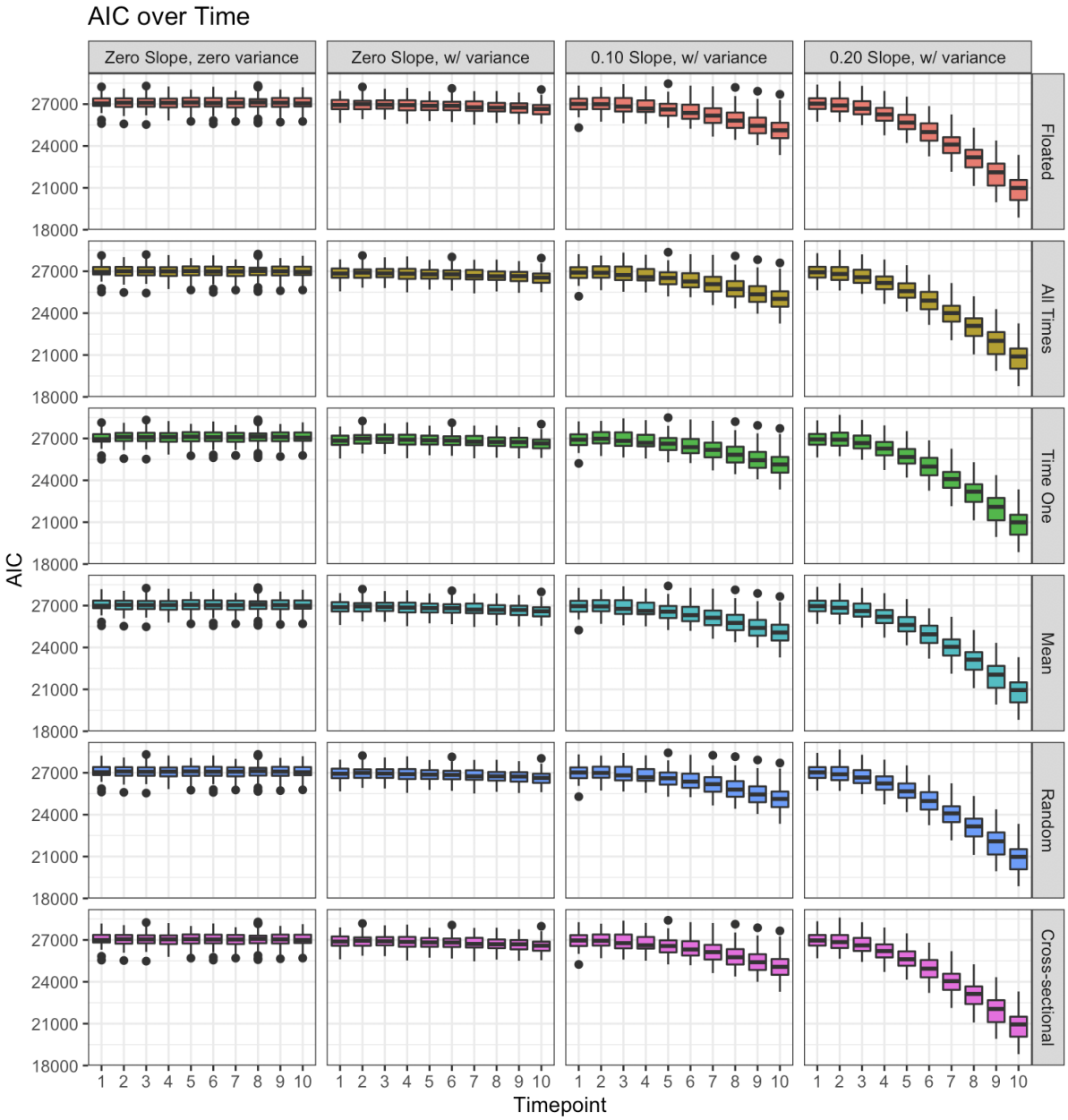

Figure 18. The distribution of AIC across the ten timepoints and conditions.

Bias in Slope Recovery

	Zero Slope, zero variance		Zero Slope, w/ variance		0.10 Slope, w/ variance		0.20 Slope, w/ variance	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Floated	-0.0002	*0.002*	-0.0001	*0.003*	0.0003	*0.003*	-0.0001	*0.003*
Racked	-0.0001	*0.000*	-0.0001	*0.000*	**-0.0989**	*0.000*	**-0.1975**	*0.000*
Time One	-0.0001	*0.002*	-0.0001	*0.003*	0.0003	*0.003*	-0.0003	*0.003*
Mean	-0.0001	*0.002*	-0.0001	*0.003*	-0.0001	*0.003*	-0.0010	*0.003*
Random	-0.0001	*0.002*	-0.0001	*0.003*	0.0002	*0.003*	-0.0004	*0.003*
Stacked	-0.0001	*0.002*	-0.0001	*0.003*	-0.0000	*0.003*	-0.0008	*0.003*

Table 2

Bias in the slope was calculated as estimate slope - simulated slope. Most bias was approximately zero, except for the Racked method (in bold). Negative values mean that the estimated slope was lower than the simulated slope.

M_2 Model Misfit Across Conditions and Methods

	Zero Slope, zero variance	Zero Slope, w/ variance	0.10 Slope, w/ variance	0.20 Slope, w/ variance
Floated	6.3%	4.8%	5.2%	6.6%
All Times	0.2%	0.5%	0.2%	0.6%
Time One	33.1%	30.0%	35.7%	32.6%
Mean	6.4%	3.1%	4.7%	5.4%
Random	29.6%	28.0%	33.1%	34.3%
Cross-sectional	6.1%	3.4%	5.3%	5.7%

Table 3

Model misfit, as measured by the M_2 statistic. For each condition and method, there were 10 timepoints for each of the 100 simulations; thus each reported percentages is out of 1,000 models.

Mean RMSEA Across Conditions and Methods

	Zero Slope, zero variance	Zero Slope, w/ variance	0.10 Slope, w/ variance	0.20 Slope, w/ variance
Floated	0.004	0.004	0.004	0.004
All Times	0.001	0.001	0.001	0.001
Time One	0.009	0.008	0.009	0.008
Mean	0.004	0.003	0.003	0.003
Random	0.008	0.008	0.008	0.009
Cross-sectional	0.003	0.003	0.003	0.003

Table 4

Mean RMSEA was very low for all methods and conditions. RMSEA was highest for the Time One and Random methods. All standard errors of RMSEA across conditions and methods were approximately 0.0001.

Mean SRMR Across Condtions and Methods

	Zero Slope, zero variance	Zero Slope, w/ variance	0.10 Slope, w/ variance	0.20 Slope, w/ variance
Floated	0.043	0.043	0.043	0.045
All Times	0.043	0.043	0.043	0.045
Time One	0.043	0.043	0.043	0.045
Mean	0.043	0.043	0.043	0.045
Random	0.043	0.043	0.044	0.045
Cross-sectional	0.043	0.043	0.043	0.045

Table 5

Mean SRMR was similar for all methods and conditions. All standard errors of SRMR across conditions and methods were approximately 0.00001.

Mean AIC Across Conditions and Methods

	Zero Slope, zero variance	Zero Slope, w/ variance	0.10 Slope, w/ variance	0.20 Slope, w/ variance
Floated	27097	26856	26381	24792
All Times	26997	26757	26282	24693
Time One	27086	26842	26370	24775
Mean	27041	26799	26326	24736
Random	27085	26845	26371	24780
Cross-sectional	27041	26799	26326	24736

Table 6

AIC varied between the anchoring methods. The All Times method always had the lowest AIC, while the Floated method always had the highest.

Discussion

Understanding how to properly anchor IRMs is critical when studying change over time. While there are many methods of anchoring IRMs, the challenge lies finding one that upholds important psychometric properties such as measurement invariance and local independence, while producing accurate estimations. Because modern IRMs are unable to handle more than 10 timepoints, new methods of producing person and item estimates and evaluating test function are necessary.

Most of the anchoring methods performed more or less similarly, except for the All Times method. For the other methods, there were only small differences in regards to the standard error of the estimate and in recovering the slope and the variance of the slope. All methods performed similarly when compared using RMSEA and SRMR. Differences arose when looking at the estimation of the item parameters, overall model fit based on the M_2 statistic, and AIC. For all of these criteria, the Mean and the Cross-sectional anchoring methods performed the best. Though the Random and the Time One methods were successful at recovering the population slope, nearly one-third of these models showed significant M_2 misfit.

The commonality of the Mean and Cross-sectional methods is that they both pool all the data to create the item difficulties. While this seems to be an obvious advantage as it provides more data for the item difficulties to be based on, the parent models also violate local independence. It remains to be seen if increasing levels of local independence reduces the efficacy of either of these methods, or if one method is more robust than the other.

The characteristics of certain anchoring methods may be better suited for different situations. For example, in the Mean method, the parent model requires estimation of n items x t time points, and p participants. In comparison, the Cross-sectional method requires estimation of p participants x t time points, and n items. If the number of participants is not much larger than the number of items, then the estimation of the items in the Mean method may be poor.

Despite the lesser performance of the Time One and Random methods in regards

to model fit and item estimation, these two methods were still able to recover the true change in the population ability parameters. While accuracy may be important when attempting to estimate "observed" variables for future growth curves, the Time One method is the simplest and least computationally intensive method, aside from floating all the models. If the Mean or Cross-sectional methods prove to be intractable to estimate, the Time One method may be a suitable option.

The Floated method performed adequately in this simulation, despite allowing all items to vary. Along with the Mean and Cross-sectional method, it had some of the lowest rates of M_2 misfit. This, though, speaks more to the performance of the Mean and Cross-sectional methods. M_2 misfit rates should be low for the Floated method as each Rasch model was run separately and without interference of the other timepoints. While this may over-specify the model to that subset of data, it should produce the best fit indices. The Mean and Cross-sectional methods performed similarly despite pooling data. In any case, the Floated method may only perform well when the data is relatively free of noise and other nuisance variables. If items showed differential functioning across timepoints, the resulting latent trait scores may not be comparable.

As hypothesized, the All Times method was an inappropriate choice when the mean trait level changed over time. The All Times method failed to recover any change in trait level in either the 0.10 or 0.20 condition. Change in the person abilities was expressed in the item difficulties, producing misleading fit statistics. The All Times method did not place the separate timepoint models all on the same scale, but rather warped the scale of the items

As evidenced by the All Times method, several methods of assessing model fit are not always reliable if a model has been accurately specified, especially for longitudinal IRMs. Incorrectly specified models can produce spuriously good fit indexes but with meaningless results. While the All Times method had comparatively good distributions of AIC, M_2 misfit, and standard error of the estimate, the All Times method produced misleading person and item estimates. The true population change over time was unable to be recovered from the person estimates, thus the All Times method should

never be recommended for longitudinal data. Even if there is not change or the data is not longitudinal, the All Times method would still be an inappropriate choice due to the shifting of variance from the persons to the items. Even the Floated method performed better than the All Times method at recovering slope, despite the low AIC values of the Floated models. The surprising "accuracy" of the All Times method among the fit indices may be specific to the one-parameter model. In the Rasch model, there is no difference in how the persons and the items are treated; it is possible to flip persons and items and receive the same outcome statistics. Thus, the All Times method makes no distinction between participants growing over time, and the items getting successively easier for the participants.

It is somewhat surprising that the Mean method performed so well when the All Times method performed so poorly. The Mean method is identical to the All Times method except that it involves the averaging of the item difficulties produced by the All Times method. As it has been demonstrated that the All Times method does not produce accurate item difficulties, the positive performance of the Mean method is unusual. Perhaps because the average true change over time was linear and invariable, averaging the item difficulties ameliorated the issues in estimation. It is possible that if change does not occur in a consistent, linear fashion that the Mean method would not perform well. It also cannot be determined at this time if any of the fit indices were similarly inflated for the Mean method as they were for the All Times method. Regardless, the Mean method did accurately recover the true population slope and item difficulties under the parameters of the present study.

There was little interaction between the change conditions and the anchoring methods. All methods performed similarly across the change conditions, except for the All Waves method. In this simulation, the size of the change did not largely influence the performance of the anchoring methods. It is unknown if different change patterns would have large influences on the performance of the anchoring methods.

Under all conditions and methods, the variance of the slope was underestimated. As this is true even for the zero mean slope condition, it is evidently not solely due to

the mismatch between persons and items. There were two sources of variance impacting the person parameters: variance from the slope and variance from the model. That is, the estimated person parameters possess error variance based on the log-odds of passing or failing items at a given ability level. This noise may have impacted the estimation of the slope variance. Yet, simultaneous estimation of longitudinal models can produce more accurate estimations of slope parameters. One drawback of longitudinal anchoring may be the underestimation of the slope variance, though this cannot be certain without directly comparing longitudinal anchoring to simultaneously estimated IRMs. In any case, most simultaneously estimated IRMs could not fit models with 10, or more, timepoints, making such a comparison difficult.

Future Directions

As with all simulation studies, there are many parameters that could be included in future analyses. First, there are variations of the anchoring methods that have not been evaluated. Currently, the Time One method anchors to the first timepoint of the data. Depending on the relative match between persons and items, another timepoint may be more or less suited to serve as the anchor. The Random method may benefit greatly by the use of bootstrapping. Currently, only one random sample of the data is used to estimate the item difficulties. A new conceptualization of the Random method could involve bootstrapping the difficulties by running the Random method multiple times and averaging the resulting item difficulties. One advantage of bootstrapping would be the reduction of potential bias from inadvertently using non-representative random samples.

In the inevitable presence of missing data, several of the anchoring methods may function differently. The Time One method presumes that the first instance of data collection has the most representative sample; this may not be true if the sample significantly changes due to attrition and subsequent refreshment. Issues of representation also arise for the Random method. If certain timepoints are over-represented in the data, then researchers must choose between using observations

from all participants and over-representing some timepoints, or representing all timepoints equally while under-representing some participants. While more data always sounds better, certain timepoints may be more or less reliable or representative. For example, participants may be more careless at the end of a study, or refreshment samples may not match the broader sample's characteristics. These issues apply not only to the Random and Time One method, but to all methods in general.

Future studies should examine parameters that are specific to longitudinal data, such as local dependence among persons and items and different levels of between and within person variance. The present study represents a relatively "clean" simulation which tests the anchoring methods themselves, rather than the limits of their performance under different circumstances. Local dependence is known to cause issues in IRMs. While the 2-step anchoring method described here may alleviate some of the issues caused by local dependence, it is unknown how large amounts of local dependence would influence the estimation of the parent models.

Finally, it is unknown if the anchoring methods perform similarly for more complex IRMs than the Rasch model. This applies both for the number of model parameters and the number of latent factors. One benefit of longitudinal anchoring is that if a multidimensional model can be run at one timepoint, longitudinal ability estimates can be obtained. This can allow longitudinal analyses to use more complex models. Overall, the present study is a validation of longitudinal anchoring for a specific set of circumstances; future research is necessary to evaluate longitudinal anchoring in a variety of situations.

Conclusions

Choosing the correct anchoring method for longitudinal IRMs is important in obtaining accurate latent trait estimations. Due to the current limitations in longitudinal IRMs, longitudinal anchoring may be a parsimonious solution for an otherwise intractable problem. Based on the results of the present simulation study, the Cross-sectional or Mean anchoring methods are recommended to be used. Further

research is necessary to determine under what circumstances either anchoring method is better than the other.

There are many benefits of establishing longitudinal anchoring as a method of longitudinal analysis. As large, multi-year longitudinal datasets come to fruition, new methods are needed to analyze data in the item response modeling framework. It is not currently feasible to run simultaneous, multi-timepoint longitudinal IRMs that contain complex latent factor structures, but it is well established that many psychological phenomenon cannot be captured through one dimensional modeling. In comparison, there is no known limit on the number of timepoints that can be analyzed using longitudinal anchoring. Should longitudinal anchoring prove itself as a reliable method of establishing latent trait estimations, researchers may be able to obtain estimates of complex multidimensional phenomena over time to further our understand of change.

Acknowledgements. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. 1842490. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

References

Ackerman, T. A., & Spray, J. A. (1986). A general model for item dependency. *ACT Research Report Series 87-9*, Iowa City, IA: ACT.

Adams, R. J., Wu, M. L., & Wilson, M. R. (2015). *ACER ConQuest: Generalised Item Response Modelling Software [Computer Software]. Version 4.* Camberwell, Victoria: Australian Council for Educational Research.

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In *Selected papers of Hirotugu Akaike* (pp. 199–213). Springer.

Andersen, E. B. (1985). Estimating latent correlations between repeated testings. *Psychometrika*, *50*(1), 3–16.

Andrade, D. F., & Tavares, H. R. (2005). Item response theory for longitudinal data: Population parameter estimation. *Journal of Multivariate Analysis*, *95*(1), 1–22.

Angoff, W. H. (1984). *Scales, norms, and equivalent scores.* Princeton, NJ: Educational Testing Service.

Anselmi, P., Vidotto, G., Bettinardi, O., & Bertolotti, G. (2015). Measurement of change in health status with Rasch models. *Health and Quality of Life Outcomes*, *13*(1).

Bacci, S. (2012). Longitudinal data: Different approaches in the context of item-response theory models. *Journal of Applied Statistics*, *39*(9), 2047–2065.

Bandalos, D. L., & Finney, S. J. (2001). Item parceling issues in structural equation modeling. In *New developments and techniques in structural equation modeling* (p. 269-275). Mahwah, NJ: Lawrence Erlbaum.

Barton, M. A., & Lord, F. M. (1981). An upper asymptote for the three-parameter logistic item-response model. *Research Bulletin 81-20*, Princeton, NJ: Educational Testing Service.

Bereiter, C. (1963). Some persisting dilemmas in the measurement of change. In C. W. Harris (Ed.), *Problems in measuring change: Proceedings of a conference sponsored by the committee on personality development in youth of the social science research council, 1962.* Madison, WI: The University of Wisconsin Press.

Bergeman, C., & Deboeck, P. R. (2014). Trait stress resistance and dynamic stress dissipation on health and well-being: The reservoir model. *Research in Human Development*, *11*(2), 108–125.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novice (Eds.), *Statistical theories of mental test scores* (pp. 397–472). Reading, MA: Addison-Wesley Publishing.

Blonigen, D. M., Hicks, B. M., Krueger, R. F., Patrick, C. J., & Iacono, W. G. (2005). Psychopathic personality traits: Heritability and genetic overlap with internalizing and externalizing psychopathology. *Psychological Medicine*, *35*(5), 637–648.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an em algorithm. *Psychometrika*, *46*(4), 443–459.

Boone, W. J., & Scantlebury, K. (2006). The role of Rasch analysis when conducting science education research utilizing multiple-choice tests. *Science Education*, *90*(2), 253–269.

Bowles, R. P., & Salthouse, T. A. (2003). Assessing the age-related effects of proactive interference on working memory tasks using the Rasch model. *Psychology and Aging*, *18*(3), 608.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & K. S. Long (Eds.), *Testing structural equation models* (p. 136–162). Sage publications.

Busemeyer, J. R., & Jones, L. E. (1983). Analysis of multiplicative combination rules when the causal variables are measured with error. *Psychological Bulletin*, *93*(3), 549–562.

Cai, L. (2010a). Metropolis-hastings robbins-monro algorithm for confirmatory item factor analysis. *Journal of Educational and Behavioral Statistics*, *35*(3), 307–335.

Cai, L. (2010b). A two-tier full-information item factor analysis model with applications. *Psychometrika*, *75*(4), 581–612.

Cai, L., Thissen, D., & du Toit, S. (2019). *IRTPRO 4.20 for Windows*. Chicago, IL: Scientific Software International.

Carlson, M., McLanahan, S., & England, P. (2004). Union formation in fragile families. *Demography*, *41*(2), 237–261.

Chalmers, R. P., et al. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29.

Chang, W.-C., & Chan, C. (1995). Rasch analysis for outcomes measures: Some methodological considerations. *Archives of Physical Medicine and Rehabilitation*, *76*(10), 934–939.

Chao, A., Tsay, P., Lin, S.-H., Shau, W.-Y., & Chao, D.-Y. (2001). The applications of capture-recapture models to epidemiological data. *Statistics in Medicine*, *20*(20), 3123–3157.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(3), 464–504.

Chen, W., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, *22*(3), 265–289.

Chiang, K. S., Green, K. E., & Cox, E. O. (2009). Rasch analysis of the Geriatric Depression Scale-Short Form. *The Gerontologist*, *49*(2), 262–275.

Cliff, N. (1989). Ordinal consistency and ordinal true scores. *Psychometrika*, *54*(1), 75–91.

Cook, L. L., & Eignor, D. R. (1991). IRT equating methods. *Educational Measurement: Issues and Practice*, *10*(3), 37–45.

Cronbach, L. J., & Furby, L. (1970). How we should measure" change": Or should we? *Psychological Bulletin*, *74*(1), 68.

Curran, P. J., Hussong, A. M., Cai, L., Huang, W., Chassin, L., Sher, K. J., & Zucker, R. A. (2008). Pooling data from multiple longitudinal studies: the role of item response theory in integrative data analysis. *Developmental Psychology*, *44*(2), 365-380.

Davison, M. L., & Sharma, A. R. (1988). Parametric statistics and levels of

measurement. *Psychological Bulletin*, *104*(1), 137–144.

Davison, M. L., & Sharma, A. R. (1990). Parametric statistics and levels of measurement: Factorial designs and multiple regression. *Psychological Bulletin*, *107*(3), 394.

Deacon, S. H., Kieffer, M. J., & Laroche, A. (2014). The relation between morphological awareness and reading comprehension: Evidence from mediation and longitudinal models. *Scientific Studies of Reading*, *18*(6), 432–451.

De Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: Guilford Publications.

DiStefano, C. (2002). The impact of categorization with confirmatory factor analysis. *Structural Equation Modeling*, *9*(3), 327–346.

du Toit, M. (2003). *IRT from SSI: BILOG-MG, MULTILOG, PARSCALE, TESTFACT*. Chicago, IL: Scientific Software International.

Ehringer, M. A., Rhee, S. H., Young, S., Corley, R., & Hewitt, J. K. (2006). Genetic and environmental contributions to common psychopathologies of childhood and adolescence: A study of twins and their siblings. *Journal of Abnormal Child Psychology*, *34*(1), 1–17.

Eid, M., & Diener, E. (1999). Intraindividual variability in affect: Reliability, validity, and personality correlates. *Journal of Personality and Social Psychology*, *76*(4), 662-676.

Embretson, S. E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika*, *56*(3), 495–515.

Embretson, S. E. (1992). Measuring and validating cognitive modifiability as an ability: A study in the spatial domain. *Journal of Educational Measurement*, *29*(1), 25–50.

Embretson, S. E. (1994). Comparing changes between groups: Some perplexities arising from psychometrics. In D. Laveault, B. D. Zumbo, M. E. Gessaroli, & M. W. Boss (Eds.), (pp. 213–248). Ottawa, Canada: Edumetrics Research Group, University of Ottawa.

Embretson, S. E. (1996). Item response theory models and spurious interaction effects in factorial ANOVA designs. *Applied Psychological Measurement*, *20*(3), 201–212.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Psychology Press.

Erbacher, M. K., Schmidt, K. M., Boker, S. M., & Bergeman, C. S. (2012). Measuring positive and negative affect in older adults over 56 days: Comparing trait level scoring methods using the partial credit model. *Journal of Applied Measurement*, *13*(2), 146.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, *37*(6), 359–374.

Fischer, G. H. (1983). Some latent trait models for measuring change in qualitative observations. In D. J. Weiss (Ed.), *New horizon testing: Latent trait test theory and computerized adaptive testing* (pp. 309–329). Academic Press.

Fischer, G. H. (1989). An IRT-based model for dichotomous longitudinal data. *Psychometrika*, *54*(4), 599–624.

Fischer, G. H. (1995). Derivations of the Rasch model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (chap. 2). Berlin, Germany: Springer-Verlag.

Fischer, L., Gnambs, T., Rohm, T., & Carstensen, C. H. (2019). Longitudinal linking of Rasch-model-scaled competence tests in large-scale assessments: A comparison and evaluation of different linking methods and anchoring designs based on two tests on mathematical competence administered in grades 5 and 7. *Psychological Test and Assessment Modeling*, *61*(1), 37–64.

Flaherty, E. G., Thompson, R., Litrownik, A. J., Theodore, A., English, D. J., Black, M. M., ... Dubowitz, H. (2006). Effect of early childhood adversity on child health. *Archives of Pediatrics & Adolescent Medicine*, *160*(12), 1232–1238.

Garcia, A. G., Lambert, M. C., Epstein, M. H., & Cullinan, D. (2018). Rasch analysis of the Emotional and Behavioral Screener. *School Mental Health*, 1–12.

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bi-factor analysis.

Psychometrika, *57*(3), 423–436.

Glas, C. A., Geerlings, H., van de Laar, M. A., & Taal, E. (2009). Analysis of longitudinal randomized clinical trials using item response models. *Contemporary Clinical Trials*, *30*(2), 158–170.

Goldstein, H. (1983). Measuring changes in educational attainment over time: Problems and possibilities. *Journal of Educational Measurement*, 369–377.

Gorter, R., Fox, J.-P., & Twisk, J. W. (2015). Why item response theory should be used for longitudinal questionnaire data analysis in medical research. *BMC Medical Research Methodology*, *15*(1), 55.

Harwell, M. R., & Gatti, G. G. (2001). Rescaling ordinal data to interval data in educational research. *Review of Educational Research*, *71*(1), 105–131.

Hopwood, C. J., Donnellan, M. B., Blonigen, D. M., Krueger, R. F., McGue, M., Iacono, W. G., & Burt, S. A. (2011). Genetic and environmental influences on personality trait stability and growth during the transition to adulthood: A three-wave longitudinal study. *Journal of Personality and Social Psychology*, *100*(3), 545.

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research*, *18*(3), 117–144.

Hoskens, M., & De Boeck, P. (2001). Multidimensional componential item response theory models for polytomous items. *Applied Psychological Measurement*, *25*(1), 19–37.

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

Iacono, W. G., Carlson, S. R., Taylor, J., Elkins, I. J., & McGue, M. (1999). Behavioral disinhibition and the development of substance-use disorders: Findings from the Minnesota Twin Family Study. *Development and Psychopathology*, *11*(4), 869–900.

Iacono, W. G., & McGue, M. (2002). Minnesota Twin Family Study. *Twin Research and Human Genetics*, *5*(5), 482–487.

Janssen, K. C., Phillipson, S., O'Connor, J., & Johns, M. W. (2017). Validation of the Epworth Sleepiness Scale for children and adolescents using Rasch analysis. *Sleep Medicine*, *33*, 30–35.

Jeon, M., & Rabe-Hesketh, S. (2016). An autoregressive growth model for longitudinal item analysis. *Psychometrika*, *81*(3), 830–850.

Jodoin, M. G., Keller, L. A., & Swaminathan, H. (2003). A comparison of linear, fixed common item, and concurrent parameter estimation equating procedures in capturing academic growth. *The Journal of Experimental Education*, *71*(3), 229–250.

Johnson, M. S., et al. (2007). Marginal maximum likelihood estimation of item response models in R. *Journal of Statistical Software*, *20*(10), 1–24.

Johnson, R. M., Kotch, J. B., Catellier, D. J., Winsor, J. R., Dufort, V., Hunter, W., & Amaya-Jackson, L. (2002). Adverse behavioral and emotional outcomes from child abuse and witnessed violence. *Child Maltreatment*, *7*(3), 179–186.

Jones, E., & Bergin, C. (2019). Evaluating teacher effectiveness using classroom observations: A Rasch analysis of the rater effects of principals. *Educational Assessment*, 1–28.

Kahler, C. W., Hustad, J., Barnett, N. P., Strong, D. R., & Borsari, B. (2008). Validation of the 30-day version of the Brief Young Adult Alcohol Consequences Questionnaire for use in longitudinal studies. *Journal of Studies on Alcohol and Drugs*, *69*(4), 611–615.

Kang, S.-M., & Waller, N. G. (2005). Moderated multiple regression, spurious interaction effects, and IRT. *Applied Psychological Measurement*, *29*(2), 87–105.

Karabatsos, G. (2000). A critique of Rasch residual fit statistics. *Journal of Applied Measurement*, *1*(2), 152–176.

Kaufman, . K. N. L., A. S. (2004). *Kaufman Assessment Battery for Children, Second edition.* Circle Pine, MN: American Guidance Service.

Kendel, F., Wirtz, M., Dunkel, A., Lehmkuhl, E., Hetzer, R., & Regitz-Zagrosek, V. (2010). Screening for depression: Rasch analysis of the dimensional structure of

the PHQ-9 and the HADS-D. *Journal of Affective Disorders*, *122*(3), 241–246.

Koenig, L. B., McGue, M., & Iacono, W. G. (2008). Stability and change in religiousness during emerging adulthood. *Developmental Psychology*, *44*(2), 532–543.

Kotch, J. B., Lewis, T., Hussey, J. M., English, D., Thompson, R., Litrownik, A. J., ... Dubowitz, H. (2008). Importance of early neglect for childhood aggression. *Pediatrics*, *121*(4), 725–731.

Kubinger, K. D. (2008). On the revival of the Rasch model-based LLTM: From constructing tests using item generating rules to measuring item administration effects. *Psychology Science*, *50*(3), 311.

Lewis, T., Kotch, J., Proctor, L., Thompson, R., English, D., Smith, J., ... Dubowitz, H. (2019). The role of emotional abuse in youth smoking. *American Journal of Preventive Medicine*, *56*(1), 93–99.

Li, C., Velozo, C., Hong, I., Li, C., Newman, J., & Krause, J. (2017). Generating rasch-based activity of daily living measures from the spinal cord injury longitudinal aging study. *Spinal Cord*, *56*(1), 14–21.

Li, H., & Hong, F. (2001). Cluster-rasch models for microarray gene expression data. *Genome Biology*, *2*(8).

Linacre, J. M. (2019). *Winsteps® Rasch measurement computer program.* Beaverton, Oregon: Winsteps.com.

Litrownik, A. J., Newton, R., Hunter, W. M., English, D., & Everson, M. D. (2003). Exposure to family violence in young at-risk children: A longitudinal look at the effects of victimization and witnessed physical and psychological aggression. *Journal of Family Violence*, *18*(1), 59–73.

Little, T. D., Rhemtulla, M., Gibson, K., & Schoemann, A. M. (2013). Why the items versus parcels controversy needn't be one. *Psychological methods*, *18*(3), 285-300.

Liu, L. C., & Hedeker, D. (2006). A mixed-effects regression model for longitudinal multivariate ordinal data. *Biometrics*, *62*(1), 261–268.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.*

Addison-Wesley: Reading, MA.

Luce, R. D., & Tukey, J. W. (1964). Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology*, *1*, 1–27.

MacKenzie, M. J., Nicklas, E., Brooks-Gunn, J., & Waldfogel, J. (2011). Who spanks infants and toddlers? evidence from the Fragile Families and Child Well-Being Study. *Children and Youth Services Review*, *33*(8), 1364–1373.

Mallinson, T. (2011). Rasch analysis of repeated measures. *Rasch Measurement Transactions*, *251*(1), 1317.

Marais, I. (2009). Response dependence and the measurement of change. *Journal of Applied Measurement*, *10*(1), 17–29.

Martin, W. L., McKelvie, A., & Lumpkin, G. T. (2016). Centralization and delegation practices in family versus non-family SMEs: a Rasch analysis. *Small Business Economics*, *47*(3), 755–769.

Maydeu-Olivares, A., Cai, L., & Hernández, A. (2011). Comparing the fit of item response theory and factor analysis models. *Structural Equation Modeling: A Multidisciplinary Journal*, *18*(3), 333–356.

Maydeu-Olivares, A., & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika*, *71*(4), 713-732.

McArdle, J. J., Grimm, K. J., Hamagami, F., Bowles, R. P., & Meredith, W. (2009). Modeling life-span growth curves of cognition using longitudinal data with multiple samples and changing scales of measurement. *Psychological Methods*, *14*(2), 126.

Michell, J. (2009). The psychometricians' fallacy: Too clever by half? *British Journal of Mathematical and Statistical Psychology*, *62*(1), 41–55.

Misajon, R., Pallant, J., & Bliuc, A.-M. (2016). Rasch analysis of the Personal Wellbeing Index. *Quality of Life Research*, *25*(10), 2565–2569.

Mislevy, R. J., Johnson, E. G., & Muraki, E. (1992). Chapter 3: Scaling procedures in NAEP. *Journal of Educational Statistics*, *17*(2), 131–154.

Morse, B. J., Johanson, G. A., & Griffeth, R. W. (2012). Using the graded response

model to control spurious interactions in moderated multiple regression. *Applied Psychological Measurement*, *36*(2), 122–146.

Muthén, L., & Muthén, B. (1998-2017). *Mplus user's guide. eighth edition.* Los Angeles, CA: Muthén & Muthén.

Norquist, J. M., Fitzpatrick, R., Dawson, J., & Jenkinson, C. (2004). Comparing alternative Rasch-based methods vs raw scores in measuring change in health. *Medical Care*, I25–I36.

OECD. (2006). *PISA 2006 Technical Report.* Paris, France: Organization for Economic Co-operation and Development.

Olsbjerg, M., & Christensen, K. B. (2015). Modeling local dependence in longitudinal irt models. *Behavior Research Methods*, *47*(4), 1413–1424.

Paek, I., Park, H.-J., Cai, L., & Chi, E. (2014). A comparison of three IRT approaches to examinee ability change modeling in a single-group anchor test design. *Educational and Psychological Measurement*, *74*(4), 659–676.

Pasternak, A., Sideridis, G., Fragala-Pinkham, M., Glanzman, A. M., Montes, J., Dunaway, S., ... others (2016). Rasch analysis of the Pediatric Evaluation of Disability Inventory–computer adaptive test (pedi-cat) item bank for children and young adults with spinal muscular atrophy. *Muscle & Nerve*, *54*(6), 1097–1107.

Pastor, D. A., & Beretvas, S. N. (2006). Longitudinal Rasch modeling in the context of psychotherapy outcomes assessment. *Applied Psychological Measurement*, *30*(2), 100–120.

Penner, L. A., Shiffman, S., Paty, J. A., & Fritzsche, B. A. (1994). Individual differences in intraperson variability in mood. *Journal of Personality and Social Psychology*, *66*(4), 712-721.

Perline, R., Wright, B. D., & Wainer, H. (1979). The Rasch model as additive conjoint measurement. *Applied Psychological Measurement*, *3*(2), 237–255.

Pillas, M., Selai, C., & Schrag, A. (2017). Rasch analysis of the carers quality of life questionnaire for parkinsonism. *Movement Disorders*, *32*(3), 463–466.

Prieto, L., Novick, D., Sacristan, J., Edgell, E., Alonso, J., & Group, S. S. (2003). A

rasch model analysis to test the cross-cultural validity of the EuroQoL-5D in the Schizophrenia Outpatient Health Outcomes Study. *Acta Psychiatrica Scandinavica*, *107*, 24–29.

R Core Team. (2013). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org`

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Oxford, England: Nielsen & Lydiche.

Reichman, N. E., Teitler, J. O., Garfinkel, I., & McLanahan, S. S. (2001). Fragile Families: Sample and design. *Children and Youth Services Review*, *23*(4-5), 303–326.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. Retrieved from `http://www.jstatsoft.org/v48/i02/`

Runyan, D. K., Curtis, P. A., Hunter, W. M., Black, M. M., Kotch, J. B., Bangdiwala, S., ... Landsverk, J. (1998). LONGSCAN: A consortium for longitudinal studies of maltreatment and the life course of children. *Aggression and Violent Behavior*, *3*(3), 275–285.

Rusch, T., & Hatzinger, R. (2009). IRT models with relaxed assumptions in eRm: A manual-like instruction. *Psychology Science Quarterly*, *51*(1), 87–120.

Russell, C. J., Pinto, J. K., & Bobko, P. (1991). Appropriate moderated regression and inappropriate research strategy: A demonstration of information loss due to scale coarseness. *Applied Psychological Measurement*, *15*(3), 257–266.

Salzberger, T. (2010). Does the Rasch model convert an ordinal scale into an interval scale. *Rasch Measurement Transactions*, *24*(2), 1273–1275.

Schneider, W., MacKenzie, M., Waldfogel, J., & Brooks-Gunn, J. (2015). Parent and child reporting of corporal punishment: New evidence from the Fragile Families and Child Wellbeing Study. *Child Indicators Research*, *8*(2), 347–358.

StataCorp. (2017). *Stata statistical software: Release 15.* College Station, TX:

StataCorp LLC.

Steiger, J. H. (1989). *Ezpath: causal modeling: a supplementary module for systat and sygraph: Pc-ms-dos, version 1.0 [computer program manual].* Evanston, IL: Systat.

Stenner, A. J. (1996). *Measuring reading comprehension with the lexile framework.* (Tech. Rep.). MetaMetrics.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*(2684), 677-680.

Sturrock, B. A., Xie, J., Holloway, E. E., Lamoureux, E. L., Keeffe, J. E., Fenwick, E. K., & Rees, G. (2015). The influence of coping on vision-related quality of life in patients with low vision: A prospective longitudinal study. *Investigative Ophthalmology & Visual Science*, *56*(4), 2416–2422.

Svetina, D., Crawford, A. V., Levy, R., Green, S. B., Scott, L., Thompson, M., ... Kunze, K. L. (2013). Designing small-scale tests: A simulation study of parameter recovery with the 1-PL. *Psychological Test and Assessment Modeling*, *55*(4), 335.

Tavares, H. R., & Andrade, D. F. (2001). *Item response theory for longitudinal data: Item parameter estimation* (Tech. Rep.). Sao Paulo, Brazil.

Testa, I., Capasso, G., Colantonio, A., Galano, S., Marzoli, I., Scotti di Uccio, U., ... Zappia, A. (2018). Development and validation of a university students' progression in learning quantum mechanics through exploratory factor analysis and Rasch analysis. *International Journal of Science Education*, 1–30.

Thompson, R., Briggs, E., English, D. J., Dubowitz, H., Lee, L.-C., Brody, K., ... Hunter, W. M. (2005). Suicidal ideation among 8-year-olds who are maltreated and at risk: Findings from the LONGSCAN studies. *Child Maltreatment*, *10*(1), 26–36.

Tomyn, A. J., Stokes, M. A., Cummins, R. A., & Dias, P. C. (2019). A Rasch analysis of the Personal Well-Being Index in school children. *Evaluation & the Health Professions*, 0163278718819219.

Turney, K., & Wildeman, C. (2015). Detrimental for some? heterogeneous effects of

maternal incarceration on child wellbeing. *Criminology & Public Policy*, *14*(1), 125–156.

Velleman, P. F., & Wilkinson, L. (1993). Nominal, ordinal, interval, and ratio typologies are misleading. *The American Statistician*, *47*(1), 65–72.

Verguts, T., & De Boeck, P. (2000). A Rasch model for detecting learning while solving an intelligence test. *Applied Psychological Measurement*, *24*(2), 151–162.

Vock, M., & Holling, H. (2008). The measurement of visuo-spatial and verbal-numerical working memory: Development of IRT-based scales. *Intelligence*, *36*(2), 161–182.

von Davier, A. A., Carstensen, C. H., & von Davier, M. (2006). Linking competencies in educational settings and measuring growth. *ETS Research Report Series*(1), i–36.

Von Davier, M., Xu, X., & Carstensen, C. H. (2011). Measuring growth in a longitudinal large-scale assessment with a general latent variable model. *Psychometrika*, *76*(2), 318–336.

Waldfogel, J., Craigie, T.-A., & Brooks-Gunn, J. (2010). Fragile Families and Child Wellbeing. *Future Child*, *20*(2), 87–112.

Waldron, J. S., Malone, S. M., McGue, M., & Iacono, W. G. (2018). A co-twin control study of the relationship between adolescent drinking and adult outcomes. *Journal of Studies on Alcohol and Drugs*, *79*(4), 635–643.

Wang, C., Kohli, N., & Henn, L. (2016). A second-order longitudinal model for binary outcomes: Item response theory versus structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(3), 455–465.

Wang, W., & Wilson, M. (2005). Exploring local item dependence using a random-effects facet model. *Applied Psychological Measurement*, *29*(4), 296–318.

Watson, J., Kelly, B., & Izard, J. (2006). A longitudinal study of student understanding of chance and data. *Mathematics Education Research Journal*, *18*(2), 40–55.

Wells, C. S., Subkoviak, M. J., & Serlin, R. C. (2002). The effect of item parameter drift on examinee ability estimates. *Applied Psychological Measurement*, *26*(1), 77–87.

Wildeman, C. (2010). Paternal incarceration and children's physically aggressive behaviors: Evidence from the Fragile Families and Child Wellbeing Study. *Social*

Forces, *89*(1), 285–309.

Wilkinson, G. S., & Robertson, G. (2006). *Wide Range Achievement Test (WRAT4).* Lutz, FL: Psychological Assessment Resources.

Wood, R., Wilson, D. T., Gibbons, R. D., Schilling, S., Muraki, E., & Bock, R. D. (2002). *Testfact: Test scoring, item statistics, and item factor analysis.* Chicago, IL: Scientific Software International Inc.

Woodcock, R. W., Johnson, M. B., & Mather, N. (1990). *Woodcock-johnson psycho-educational battery–revised.* DLM Teaching Resources.

Wright, B. (1996). Time 1 to time 2 (pre-test to post-test) comparison and equating: Racking and stacking. *Rasch Measurement Transactions*, *10*(1), 478.

Wright, B. (2003). Rack and stack: Time 1 vs. Time 2. *Rasch Measurement Transactions*, *17*(1), 905–906.

Yang, C., Nay, S., & Hoyle, R. H. (2010). Three approaches to using lengthy ordinal scales in structural equation models: Parceling, latent scoring, and shortening scales. *Applied Psychological Measurement*, *34*(2), 122–142.

Yang, C., Olsen, J. A., Coyne, S., & Yu, J. (2017). Latent growth curve modeling of ordinal scales: A comparison of three strategies. *Journal of Biometrics & Biostatistics*, *8*(6).