Comparing Traditional Growth and Time-to-Criterion (T2C) Latent Variable Models

Micaela B. Moricet

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Education

University of Washington

2021

Committee:

Elizabeth Sanders, Chair

Oscar Olvera Astivia

Program Authorized to Offer Degree:

Measurement & Statistics

College of Education

University of Washington

Abstract

Comparing Traditional Growth and Time-to-Criterion (T2C) Latent Variable Models

Micaela B. Moricet

Chair of the Supervisory Committee:

Elizabeth Sanders

Department of Education

This paper compared traditional growth models with the time-to-criterion (T2C) latent variable models to examine if missingness and sample size would have differential effects on relative bias, model convergence rates, and power of growth parameters. Additionally, w the criterion location was also investigated for potential interactions with different levels of missingness and sample size. Results showed both models were comparable in terms of relative bias and power of mean growth estimates as well as predictor effects on growth estimates irrespective of conditions. However, for the T2C model, the time-to-criterion factor and its predictor effect exhibited positive bias in the small sample condition, and this was exacerbated as missingness levels increased as well as when the criterion location was farther out in time. Moreover, the T2C model approach was also associated with increased convergence issues when coupled with increased missing data and a smaller sample size. These results suggest the T2C approach should

only be used in circumstances where there is either ample sample size or modest (or no) levels of missingness. Limitations and future research directions are discussed.

Keywords: time-to-criterion model, latent variable growth modeling, missing data

Comparing Traditional Growth and Time-to-Criterion (T2C) Latent Variable Models

Many large-scale educational evaluations rely on measuring learners at a single point in time. For example, under the U.S. No Child Left Behind Act passed in 2002, a state's yearly educational effectiveness level (accountability) is measured by the proportion of students demonstrating proficiency on state-specific tests (No Child Left Behind [NCLB], 2002). Likewise, the National Assessment of Educational Progress (NAEP) measures U.S. students every other year in grades 4, 8, and 12 on a variety of subjects to provide insight into the nation as a whole, but also to compare states on student math and reading performance in particular (National Center for Educational Statistics, 1992). Similarly, every three years, the OECD's Programme for International Student Assessment (PISA) assesses 15-year-olds across over 90 countries on reading, mathematics, and science performance (Ray & Margaret, 2003). In all these assessments, understanding differences over time, not just differences within a particular year, requires that state or nation-level aggregate scores be computed (and weighted for representativeness) for use in any analyses aimed at evaluating change over time.

There are fundamentally two issues with only assessing a sample at one point in time when change processes over time are of interest. First, we cannot treat samples from one time to another, when different individuals form the samples, as exchangeable (i.e., we cannot match a sample from one year to a sample from another year at the individual level). In other words, even if a given sample is representative of a population at a given time point, the population itself changes over time due to physical movement within and across states and countries, as well as new generations being born into different social, political, educational, and environmental systems. Because of this first point, we aggregate samples to a particular macro-unit level (e.g., state or nation averages) for longitudinal analysis. However, in so doing, we lose rich

information about individual level variation and simultaneously introduce added measurement error to analysis estimates. In other words, there is no ability for analyses to disentangle within- and between-level differences in growth trajectories, and so any estimate of growth is a mix of both within- and between-level differences (e.g., Hamaker & Muthen, 2020). Last but not least, results comparing states or countries on growth trajectories may be fraught with construct-irrelevant confounds, such as differences testing conditions (allowing use of technology for example, or perhaps increased use of technology over time that inflates true growth) or outright cheating (e.g., countries trying to purposefully compete may provide their students with practice or may purposefully ensure that only very high performing students are sampled).

Of course, the alternative to single time point assessment is utilizing a multiple time point longitudinal study. Although such studies are more costly in terms of time and resources required to track individuals over time, such designs provide a far richer source of data for validly assessing change over time at both individual- and macro-levels, as well as predictors of change over time that could help inform policy (rather than just accountability). Further, such models are arguably fairer methods of holding states accountable compared to the single time point assessment, since individual and contextual effects on change over time could be disentangled analytically (McCoach et al., 2013). For example, if NAEP used a cohort-based design in which it sampled one cohort every few years, and each cohort involved the same students followed from grades 4 to 12, questions about contextual contributions, such as school district or state membership and related policies, could be directly evaluated.

Traditional Growth Models

In statistics, *growth modeling* is a class of methods used for analyzing change over time in longitudinal data, or data collected at multiple points in time. In recent years, the use of

growth models to assess students' change in assessment scores across time has increased among researchers and policymakers in education (e.g., Anthony & Ogg, 2020; Coley & Votruba-Drzal, 2020; Guglielmi, 2012; Guglielmi & Brekke, 2018; McCoach et al., 2013). This increase in use can be attributed in part to technological advances in data analysis with modern computing capabilities, and consequently, increased educational training in advanced statistical methods.

Traditionally, growth modeling has been conducted using multilevel or "mixed" effect models (this includes repeated measures analysis of variance models), or as multivariate (unilevel) structural equation models. In either case, "time" is typically structurally coded as a fixed effect corresponding to the actual time span in which assessments have occurred, with some reference point for the intercept (Biesanz et al., 2004). Once coded, time is specified as a predictor of the outcome; assuming a linear model, the estimated coefficient for "time" corresponds to the expected change in the outcome for each one unit increase in "time" (i.e., a growth rate). Predictors of change are specified as interactions with "time" on the outcome such that those estimates are either the difference among predictor categories (if a binary predictor) on growth rates, or the expected change in the growth rate for each one unit increase in the predictor (if a metrical predictor).

Structural Equation Modeling Approach

Compared to multilevel modeling (MLM) approaches to estimating change over time, structural equation modeling (SEM) approaches to growth estimation are more constrained by number of levels possible. Mathematically, the two approaches are identical under certain constraints (Curran, 2003). Indeed, both MLM and SEM estimate intra- and inter-individual change over time (e.g., separating growth within and between individuals), and both can incorporate predictors at either level of analysis (e.g., time-varying and time-invariant).

However, the SEM approach has two important advantages: it directly handles measurement error as part of the modeling process (MLM treats all variables as if they are perfectly measured), and further, the SEM approach can incorporate mediation modeling such as using growth rates (as a factor) to predict a separate distal outcome (Bollen & Curran, 2006). So, while both MLM and SEM approaches to growth modeling can answer the following questions:

1. At what skill level do individuals start out, on average? How much variability is there within and across students' baseline skills? What predicts how individuals start out?
2. What is the average growth rate across individuals? How much variability is there within and across students' growth rates? What predicts growth rate?
3. Do individuals who start at a higher skill level exhibit higher or lower growth rates?

Only an SEM approach can specifically answer the questions above *while taking into account measurement error*, as well as these questions:

4. Do individuals who start out at a higher skill level on one outcome predict having a higher skill on a different outcome?
5. Do individuals who have higher growth rates on one outcome predict having a higher skill on a different outcome?

Model Specification

A simple linear growth model (with no covariates) using multilevel or "mixed" model notation is given below.

$$y_{ij} = \alpha_j + \beta_j t_i + \varepsilon_{ij} \quad \begin{cases} \text{i} = \text{time} \\ \text{j} = \text{case} \end{cases} \tag{1}$$

Specifically, the expected outcome (y_{ij}) for individual *j* at time point *i* is the sum of: the mean response on outcomes at a reference time point (α_j) often set as *i* = 0 (the intercept factor); the

expected linear growth rate (β_j) for the outcome across time for individual *j*, (slope factor); and the error term (ε_{ij}) for individual *j* at time *i*.

In a multivariate SEM framework, the model shown in (1) can be reformulated as a unilevel model involving matrices, as follows.

$$\mathbf{y_j} = \boldsymbol{\upsilon} + \boldsymbol{\Lambda\eta_j} + \boldsymbol{\varepsilon_j} \quad (2)$$

In the model above, the vector of outcomes **y** for person *j* measured at *t* time points (*i* = (0,1,2,3…*t*-1)) are a function of a vector of intercepts $\boldsymbol{\upsilon}$ (*t* x 1 vector often fixed to **0** to identify the intercept and growth factor means[1]; a *t* x *P* matrix of factor loadings $\boldsymbol{\Lambda}$, where *P* = the number of growth factors (usually constrained to a design matrix layout within which the first column is fixed to **1** and the next column is fixed to **t** (if the first time point is serving as the reference for the intercept, **t** may be coded **t-1**); a *p* x 1 vector of weights $\boldsymbol{\eta}_j$ constraining each of the individual-specific latent factors (e.g., α_i and β_j with distribution $\boldsymbol{\eta}_j \sim N(\boldsymbol{\mu},\boldsymbol{\Sigma})$); and a *t* x 1 vector of time point-specific errors $\boldsymbol{\varepsilon}_j$ with distribution $\boldsymbol{\varepsilon}_j \sim N(\mathbf{0},\boldsymbol{\Theta})$. The *t* x *t* covariance matrix $\boldsymbol{\Theta}$ is typically fixed to assume local independence for person *j* across *t* time points, conditional on the latent factors $\boldsymbol{\eta}_j$ as follows:

$$\boldsymbol{\Theta} = \begin{bmatrix} \theta_0 & & \\ 0 & \theta_1 & \\ 0 & 0 & \theta_{t-1} \end{bmatrix}.$$

Perhaps a more intuitive understanding of latent models can be seen using path diagrams. Figure 1 is a path diagram adapted from Preacher et al. (2010) that illustrates a latent linear growth model for a scenario with five time points. As can be seen, both the intercept and the

[1] The intercept factor mean is the estimated mean value of Y for the time point coded 0, which is often (but not always) the first time point at which Y was measured; the growth factor mean, if assumed linear, is defined as the estimated mean change in Y for a one-unit increase in time. The difference between the SEM and MLM approaches is that the SEM approach estimates these means while taking the measurement error of Y at each time point into account.

slope are symbolized in ovals because they are latent, or unobserved, while the outcomes (y_i) are our observed data symbolized as boxes. The five observed outcomes are assumed to be caused by both the intercept and the slope, as indicated by the arrows or "paths" from each factor to each observed variable. The error term is unspecified as it is free to vary (i.e., be estimated separately) for each time point, which essentially provides us with structural model estimates that are separate from observed variable measurement error. Additionally, in our path diagram, we assume the covariance between the intercept and slope factors is zero (the path between them is fixed to zero). While this may be an unrealistic assumption, it can be a simple place to start the modeling process and can be relaxed in subsequent models (e.g., for testing the research question 3 described earlier).

Importantly, in all SEM, the scales of the latent variables depend on how we relate the observed variables to them, as well as our assumptions about the latent variables themselves (i.e., a latent variable need not be a normal distribution). In "regular" SEM (e.g., a confirmatory factor analysis), we might assume the latent variables are unit normal and only be interested in how the observed variables relate to the latent variables via their loadings (path coefficients), or the relationships (covariation) among the latent variables themselves. In SEM growth modeling, however, we are highly interested in estimating the latent means and variances. To accomplish this, we must put in place model constraints. Compared to regular latent variable models in which path coefficients are estimated, in latent growth models, the paths are (typically) constrained or "fixed" to reflect the structure of the data. In Figure 1, the paths from the intercept to each observed variable are constrained to 1, and the paths from the slope to each observed variable reflect the time point and distance of the measurements. This specification forces the estimation of the mean of the intercept to be the estimated mean at the first time point (because it

has no relationship with the slope per our specification), and the slope mean to be the estimated change in the outcomes per one year increase in time.

Lastly, we wish to make a few other points of clarification. For brevity throughout this paper, the term "linear" is used because it is assumed that the effect of time on the outcome is linear/additive, and not some other polynomial or piecewise form (such models would incorporate additional latent factor(s) to represent the functional growth form assumed). Also, again for brevity, for the remainder of this paper we limit our focus to observed variables assumed to be (conditionally) multivariate normally distributed, and equidistantly measured. This said, such assumptions can be relaxed in more complex model specifications.

Time to Criterion (T2C) Models

A limitation of the traditional latent growth model is that it estimates an intercept and slope (notably the intercept can be located as the final time point rather than initial time point, if desired) – it is does not focus at all on how long it might take individuals to reach a pre-defined level of an outcome. This is a fundamentally different question that may indeed reflect substantive policy interests (e.g., how long does it take children to become proficient?). Recently, Johnson & Hancock (2019) provided a means within which to model data to inform such a question by mathematically reparameterizing the traditional latent growth model as a time-to-criterion model (T2C) (for a technical discussion of the derivation see Johnson & Hancock, 2019). The mathematical representation of the reparameterized function is shown below in traditional multilevel or "mixed" notation.

$$y_{ij} = (c - \beta_j \tau_j) + \beta_j t_i + \varepsilon_{ij} \quad \begin{cases} i = \text{time} \\ j = \text{case} \end{cases} \tag{3}$$

We can see that the only difference between Equations 1 and 3 is that the intercept α_j has been replaced by the expression $(c - \beta_j \tau_j)$, where *c* is a predefined (known) criterion of the

outcome (y_{ij}). The mean linear growth factor β_j is the same in both models, but now τ_j is the focus of the model – it is the expected value of time *t* at which individual *j* reaches criterion *c*. In essence, both Equations 1 and 3 describe the same growth process (i.e., slope does not change), but Equation 3 estimates τ_j instead of α_j. (Note: an alternative formulation could also be to sacrifice the growth rate to estimate the intercept instead, if the intercept is of more interest; see Johnson & Hancock, 2019, p. 694 for further details.)

The multivariate characterization of Equation (2) still holds for the T2C model; however, the elements of the $\boldsymbol{\upsilon}$ and $\boldsymbol{\Lambda}$ matrices are altered to accommodate the estimation of τ rather than α. The resulting model equation is as follows (Johnson & Hancock, 2019).

$$\begin{bmatrix} y_{0j} \\ \dots \\ y_{(t-1)j} \end{bmatrix} = \begin{bmatrix} c + \mu_\beta \mu_\tau \\ \dots \\ c + \mu_\beta \mu_\tau \end{bmatrix} + \begin{bmatrix} 0 - \mu_\tau & -\mu_\beta \\ \dots & \dots \\ T_{(t-1)-\mu_\tau} & -\mu_\beta \end{bmatrix} \begin{bmatrix} \beta_j \\ \tau_j \end{bmatrix} + \begin{bmatrix} \varepsilon_{0j} \\ \dots \\ \varepsilon_{(t-1)j} \end{bmatrix} \quad (4)$$

Figure 2 depicts the model graphically. We see the intercept factor was replaced by the time-to-criterion factor. Moreover, we see that to make the model identifiable, constraints on the factor loadings and item intercepts had to be changed. For example, the factor loadings of the time to criterion factor had to be fixed at ($-\mu_\beta$) which is the mean growth rate factor mean. In addition, the factor loadings of the slope had to be fixed at $t_i - \mu_\tau$ which is the time at *i* (t_i) minus the mean time to reach the criterion (μ_τ). Lastly, the item intercepts had to be fixed at $c + \mu_{\beta*}\mu_\tau$ (not shown in figure).

Assumptions of this model are the same as that for the traditional multivariate growth model, including that both latent factors (i.e., slope and the time-to-criterion factor) and observed outcomes (y_{ij}) are metrical in nature (interval or ratio scale), and that the model is linear in nature; nonlinear modeling specifications (e.g., survival models) are viable but more difficult to parameterize and estimate (Johnson & Hancock, 2019).

Potential Model Estimation Issues

Missing Data

Within both the SEM and MLM framework, model estimation when missing data on the outcome (the endogenous variable) exists can be appropriately handled using full information maximum likelihood estimation (FIML). This approach is superior to listwise/casewise deletion in terms of statistical power/precision, and yields unbiased model estimates if the data can be assumed to be missing completely at random (MCAR) or just missing at random (MAR) (e.g., Enders & Bandalos, 2001; Kenward & Molenberghs, 1998). Specifically, MCAR refers to situations where missingness on a given outcome, *y*, is not systematically related to values of the *y* variable (e.g., individuals with higher skill deficits do not have more missing data on that skill), nor any other known variable, observed or unobserved (e.g., individuals classified as part of a particular subgroup, *x*, do not have more missing data on the focal skill, *y*). MAR, on the other hand, allows for the latter situation, so long as the external variable related to missingness is measured and included in the model. Otherwise, the missingness mechanism is considered not at random (MNAR); in such cases, parameter estimate biases can occur (Rubin, 1976). Although MAR may be a reasonable assumption for some models, including SEM growth models, any presence of missing data adds uncertainty to FIML model parameter estimates, yielding higher parameter standard errors than would occur if full data were available; in other words, power and precision for parameter effects can be reduced (Enders & Bandalos, 2001).

Both the type of missingness as well as missingness effects on power are crucial considerations for modeling growth processes since missingness may occur much more frequently in longitudinal data analyses as a result of people moving or discontinuing participation for a host of reasons. Moreover, it is natural for attrition to increase the longer the

study goes on (i.e., missingness will depend on time). Indeed, the only instance of MCAR that would be plausible for a longitudinal study are situations where participants may opt into and out of participation at any given time point. Although growth models do incorporate "time" as part of the modeling process (which therefore allows for assuming MAR, but not MCAR), it is unclear how missingness levels might differentially affect a T2C model, since the T2C depends on time points that are further out and therefore more prone to being missing due to attrition.

Sample Size

In structural equation modeling, sample size matters more than in traditional models for both statistical power as well as model convergence (i.e., arriving at a solution), but always in the context of complexity of the model being estimated (i.e., number of parameters and assumptions about the variables in the models). Rules of thumb recommended have ranged from greater than 100 up to 1000 (Comrey & Lee, 1992; Gorsuch, 1983; Guilford, 1954; Kline, 1979). However, for growth modeling, the number of individuals alone is not the essential consideration as "sample size"; rather it is the number of individuals as well as the number of time points per individual, the average correlation among any randomly selected pair of outcomes across time points (within individuals; in other words, the ICC), and choice of model complexity, that ensure sufficient precision for estimating model parameters (Hecht & Zitzmann, 2020). The higher the correlation is in outcomes across time points, the more subjects *or* time points (either – in other words, size of the full dataset) are needed.

Current Study

Although the traditional latent growth and T2C models are mathematically equivalent (yet answer at least one different research question), the present paper investigated whether missingness and sample size would have differential effects on relative bias, model convergence

rates, and power. I also wondered whether the criterion location might interact with missingness and sample size. Specifically, my research questions were:

1) (a)What are the effects of missingness, sample size, and level of criterion on relative bias and power for *mean growth* (β_j) for the T2C model (compared to the traditional growth model)? (b) Similarly, what is the bias in, and power for detecting, *a predictor effect* (group differences) on growth rates?
2) What are the effects of missingness, sample size, and level of criterion on relative bias and power for *mean T2C* (τ_j) for the T2C model? (b) Similarly, what is the bias in, and power for detecting, *a predictor effect* (group differences) on T2C?
3) Are there differences in convergence rates, error messages, and time for estimation between traditional and T2C models?

Method

To answer the research questions above, a small-scale Monte Carlo simulation was conducted in *Mplus*. Again, although the traditional and time-to-criterion models themselves, under the specification constraints described above, are mathematically equivalent, they may not yield equivalent power/precision when missing data is present, particularly when missing data is related to time (i.e., missingness at random, MAR). Moreover, the choice of the cutoff criterion for the T2C model, in combination with MAR, may also have impacts on model estimates and convergence problems.

Data Generation

Data were generated as a traditional SEM growth model (see Equation 2). In order to isolate the conditions of interest as well as produce a realistic situation for educational-related research, the following conditions were held constant:

1. There are five time points, equally spaced, to represent an annual measurement.
2. The measured variable, y_{ij}, is normally distributed.
3. The reliability of the measured variable is .80, which is a typical threshold of acceptability for U.S. norm-referenced achievement test measures.
4. The data *generation* model assumes a traditional growth model:
    a. The intercept has $M = 100$, $SD = 15$, much like a typical U.S. norm-referenced achievement measure.
    b. The true average growth rate, or slope, β_j is assumed linear and normally distributed with $M = 2$ (i.e., a 2-point increase per year) and $SD = 3$ (average deviation in growth rates). This induces an average observed correlation among time points of approximately .75, depending on the other conditions in the model; however, the correlation is not uniform: among nearer time points (e.g., y_1 and y_2) the relation is stronger than time points farther from each other (e.g., y_1 vs. y_5).
5. One binary predictor approximately equal in category proportions is assumed to have a 1-point effect on the intercept (i.e., a 1-point difference between groups, which translates to ±0.5 points difference between each category and the mean intercept) and a 2-point effect on the slope (i.e., a 2-point difference between groups, which translates to ±1 point difference between each category and the mean slope). This represents a tiny difference in groups at baseline, and a modest difference in groups on growth rates, approximately 1/3 of a standard deviation.

Sample size, missingness, and time-to-criterion values were manipulated. Specifically, the following conditions were varied.

1. ***Sample size***: two levels were used (n = 100 vs. n = 500). These sizes were chosen to reflect smaller-scale intervention-type data and larger-scale policy-type data.
2. ***Missingness***: three levels were used (0% missing, approximately 10% per year, and approximately 20% per year). In other words, 10% per year starts with approximately 10% missing by the second time point, then an additional 10% of the remainder, and so forth. We note that the term "approximately" is used here because generating missingness based on time points is not easily controlled in *Mplus'* Monte Carlo options, and there are several missingness patterns possible; however, the largest pattern generated is reflected in the prescribed missingness levels and in all conditions, there was no missing data for the first time point or the binary predictor. Note that the assumed missing data mechanism is MAR because missingness is only related to time, and time is incorporated in the model.
3. ***Criterion***: when we applied the T2C model to the data, two levels of the time to the criterion (τ_j) factor were used (3 years and 5 years, on average). We used these because one represents a value that might be less affected by missingness since it is achieved earlier in time, and one represents a value that might be more affected. Mathematically, since we know the growth rate is 2 points per year, and the intercept is 100 points, this translates to a criterion cutoff of 106 for $\tau_j = 3$ and 110 for $\tau_j = 5$.

   Of note, the location of the criterion impacts the expected predictor effect on τ_j. The data generation procedure included a group effect (the binary predictor) of 1 point difference on the intercept and 1 point difference on the growth rate. Recall that the true mean trajectory was generated as 100 + 2*time. Importantly, Mplus generation procedures in Monte Carlo simulations assume that X is an effect coded,

rather than dummy coded variable. As such, the true mean trajectory of the group coded -1 is 99.5 + 1.5*time (because the group is assumed 0.5 points lower on the intercept and 0.5 points lower on the growth rate *compared to average*). And the true mean trajectory of the group coded +1 is 100.5 + 2.5*time (because the group is assumed 0.5 points higher on the intercept and 0.5 points higher on the slope). One can then solve for the expected value of each group's time-to-criterion by setting the expressions equal to c. Thereafter, the distance between the two groups can be computed. This is the expected effect of the predictor on the time-to-criterion factor. For the $\tau_j = 3$ condition, the expected difference between groups is approximately 2.13. For the $\tau_j = 5$ condition, the expected difference between groups is approximately 3.20. In the model, we would find both of these to be negative values, since the group coded +1 grows at a faster rate and therefore has a lower time to criterion than the group coded -1.

Across the simulation design elements, we have a 2 (sample size) x 3 (missingness level) x 3 (level of τ_j) = 12-condition design. We applied the traditional growth model to each of the first two design elements, and the T2C model to all three design elements. We used the same seed to randomly generate the data to ensure results could be replicated and were as comparable across conditions as possible. For each condition and model approach, 1,000 replications were conducted.

Data Analysis

Because the present study is a small-scale Monte Carlo simulation, descriptive statistics and related data visualizations were used to evaluate the research questions. Descriptive statistics across the two modeling approaches focused relative bias and power for the growth rate and the

growth rate predictor effect, as well as model convergence time and number of error messages. Within the two T2C τ_j conditions, relative bias and power were examined for the time-to-criterion factor and the predictor effect on it, as well as convergence and number of error messages. Relative bias is defined as the difference between the mean estimate across 1,000 replications and the true value, divided by the true value. Power is defined as the proportion of parameter estimate effects in which the null was rejected (all our conditions are non-null). Lastly, convergence issues were assessed across each of the 1,000 replication sets, for each condition and modeling approach.

Results

Traditional vs. T2C

Results showed that model fit indices across the traditional and time-to-criterion (T2C) model approaches were identical for each of the conditions, including the Chi-Square test of model fit, AIC, BIC, Adjusted BIC, RMSEA, and SRMR. Although this was expected given that the models themselves are equivalent, it also served as a check that the models were specified correctly.

Table 1 displays mean relative bias for the growth factor mean and the predictor effect on growth, across the two approaches and sample sizes. As can be seen, there was little relative bias for any of the models, irrespective of sample size or missingness conditions (less than 5% is typically considered acceptable for coefficient bias; Hoogland & Boomsma, 1998).

Table 2 displays the empirical power for the growth factor mean and predictor effect on growth. Although power decreased for higher amounts of missingness (as expected), for the

mean growth rate specifically, this was only observed for the smaller sized sample condition, and it did not vary across modeling approach.

This said, convergence issues were apparent for the T2C models for conditions with missingness under the sample size condition (n = 100) (see Table 3). Although the pattern was not entirely consistent, there were more convergence issues with the T2C models compared with the traditional model for data with any missingness, and the number of error messages did increase consistently as for the higher τ_j level and increased missingness level.

T2C: Earlier vs. Later Criteria

Tables 4 and 5 display the mean relative bias and power for the time-to-criterion factor mean τ_j and its predictor. As can be readily seen in Table 4, for the small sample size condition (n = 100) there was positive bias for estimating the factor mean (on average, 13%), and a very pronounced positive bias for estimating the predictor effect on the factor (on average, 111%). Importantly, the overestimation was present even with no missingness, but increased especially for the higher missingness condition (see Figure 3 A & B). This said, no overestimation was observed for the larger sample condition (n = 500): average bias was 2-3% for the factor mean and predictor effect, respectively.

Although there were substantial biases in the parameter coefficient estimates, it is clear that there was also a large degree of variability in the estimates given the lower than optimal power levels for detecting the factor mean, and the nearly non-existent power levels for the predictor effect on the time-to-criterion factor (see Table 5). This said, the larger sample size condition (n = 500) exhibited better power (in addition to relatively little bias).

Discussion

Overall, the current study found that both the traditional growth and the time-to-criterion (T2C) latent variable model approaches were highly comparable. Both approaches' growth parameter-related results were similarly sensitive to sample size, and to a lesser extent, missingness levels. We found little relative bias for the growth rate factor or the predictor of growth rate for either approach (under any condition), and no difference in fit or power across approaches for any given condition. As expected, power for detecting parameters decreased with higher amounts of missingness.

There are two important caveats in the results, however. First, results for the T2C model approach, when coupled with missing data and a smaller sample size, were associated with greater failures to converge and error messages, and considerably larger convergence times. Second, the focal outcome for T2C models – the tau factor and its predictor – exhibited positive bias in the small sample condition, which increased as missingness levels increased as well as when the cutoff criterion was farther out in time. In other words, as the cutoff was more affected by missingness, there was more over-estimation of the parameter value. Although we might be worried about such bias, the power estimates revealed that these conditions also produced increased variability to the extent that there was little power for detecting even the biased effects. Again, however, this finding was not observed for the larger sample size.

Although we already know that the T2C model is a reparameterization of the traditional latent growth model (i.e., they are mathematically equivalent), our findings found that the T2C approach is relatively more sensitive to sample size and missingness than the traditional latent growth model. Moreover, when different start values were used, or were dropped, results could vary considerable in terms of convergence rates, time, and error messages. Of course, the closer

the start values were the same as the "true" parameter, the faster the convergence times and the fewer the convergence issues. In the real world, however, we do not know the true growth parameters, so convergence issues are most likely underestimated in this study. Further research should be done to figure out how to find appropriate starting values when working with real data.

The limitations in our results are directly related to the conditions set for this study. First, this paper only looked at linear growth, and results may not be generalizable to nonlinear models. Second, we used one binary predictor and a relatively small effect size for the predictor on the growth and T2C factors. Because the predictor effect on the growth and T2C factors was not zero for all conditions, we lack a condition evaluating Type I error. Third, the covariance between the intercept and slope was constrained to 0. In the real world, this is a rare occasion as typically where you start determines your growth rate. Fourth, the levels of missingness used were modest, as was the growth rate itself. Some longitudinal studies, especially data for vulnerable populations with high mobility (e.g., emergent bilingual children), can have upwards 80% of the data missing. Relatedly, we used FIML to handle missingness throughout this study; nevertheless, we have no reason to believe using multiple imputation would alter results.

Finally, for the T2C model, the two time-to-criteria levels tested (3 and 5 years) were within the five time points' observed range. Thus, we cannot say how results might differ if the criterion were set farther out, after the last time point. Relatedly, we used five time points; the minimum is three time points, and we would guess that fewer time points would yield even more instability for smaller sized samples. In sum, future work should investigate the effects of the following conditions on parameter power and model convergence: (a) nonlinearity, (b) null effect of the predictor on the growth parameters, (c) effects of differing levels of intercept-slope

covariance, (d) higher levels of missingness, and (e) time-to-criterion levels extending beyond the observable time points.

Under the right conditions, both model approaches can be useful in understanding individual growth trajectories over time, and with the T2C model, understanding differences in achieving a specific criterion or proficiency level in achievement. The current study results shed light on the instability of the T2C model in circumstances with relatively small sample sizes, coupled with systematic attrition over time. As such, we caution researchers to use the T2C in circumstances where there is either ample sample size or modest (or no) levels of missingness.

Whether one can use the T2C model to ask questions relating to individuals' time to reach a proficiency level, or differences in time to reach a proficiency level between groups, is highly dependent on the quality of the data that is available to the researcher. As the use and availability of these sophisticated models continues to grow in education and the social sciences, we are hopeful that intentional data collection methods will follow so that researchers can accurately test the questions they have embarked upon.

References

Anthony, C. J., & Ogg, J. (2020). Executive function, learning-related behaviors, and science growth from kindergarten to fourth grade. *Journal of Educational Psychology*, *112*(8), 1563–1581. https://doi.org/10.1037/edu0000447

Biesanz, J. C., Deeb-Sossa, N., Papadakis, A. A., Bollen, K. A., & Curran, P. J. (2004). The role of coding time in estimating and interpreting growth curve models. *Psychological methods*, *9*(1), 30–52. https://doi.org/10.1037/1082-989X.9.1.30

Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation perspective*. Hoboken, NJ: Wiley

Coley, R. L., Kruzik, C., & Votruba-Drzal, E. (2020). Do family investments explain growing socioeconomic disparities in children's reading, math, and science achievement during school versus summer months?. *Journal of Educational Psychology*, *112*(6), 1183–1196. https://doi.org/10.1037/edu0000427

Comrey, A. L., and Lee, H. B. (1992). *A First Course in Factor Analysis*. Hillsdale, NJ: Erlbaum.

Curran, P. J. (2003). Have multilevel models been structural equation models all along?. *Multivariate Behavioral Research*, *38*(4), 529–569. https://doi.org/10.1207/s15327906mbr3804_5

Enders, C. K., & Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural equation modeling*, *8*(3), 430–457. https://doi.org/10.1207/S15328007SEM0803_5

Gorsuch, R. L. (1983). *Factor Analysis*, 2nd Edn. Hillsdale, NJ: Erlbaum.

Guglielmi, R. S. (2012). Math and science achievement in English Language Learners: Multivariate latent growth modeling of predictors, mediators, and moderators. *Journal of Educational Psychology*, *104,* 580–602. https://doi.org/10.1037/a0027378

Guglielmi, R. S., & Brekke, N. (2018). A latent growth moderated mediation model of math achievement and postsecondary attainment: Focusing on context-invariant predictors. *Journal of Educational Psychology*, *110*(5), 683–708. https://doi.org/10.1037/edu0000238

Guilford, J. P. (1954). *Psychometric Methods* (2nd Edn.). New York, NY: McGraw-Hill.

Hamaker, E. L., & Muthén, B. (2020). The fixed versus random effects debate and how it relates to centering in multilevel modeling. *Psychological methods*, *25*(3), 365–379. https://doi.org/10.1037/met0000239

Hecht, M., & Zitzmann, S. (2020). Sample size recommendations for continuous-time models: Compensating shorter time series with larger numbers of persons and vice versa. *Structural Equation Modeling: A Multidisciplinary Journal*, 1–8. https://doi.org/10.1080/10705511.2020.1779069

Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling: An overview and a meta-analysis. *Sociological Methods & Research*, *26*(3), 329-367. https://doi.org/10.1177/0049124198026003003

Johnson, T. L., & Hancock, G. R. (2019). Time to criterion latent growth models. *Psychological methods*, *24*(6), 690–707. https://doi.org/10.1037/met0000214

Kenward, M. G., & Molenberghs, G. (1998). Likelihood based frequentist inference when data are missing at random. *Statistical Science*, *13*(3), 236–247. https://doi.org/10.1214/ss/1028905886

Kline, P. (1979). *Psychometrics and Psychology*. London: Academic Press.

McCoach, D. B., Rambo, K. E., & Welsh, M. (2013). Assessing the growth of gifted students. *Gifted Child Quarterly*, *57*(1), 56–67. https://doi.org/10.1177/0016986212463873

National Center for Education Statistics., National Assessment of Educational Progress (Project), Educational Testing Service., & United States. (1992). *NAEP ... reading report card for the nation and the states*. Washington, D.C: National Center for Education Statistics, Office of Educational Research and Improvement, U.S. Dept. of Education.

No Child Left Behind Act of 2001, P.L. 107-110, 20 U.S.C. § 6319 (2002).

Preacher, K. J., Zyphur, M. J., & Zhang, Z. (2010). A general multilevel SEM framework for assessing multilevel mediation. *Psychological methods*, *15*(3), 209–233. https://doi.org/10.1037/a0020141

Ray, A., & Margaret, W. (Eds.). (2003). *PISA Programme for international student assessment (PISA) PISA 2000 technical report: PISA 2000 technical report*. OECD Publishing.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592. https://doi.org/10.1093/biomet/63.3.581

Table 1

Comparing Models on Relative Bias for Growth Rate-Related Estimates

Parameter	Sample size n = 100			Sample size n = 500		
	Trad	T2C, $\tau_j = 3$	T2C, $\tau_j = 5$	Trad	T2C, $\tau_j = 3$	T2C, $\tau_j = 5$
Mean Growth Rate ($\beta_j = 2$)						
Missing Rate = 0%	-.01	-.01	-.01	.00	.00	.00
Missing Rate = 10%	-.01	-.01	.00	.00	.00	.00
Missing Rate = 20%	.00	.00	.00	.00	.00	.00
Predictor Effect on Growth Rate (Coeff = 1)						
Missing Rate = 0%	.03	.03	.03	-.01	-.01	-.01
Missing Rate = 10%	.01	.01	.00	-.02	-.02	-.02
Missing Rate = 20%	-.01	-.02	-.02	-.01	-.01	-.01

Note. N = 1,000 replications per cell. Relative Bias = (model-estimated mean - true parameter)/true parameter.

Table 2

Comparing Models on Power for Growth Rate-Related Estimates

Parameter	Sample size n = 100			Sample size n = 500		
	Trad	T2C, $\tau_j = 3$	T2C, $\tau_j = 5$	Trad	T2C, $\tau_j = 3$	T2C, $\tau_j = 5$
Mean Growth Rate ($\beta_j = 2$)						
Missing Rate = 0%	.93	.93	.93	1.00	1.00	1.00
Missing Rate = 10%	.87	.88	.88	1.00	1.00	1.00
Missing Rate = 20%	.78	.79	.78	1.00	1.00	1.00
Predictor Effect on Growth Rate (Coeff = 1)						
Missing Rate = 0%	.23	.23	.23	.77	.77	.77
Missing Rate = 10%	.21	.21	.20	.70	.70	.70
Missing Rate = 20%	.18	.17	.18	.59	.59	.59

Note. N = 1,000 replications per cell. Power = % of parameter estimates found significant for α = .05, 2-tailed.

Table 3

Comparing Models on Convergence Issues

Issue	Sample size n = 100			Sample size n = 500		
	Trad	T2C, $\tau_j = 3$	T2C, $\tau_j = 5$	Trad	T2C, $\tau_j = 3$	T2C, $\tau_j = 5$
Total Time across Replications (Seconds)						
Missing Rate = 0%	4	18	21	6	19	25
Missing Rate = 10%	10	45	705	13	29	32
Missing Rate = 20%	12	412	48	17	40	38
Non-Convergences						
Missing Rate = 0%	0	0	0	0	0	0
Missing Rate = 10%	0	2	141	0	0	0
Missing Rate = 20%	0	76	3	0	0	0
Error Messages for Converged Solutions						
Missing Rate = 0%	0	0	0	0	0	0
Missing Rate = 10%	0	6	6	0	0	0
Missing Rate = 20%	0	45	49	0	0	0

Note. N = 1,000 replications per cell. All values are totals across 1,000 replications.

Table 4

Comparing Different T2C Criteria on Relative Bias for Tau-Related Estimates

Parameter	Sample size n = 100		Sample size n = 500	
	$\tau_j = 3$	$\tau_j = 5$	$\tau_j = 3$	$\tau_j = 5$
Mean Time-to-Criterion (τ_j)				
Missing Rate = 0%	.10	.11	.02	.02
Missing Rate = 10%	.12	.11	.02	.02
Missing Rate = 20%	.16	.18	.02	.02
Predictor Effect on Time-to-Criterion (Coeff)*				
Missing Rate = 0%	.67	.77	.01	.02
Missing Rate = 10%	.99	1.01	.02	.04
Missing Rate = 20%	1.52	1.72	.04	.07

Note. N = 1,000 replications per cell. Relative Bias = (model-estimated mean - true parameter)/true parameter. Values in boldface represent substantial bias.

* True value depends on group differences on slope (β_j) and sacrificed intercept, as well as the mean τ_j; for $\tau_j = 3$, coeff ≈ 2.13.

Table 5

Comparing Different T2C Criteria on Power for Tau-Related Estimates

Parameter	Sample size n = 100		Sample size n = 500	
	$\tau_j = 3$	$\tau_j = 5$	$\tau_j = 3$	$\tau_j = 5$
Mean Time-to-Criterion (τ_j)				
Missing Rate = 0%	.63	.91	1.00	1.00
Missing Rate = 10%	.60	.86	1.00	1.00
Missing Rate = 20%	.53	.78	1.00	1.00
Predictor Effect on Time-to-Criterion (Coeff)*				
Missing Rate = 0%	.01	.00	.32	.40
Missing Rate = 10%	.00	.00	.24	.25
Missing Rate = 20%	.00	.00	.14	.10

Note. N = 1,000 replications per cell. Power = % of parameter estimates found significant for α = .05, 2-tailed.

* True value depends on group differences on slope (β_j) and sacrificed intercept, as well as the mean τ_j; for $\tau_j = 3$, coeff ≈ 2.13.

Figure 1

A linear latent growth model with y_{ij} measured at I time points, a random slope factor (β_j) and a random intercept factor (α_j). Slope factor loadings are fixed at 1-year intervals, and intercept factor loadings are fixed at 1.

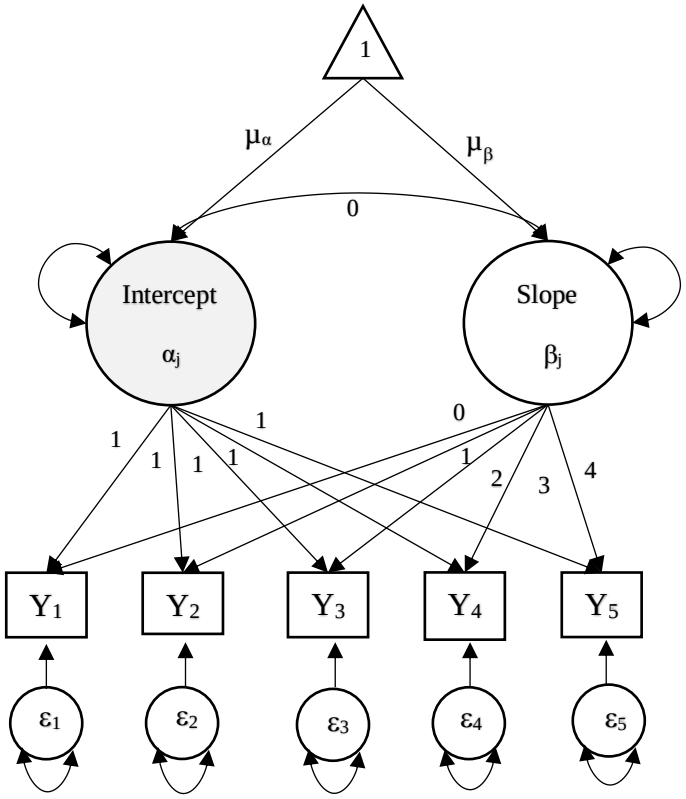


Figure 2

A reparameterized linear latent growth model with y_{ij} measured at I time points, a random slope factor (β_j) and a random Time to Criterion factor (τ_j). Slope factor loadings are fixed at 1-year intervals, and intercept factor loadings are fixed at 1.

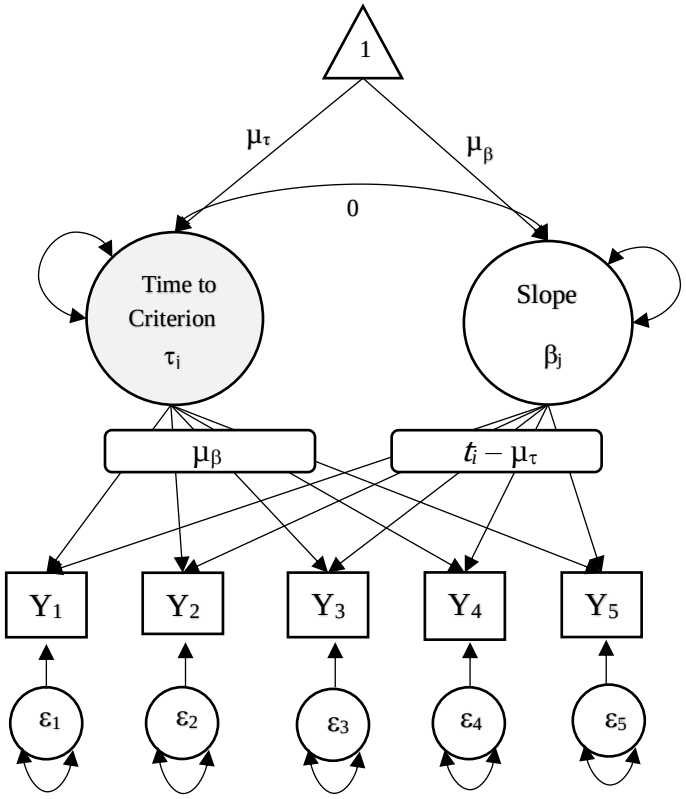
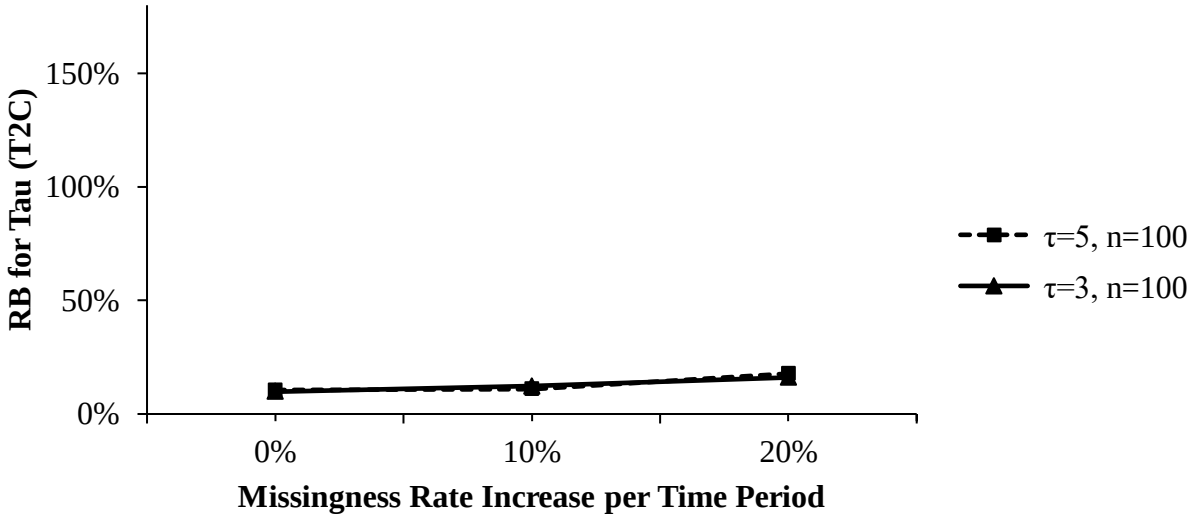


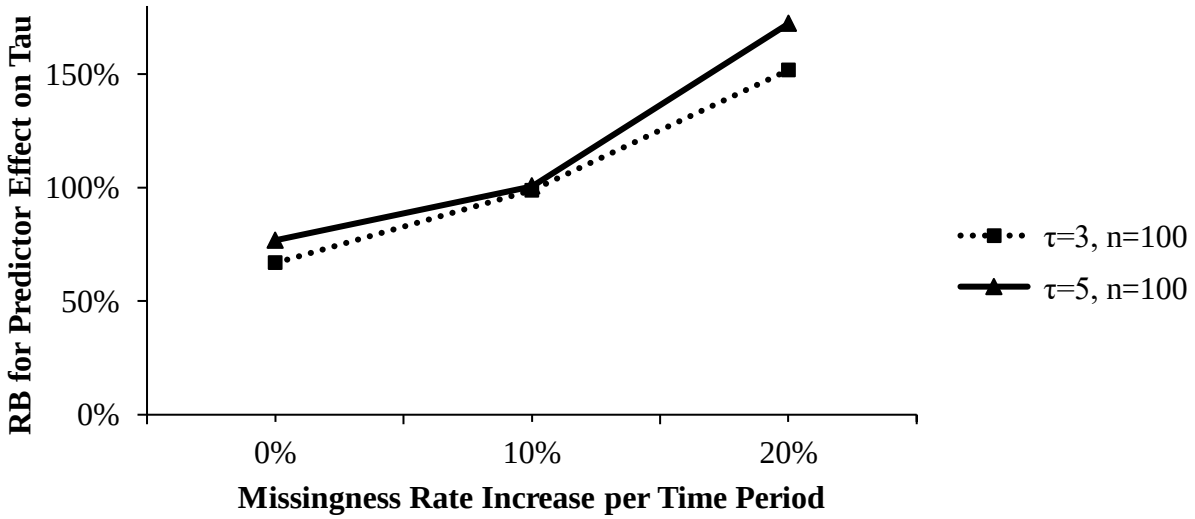
Figure 3

Relative Bias of Time-to-Criterion Factor Means (A) and Predictor Effects (B) for n = 100

A



B



Appendix

Sample *Mplus* MC Simulation Code for T2C Analysis: τ = 3, *n* = 100, 10% missing rate

```
TITLE: MC T2C Model, 5 time points, missing=10, n=100, t=3

MONTECARLO:
NAMES = y1-y5 x;! initial variables to generate, unit normal
CUTPOINTS = x(0); ! binary variable 50/50 split at z=0
NOBSERVATIONS = 100;! sample size, hold constant
NREPS = 1000; ! number of replications (1000 standard)
SEED = 293117;! arbitrary random seed
MISSING = y2-y5;! not missing completely at random due to attrition

MODEL POPULATION:
i b | y1@0 y2@1 y3@2 y4@3 y5@4; ! assuming equidistant time points, let's say years
[i*100];! true mean of latent intercept (value of Y at time0)
[b*2];! true mean of latent linear slope (change in Y/+1year)
i*225;! true variance of latent intercept factor
b*9;! true variance of latent slope factor
i with b*0; ! true factor covariance unrealistic but for simplicity
y1-y5*81; ! true resid err of obs items = Y var*(1 - reliab^2)
[x*0];! true mean of predictor = avg of effect-coded variable
x*1;! true var of predictor = var of effect-coded var=1
!will generate dummy variable for analysis, however
i ON x*1; ! true effect of predictor on i (diff btwn grps at t1)
b ON x*1; ! true effect of predictor on b (diff btwn grps on rate)

MODEL MISSING:! 10% missing increase over time
[y2@-2.20];! mean 10% missing (in logits)
[y3@-1.45];! mean 19% missing (in logits)
[y4@-0.99];! mean 27% missing (in logits)
[y5@-0.65];! mean 34% missing (in logits)

ANALYSIS:
ESTIMATOR = ML;
COVERAGE = .05;
ITERATIONS = 10000;

MODEL:
! factor-item relationships
tj BY y1-y5* (t); ! tau = time to criterion factor with label for constraint
bj BY y1-y5* (b1-b5); ! beta = linear change per year factor with labels for
constraints
! factor means
[tj*3] (mt);! mean time to reach criterion factor mean, estimated, based on c
! start value avg t2c for control grp (x=0)
[bj*2] (mb);! mean growth rate factor mean, estimated
! start value is average for control grp
! factor variances and covariances
tj*81;!! res var T2C factor freely estimated, start val = resvar(Y)
bj*9; !! res var of slope factor freely estimated, use start val var(b)
tj with bj*-13.5; !! covar of factors, freely estimated, approx (-mt/mb)*var(b)
! obs item means and variances
[y1-y5*108] (n);! obs item mean at tau, label for constraint, start value
! start value = i (or i + 1/2*effect of x on i, if x effects i)
y1-y5*81; ! obs item resid error variance, no constraints, start value
!!!! predictor effects <---- START VALUES AFFECTED BY THE CHOICE OF C (BELOW)
bj ON x*1;! predictor effect on b, start value (same as bj in pop model)
tj ON x*-1; ! predictor effect on t, start value (note: the actual effect is -2.13)

MODEL CONSTRAINT:
```

```
NEW(c ma);
!!!! the criterion <---- THIS IS WHAT WE MANIPULATE
c = 106;! used i + b*time, where time=3 = 100+2*3 = 106
ma = c - mb*mt; ! mean intercept (not estimated, mathematically derived)
! intercept constraints
n = c + mb*mt;! same for all item-factor relationships
! tau factor loading constraints: t2c
t = -mb;! same for all item-factor relationships
! beta factor loading constraints function of mean t2c
b1 = 0-mt;! time @y1 -mt
b2 = 1-mt;! time @y2 -mt
b3 = 2-mt;! time @y3 -mt
b4 = 3-mt;! time @y4 -mt
b5 = 4-mt;! time @y5 -mt
```