Using Bayesian IRT for Multi-Cohort Repeated Measure Design to Estimate Individual Latent Change Scores

Chun Wang
Ruoyi Zhu
College of Education, University of Washington, Seattle, WA, USA

Paul K. Crane
Seo-Eun Choi
[1]Department of Medicine, University of Washington, Seattle, WA, USA
Richard N. Jones
Douglas Tommet
[13]Department of Psychiatry, Brown University, Providence, RI, USA

Correspondence concerning this manuscript should be addressed to Chun Wang at:
312E Miller Hall
Measurement and Statistics
College of Education, University of Washington
2012 Skagit Ln, Seattle, WA 98105
e-mail: wang4066@uw.edu
phone: 217-722-7037

Acknowledgements
Analyses were funded by R01 AG029672ADNI Psychometrics, P Crane, PI) and a supplement to that grant.

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.;Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.;Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Author's Note. This full study was presented at an invited talk at the 2023 Conference of Advanced Psychometric Methods of Cognitive Aging Research (ψMCA), Granlibakken, CA

Using Bayesian IRT for Multi-Cohort Repeated Measure Design to Estimate Individual Latent Change Scores

Abstract

Repeated measure data design has been used extensively in a wide range of fields, such as brain aging or developmental psychology, to answer important research questions exploring relationships between trajectory of change and external variables. In many cases, such data may be collected from multiple study cohorts and harmonized, with the intention of gaining higher statistical power and enhanced external validity. When psychological constructs are measured using survey scales, a fundamental psychometric challenge for data harmonization is to create commensurate measures for the constructs of interest across studies. Traditional analysis may fit a unidimensional item response theory (IRT) model to data from one time point and one cohort to obtain item parameters and fix the same parameters in subsequent analyses. Such a simplified approach ignores item residual dependencies in the repeated measure design on one hand, and on the other hand, it does not exploit accumulated information from different cohorts. Instead, two alternative approaches should serve such data designs much better: an integrative approach using multiple-group two-tier model via concurrent calibration, and if such calibration fails to converge, a Bayesian sequential calibration approach that uses informative priors on common items to establish the scale. Both approaches use a Markov chain Monte Carlo (MCMC) algorithm that handles computational complexity well. Through a simulation study and an empirical study using Alzheimer's Diseases Neuroimage Initiative (ADNI) cognitive battery data (i.e., language and executive functioning), we conclude that latent change scores obtained from these two alternative approaches are more precisely recovered.

Key words: Bayesian item response theory, two-tier model, latent change score, bi-factor model

The National Institute of Health actively endorses the sharing of data sets between research teams. For instance, National Institute on Aging (NIA) lately funded a Harmonized Cognitive Assessment Protocol (HCAP) to measure and understand dementia risk within ongoing longitudinal studies of aging around the world to harmonize data, methods, and content to facilitate cross-national comparisons.[1] NIA also put out a new grant competition on harmonization of Alzheimer's disease and related dementias (ADRD) genetic, epidemiologic, and clinical data to optimize the ability to identify well-targeted therapeutic approaches for ADRD. The position and efforts of prominent national agencies signal the emerging trend of leveraging scarce resources to assemble cross-study data sets to efficiently address overarching research questions. Required to accomplish these goals are methods to support principled analysis of pooled datasets, one of which is to reconcile between-study differences in the measurement of key constructs, such as cognitive ability (Vonk et al., 2022), HIV stigma (Kemp et al., 2019), alcohol use (Huh et al., 2015; Witkiewitz et al., 2016), marijuana use (Silins et al., 2015), among others. Only when putting the factor scores of those constructs from different validated instruments or the same instruments from different populations on the commensurate scale can the subsequent statistical analysis of intervention effects on the pooled data be valid. In addition, a desirable potential advantage is the broader psychometric assessment of theoretical constructs resulting from the use of different item sets across study.

Integrative data analysis (IDA) is a novel framework for conducting the simultaneous analysis of raw data pooled from different studies. It offers many advantages, including increased power due to larger sample size, enhanced external validity and generalizability due to greater heterogeneity in demographic and psychosocial characteristics, cost effectiveness due to

[1] https://hrs.isr.umich.edu/data-products/hcap

reuse of extant data, and potential to address new research questions not feasible by a single study, etc. (Curran & Hussong, 2009; Curran et al., 2010). However, significant methodological challenges must be addressed when pooling data from independent studies, and one such challenge is to establish commensurate measures for the constructs of interest (Nance et al., 2017). When data from different yet overlapping instruments and diverse samples are pooled, psychometric analysis needs to be general and flexible enough to accommodate idiosyncrasies in the population and instrument designs.

As psychological constructs of interest cannot be directly observed, they are often measured by a set of survey items, making item response theory (IRT) an appropriate candidate for analysis. Traditionally, data from a single representative sample in one study is first analyzed, such as using a specific IRT model to calibrate item parameters. Then such item parameters are carried forward in subsequent analyses as if they are free of measurement and estimation errors (Choi et al., 2020; Crane et al., 2012, 2021; Schober & Vetter, 2018). With the advancement of multilevel and multiple-group IRT models (Cai et al., 2011; Wang & Nydick, 2020), a better approach to such data would be to take an integrated approach that simultaneously handle data from different studies together.

Aside from the multi-cohort design that naturally arises when pooling data from different studies (e.g., Davoudzadeh, et al., 2020), this paper also considers a compounded repeated measure design that intends to measure individual change. When the research focus is on tracking individual's change on the latent construct over time, such as in the field of developmental, clinical, educational, and applied psychology (e.g., Grimm et al., 2013, McArdle, 1988; McArdle et al., 2009), either second-order latent growth curve (LGC) models (e.g., Bauer & Curran, 2016; Caprara et al., 2011; Hancock et al., 2001; McArdle, 1988; Meredith & Tisak,

1990; Soland et al., 2019; Soland & Kuhfeld, 2019) or longitudinal IRT models (Cai & Houts, 2021; Wang & Nydick, 2020; Paek et al., 2014) are the *de facto* tools. Mathematically, longitudinal IRT models can be reparametrized as second-order or higher-order latent growth curve models (e.g., Edwards & Wirth, 2009; Wang et al., 2016), which implies that they are inherently equivalent, and both can be fitted with general-purpose software packages such as M*plus* (Muthén & Muthén, 2023), OpenMx (Boker et al., 2023), and Lavvan (Rosseel, 2012). However, when primary measurements differ from one occasion to the next, due to age appropriateness (McArdle, et al., 2009), new and improved test batteries (e.g., Edwards & Wirth, 2009), or test security, longitudinal IRT modeling framework is preferred. That is because the IRT approach can naturally separate "differences in the scales over time from changes in the constructs over time" (McArdle, et al., 2009, p.129), via the IRT linkage of common items (Edwards & Wirth, 2009; Wang, et al., 2016).

Although multi-cohort repeated measure design is common when one pools longitudinal data from different studies, most of the current published analyses still use a simplified approach. That is, a unidimensional IRT model is used to estimate item parameters from a single representative sample. Such item parameters are then used in other studies via a fixed-parameter calibration (Choi et al., 2020; Crane et al., 2012; Kim & Kolen, 2016; Wang et al., 2019). The second-order LGC model or longitudinal IRT models are still much less adopted in real data analysis (Kuhfeld & Soland, 2022; Isiordia & Ferrer, 2018), partly because simultaneously modeling measurement and change is hindered by computational and practical concerns (Bauer & Curran, 2016), such as the large samples needed to obtain stable parameter estimates. However, little evidence exists about the precision loss, if any, from such a simplified yet probably more feasible approach compared to a more sophisticated approach in terms of

individual change scores. Such change scores can be used in subsequent analyses, for instance, to establish links between treatment and effect, or to cluster individuals based on different change patterns and then use clusters to explain outcome variables (i.e., varying levels of cognitive change indicate different risks of conversion from mild cognitive impairment to Alzheimer's disease, Choi et al., 2020). Hence, quantifying and minimizing measurement errors of individual change score is essential to increase power of subsequent analysis. Indeed, there are so many ways scores can be extracted from a single data source, and these different choices (i.e., using a simplified approach such as a unidimensional IRT model vs. using a more sophisticated approach such as a longitudinal IRT model) may contribute to the now well-documented replication issues in psychological studies (Fried & Flake, 2018).

Our study involves simulation and empirical studies to investigate how to leverage Bayesian Markov chain Monte Carlo (MCMC) to optimally extract individual change scores from multi-cohort repeated measure design in which instruments may differ across time and across studies. We choose Bayesian MCMC over the popular full-information maximum likelihood (FIML) here because FIML is computationally prohibitive when the number of latent factors in the model is high (Fox, 2010). The high dimensionality issue occurs when either the number of time points and/or the number of repeated-administered items (hence the number of nuisance factors) are large. Instead, Bayesian estimation serves as a viable alternative for IDA not only because it deals with missing data as well as FIML while obviating FIML's high-dimensional challenge, but also due to its nature of incorporating varying degrees of prior information. When estimation of a complex model proceeds in stages, it can incorporate uncertainties in separate stages of the estimation whereas other estimation methods would have to ignore the uncertainties across stages undesirably. McArdle et al. (2009) first used MCMC on

their proposed longitudinal invariant Rasch test (LIRT) model. Our paper differs significantly from theirs in that (1) we use the graded response model that allows for discrimination parameters to differ across items and that can handle Likert scale items; (2) we consider a multi-group scenario to account for between-study population differences (e.g., Davoudzadeh, et al., 2020); and (3) we include nuisance factors in the model to account for time dependency of the same items administered repeatedly. To model growth trajectories, we choose not to impose any second-order structural models because they are studied elsewhere (e.g., Wang & Nydick, 2019; Wang et al., 2016), and we restrict our discussion to two and three time point models, which are commonly seen in quick-paced clinical trials.

In addition to computational benefits of the Bayesian approach, Bayesian latent variable modeling, in general, has been demonstrated to be a more flexible representation of substantive theory because it allows to "replace parameter specifications of exact zeros with approximate zeros based on informative, small-variance priors." (Muthen & Asparouhov, 2012). Freeing these parameters in conventional maximum likelihood estimation would render the model non-identified, whereas in Bayesian estimation, substantively driven small-variance priors bring information into the analysis which alleviates the nonidentification issue. As we will explicate in the following sections, the main idea of using informative priors to ensure model identification is applied in our proposed multi-stage estimation, in which informative priors are imposed on common items shared between different study cohorts to place their parameters on the same scale.

To summarize, we focus on three approaches in the study: the traditional unidimensional IRT model using data from a single cohort and single time point; the integrated model (namely, the multiple-group two-tier model) with concurrent calibration, and the integrated model with

multi-stage calibration. The latter two approaches complement each other in that if study design (or sample size) supports concurrent calibration, it is a statistically ideal choice. Otherwise, the multi-stage calibration is a robust alternative and more importantly, the Bayesian framework provides a principled way to conduct staged calibration that considers estimation uncertainties at each stage. The simulation results illuminate the conditions under which the simple unidimensional model approach produces comparable results, as well as quantifying the precision loss in other conditions. The real data example provides detailed analysis protocols for estimating individual change scores from the multi-cohort repeated measure design, including steps of evaluating model fit. The recommended two approaches perform strikingly well as the change score extracted therefrom are much better separated from measurement errors, thereby providing stronger signals to relate change scores with external variables.

The Multiple-Group Two-Tier Model (MGTT)

For the multi-cohort repeated measure design, a multiple-group two-tier model (MGTT), which is a straightforward extension of the two-tier model (Cai, 2010b; Cai et al, 2011), will serve as an integrated model. The well-studied multiple group IRT model (Davoudzadeh et al., 2020) is a special case of the MGTT model. However, implementing the MGTT model in practice is challenging due to model complexity. In what follows, we introduce the formal parameterization of the MGTT model, followed by a discussion of model estimation that sets the stage for our proposal of multi-stage Bayesian estimation.

The MGTT model is built upon the multidimensional graded response model (MGRM; Hsieh et al., 2010, Jiang et al., 2016; Wang et al., 2018) that includes the graded response model (Samejima, 1969), the two-parameter logistic (2PL) model and the multidimensional 2PL model (Reckase, 2009) as special cases. The model is suitable for outcomes such as symptom presence

(yes/no), symptom severity/frequency, or other types of polytomous responses. Assume item j has C_j *ordered* response categories that any response, l, to item j falls within the set of $l \in \{0, \dots, C_j - 1\}$.[2] For an individual i with a K-dimensional latent trait level $\boldsymbol{\theta}_i \in \mathbb{R}^K$, the probability of a response to item j, y_{ij}, is a function of $\boldsymbol{\theta}_i$, the item's K-dimensional discrimination parameter $\boldsymbol{a}_j \in \mathbb{R}_+^K$ and the boundary parameters $d_{j,1}, \dots, d_{j,C_j-1} \in \mathbb{R}^1$. The MGRM starts out by defining the boundary response function as follows, which is the probability of responding to response category l or higher, i.e., $P(y_{ij} \geq l)$:

$$P(y_{ij} \geq l | \boldsymbol{\theta}_i, \boldsymbol{a}_j, \boldsymbol{d}_j) = \frac{1}{1+\exp(-(\boldsymbol{a}_j^\top \boldsymbol{\theta}_i - d_{j,l-1}))} . \quad (1)$$

Then the probability for each response class can be given by the difference between two adjacent boundary response functions,

$$P(y = l) \equiv p_{jl} = P(y \geq l) - P(y \geq l + 1).$$

It is assumed that $\boldsymbol{\theta}_i$ follows a multivariate normal distribution with mean of μ and covariance matrix of $\boldsymbol{\Sigma}$.

Extending MGRM to a multiple-group two-tier version, let us assumes K=2, implying the existence of two main factors. It could be a single main factor measured twice. Then, a series of nuisance factors are added so that Equation 1 is updated as

$$P\left(y_{ij} \geq l \middle| \boldsymbol{\theta}_i, \boldsymbol{a}_j, \boldsymbol{d}_j, \boldsymbol{\eta}_i\right) = \frac{1}{1 + \exp\left(-\left(\boldsymbol{a}_j^\top \boldsymbol{\theta}_{i(g)} - d_{j,l-1} + \lambda_{j(g)}\eta_{ij}\right)\right)} , \quad (2)$$

[2] Here we assume the lowest score is 0, and the highest score is the number of total response categories minus 1. That is, for an item with 4 response categories, the list of possible scores would be: 0, 1, 2, and 3. We chose this parameterization just to be consistent with the convention of graded response models. If users use other parameterizations (such as 1, 2, 3, 4 for scores), that will yield equivalent results as well albeit slightly modified notations.

where η_{ij} is the nuisance factor for person i on item j assuming item j is administered repeatedly. In many cases, it is assumed that $\eta_{ij} \sim N(0,1)$ and all η_{ij} (for any item j)'s are mutually independent and they are all independent of $\boldsymbol{\theta}_i$. The subscript "g" denotes group, $\boldsymbol{\theta}_{i(g)}$ follows a multivariate normal distribution with mean of μ_g and covariance matrix of $\boldsymbol{\Sigma}_g$. Without group subscripts on item parameters $\boldsymbol{a}_j$ and $\boldsymbol{d}_j$, Equation 2 implicitly assumes that all items function the same for all people (i.e., differential item functioning [DIF] does not exist; Holland & Thayer, 1988; Penfield & Camilli, 2006; Woods, 2009; Woods & Grimm, 2011). The only exception is $\lambda_{j(g)}$, which implies that the loading of item j on the nuisance factor η_{ij} can vary across groups. This flexibility is intentionally built in the model for two reasons: (1) it is hard to justify the equality of loadings on nuisance factors over time; and (2) different $\lambda_{j(g)}$ per group does not lead to undesirable DIF because the introduction of nuisance factors is merely a re-partition of the residual covariances among items after controlling for the main factors. We further explain this second point in a remark below to show that the main factor scores will not be affected with the inclusion of the nuisance factor, highlighting the inherent connection between the bi-factor model and unidimensional model. Figure 1 presents an illustrative diagram of the model. In this figure, executive functioning (EF) serves as the primary factor that is measured twice and hence, they are correlated over repeated measures. In addition, there is another nuisance factor called "clock method" (Crane et al., 2012) that accounts for the shared commonality among five measures using the same clock stimulus, and it covaries over time as well. Further, nine additional nuisance factors are introduced to explain residual dependence between the same items administered twice. In this multi-cohort setting, the distribution of both EF and clock method factors can vary across groups.

Insert Figure 1 Here

Concurrent Estimation

Let's return to our motivation example in which there are two cohorts and two time points, and some items are shared between cohorts and measurement waves to establish a common scale. Whenever an item j is administered repeatedly, an η_{ij} is introduced to account for residual dependence. Without loss of generality, we assume the first group is the reference group, and then for model identifiability, $\mu_{11} = 0$ (i.e., the mean of the main factor at time 1, group 1), and the diagonal elements of Σ_{11} is set to 1, implying that the main factor (e.g., EF at time 1, group 1) has a mean of 0 and variance of 1[3]. All remaining elements in μ_1 and $\boldsymbol{\Sigma}_1$ as well as μ_g and $\boldsymbol{\Sigma}_g$ ($g \neq 1$) are freely estimable. As to the nuisance factors, because we only consider the same items administered twice, $\lambda_{j(g)}$ is constrained to be equal over time for model identification. Even so, the sign of $\lambda_{j(g)}$ is still indeterminant because flipping the sign of both $\lambda_{j(g)}$ and η_{ij} would yield equivalent models. There are two ways to resolve this indeterminacy: constraining all $\lambda_{j(g)}$'s to be non-negative or fixing $\lambda_{j(g)} = 1$ but estimating the variance of η_{ij} per group (denoted as $\sigma^2_{j(g)}$). We use the latter one in our study because it is easier to implement in Bayesian MCMC. The means of all nuisance factors are constrained to be 0.

Model estimation can proceed using either marginal maximum likelihood (MML) estimation or Bayesian MCMC. Although the MGTT model may be high-dimensional due to a potentially high number of nuisance factors, MML estimation will be greatly simplified using analytic dimension reduction (Cai et al., 2011; Gibbons & Hedeker, 1992, 2007; Rijmen et al., 2008) such that only a three-dimensional integral is necessary. Specifically, for the model

[3] The model shown in Figure 1 is slightly more complicated due to the introduction of "clock method" factor. As we will make it clear in the real data example, the mean and variance of the clock method factor at time 1 group 1 is set to 0 and 1 respectively, whereas the mean and variance of it in the second time points are freely estimated. The mean and variance of the clock method factor do not differ across the two cohorts as otherwise the model cannot converge. Clock method is uncorrelated with any other factors.

expressed in Equation 2, the marginal likelihood of model parameters, $\boldsymbol{\Delta}$, which includes $\boldsymbol{a}_j$, $\boldsymbol{d}_j$ $\lambda_{j(g)}$, for $j = 1, .., J$, μ_g, $\boldsymbol{\Sigma}_g$, for $g = 1, \dots, G$ subject to identification constraints stated above, is expressed as

$$L(\boldsymbol{\Delta}) =$$

$$\prod_{g=1}^{G}\prod_{i=1}^{N_g}\left[\int\cdots\int\prod_{j=1}^{J}\left\{\prod_{l=0}^{C_j-1}P\left(y_{ij(g)}=l\middle|\boldsymbol{\theta}_i,\boldsymbol{a}_j,\boldsymbol{d}_j,\lambda_{j(g)},\eta_{ij}\right)^{\mathrm{I}(y_{ij}=l)}\right\}N\left(\boldsymbol{\theta}_i\middle|\mu_g,\boldsymbol{\Sigma}_g\right)d\boldsymbol{\theta}_i d\eta_{i1}..d\eta_{iJ}\right] \quad (3)$$

Here $\mathrm{I}(y_{ij}=l)$ is an indicator function and it takes the value of 1 when $y_{ij}=l$ and 0 otherwise. Equation 3 is maximized to obtain model parameter estimates, $\widehat{\boldsymbol{\Delta}}$. The marginal likelihood in its plain form in Equation 3 involves a $(2+J)$-dimensional integral (assuming $\boldsymbol{\theta}_i$ is 2-dimensional), which is computationally prohibitive when J, the number of common items, is large. Instead, the marginal likelihood can be rewritten as follows, exploiting dimension reduction,

$$L(\boldsymbol{\Delta}) =$$

$$\prod_{g=1}^{G}\prod_{i=1}^{N_g}\int N\left(\boldsymbol{\theta}_i\middle|\mu_g,\boldsymbol{\Sigma}_g\right)\prod_{j=1}^{J}\left[\int\left\{\prod_{l=0}^{C_j-1}P\left(y_{ij(g)}=l\middle|\boldsymbol{\theta}_i,\boldsymbol{a}_j,\boldsymbol{d}_j,\lambda_{j(g)},\eta_{ij}\right)^{\mathrm{I}(y_{ij}=l)}\right\}d\eta_{ij}\right]d\boldsymbol{\theta}_i. \quad (4)$$

The derivation from Equation 3 to Equation 4 uses the simple calculus conclusion that we have $\int\int f(x_1)f(x_2)dx_1dx_2 = \int f(x_1)\,dx_1 \times \int f(x_2)\,dx_2$ when x_1 and x_2 are independent. As shown in Equation 4, the joint distribution conveniently factors into $(J+1)$ terms that are mutually independent, hence the $(2+J)$-dimensional integral is converted into a series of iterated integrals whose dimensionality is 3.

Despite the computational efficiency of analytic dimension reduction, if readers want to fit the MGTT model using general-purpose software, such as M*plus*[4], they need to be aware that

[4] Other software packages, such as *flex*MIRT, may handle dimension reduction differently.

this benefit may not always be fully utilized. That is, M*plus* will only enable dimension reduction when the model is expressed in a typical bi-factor form, i.e., (1) the loadings $\lambda_{j(g)}$ are freely estimated while fixing the nuisance factor variances to 1, and (2) there is only one group. Without dimension reduction, the number of numeric integrations is high enough that only Monte Carlo integration (in contrast to quadrature integration) is feasible. The idea of MC integration is to draw a random sample from a given distribution and then compute sample average. Based on the MC principle, the sample average provides a consistent estimate of the integral as sample size goes to infinity. As one can conveniently draw samples from a multivariate distribution, the MC integration is much more feasible, computation wise, than quadrature-based integration, and even so, it is extremely slow.

Two other full information[5] based methods that serve as alternatives to marginal ML are Metropolis-Hastings Robbins-Monro (MH-RM; Cai, 2008, 2010) and Markov chain Monte Carlo (MCMC; Patz & Junker, 1999; Robert & Casella, 1999; Wang et al., 2013; Wang & Nydick, 2015). MCMC methods circumvent intractable analytic or numerical integrations; however, they can be computationally intensive for complicated models because they typically require a large Monte Carlo sample size or a long chain to converge. MH-RM, as the name entails, combines elements from MCMC with stochastic approximation. It has a strict convergence criterion reminiscent of conventional maximization routines, and it has been successfully used for calibrating multigroup, multilevel, and multidimensional IRT models (Cai, 2010b). Although MH-RM is computationally much faster than MCMC (Cai, 2010a; Edwards, 2010), other studies found that MCMC outperformed MH-RM in terms of estimation accuracy in

[5] We used "full-information" to be in contrast to "limited-information" weighted least square methods, see Remark II below for details.

non-linear factor models in the presence of item cross-loadings (i.e., an item measures multiple factors) and when latent dimensions are highly correlated (Wang & Nydick, 2015).

In this paper, we will focus on the MCMC algorithm by treating MGTT from a Bayesian structural equation modeling (SEM) perspective. The advantage is, when the concurrent calibration fails to converge either due to data idiosyncrasy or model complexity (in both cases, MH-RM also encounters convergence challenges), MCMC can easily accommodate a divide-and-conquer type of staged estimation approach by handling uncertainties in separate stages of the estimation well, whereas other methods would have treated parameters obtained from a preceding stage as "fixed." Details regarding this advantage are presented in the subsection below. For MCMC, non-informative or weakly informative priors are used for all model parameters to reduce the effect of priors on parameter estimation to the largest extent.

Multi-Stage Estimation

Multi-stage estimation alleviates the challenge of estimating MGTT concurrently by only needing to estimate a single group two-tier model that is well studied (Cai, 2010b). Take a two-cohort design as an example. In stage I, a single group two-tier model is fitted using MCMC on the data set from the first cohort, although it could be any cohort arbitrarily chosen from the full data set. MCMC outputs the posterior mean and standard deviation of all model parameters. In stage II, the estimated posterior mean and standard deviation of the main loadings and threshold parameters of *common* items (shared between two cohorts) are used as informative priors and again, a single group two-tier model is fitted on the data set from the second cohort. The informative priors help fix the scale in stage II, and as we make it explicit in Remark I below, only informative priors of the loadings on the main factors need to be fixed whereas the loadings on the nuisance factors (or equivalently the variance of the nuisance factors) do not necessarily

need to be fixed, which allows more flexibility to maximize model fit[6]. The specific non-informative priors for all other model parameters are presented in the simulation section. Despite the informative priors, the posterior means of those common item parameters may still be updated, albeit slightly, in stage II estimation for cohort 2 data. To make sure that the change score estimates from both cohorts are on the same exact scale, a stage III estimation proceeds by fitting the single group two-tier model again on cohort 1 data but fixing the common item parameters (i.e., loadings on the main factors and thresholds) to their estimated posterior means from stage II. For data sets that contain more than two cohorts, this multi-stage estimation will proceed similarly but with more stages. For instance, with a three-cohort design, one will start with data from cohort 1, then informative priors obtained from cohort 1 will be passed on to cohort 2, and then the priors will be updated again and passed on to cohort 3. After this forward passage of information is finished, the common item parameters will be fixed and the fixed parameters will be passed on back to cohort 2 and cohort 1 to ensure that the latent trait estimates from three cohorts are on the exactly same scale. Please see Figure 2 for an illustration.

Insert Figure 2 Here

Remark I. When ignoring the residual dependence, the two-tier model reduces to a simple two-dimensional IRT model, or in a simplest case, a bi-factor model reduces to a unidimensional model. Here, we want to emphasize that the introduction of nuisance factors helps repartition the residual variances, without affecting the strength of relationships (i.e., standardized loadings) between the main factor and item responses. This point is essential to understand the constraints needed to establish invariance of bi-factor models across groups or

[6] The informative priors are also imposed on the threshold parameters to help fix the origin of the scale.

longitudinally. That is, in a simple unidimensional two-parameter IRT model, the constraints needed to establish the scale is either to fix one item discrimination and difficulty parameter, or to fix the mean and variance of θ at certain constants. For a bi-factor model structure, written in factor analytic format as

$$y_{ij}^{*} = \lambda_j \theta + \gamma_j \xi + \varepsilon_{ij},$$

fixing a λ_j is sufficient to fix the metric of θ, and there is no need to fix γ_j. The contribution of $\gamma_j \xi$ is to repartition the residual variance of ε_{ij} relative to the variance of y_{ij}^{*}, i.e., $var(\varepsilon_{ij}) = var(y_{ij}^{*}) - \lambda_j^2 var(\theta) - \gamma_j^2 var(\xi)$, whereas in a simple unidimensional model written as

$$y_{ij}^{*} = \beta_j \theta + \zeta_{ij},$$

the residual variance is $var(\zeta_{ij}) = var(y_{ij}^{*}) - \beta_j^2 var(\theta)$. Note that $\beta_j \neq \lambda_j$ but the standardized factor loadings are equivalent, i.e., $\frac{\lambda_j \sqrt{var(\theta)}}{\sqrt{\lambda_j^2 var(\theta) + \gamma_j^2 var(\xi) + var(\varepsilon_{ij})}} = \frac{\beta_j \sqrt{var(\theta)}}{\sqrt{\beta_j^2 var(\theta) + var(\zeta_{ij})}}$. There are two take-away messages from this remark. First, when fitting a bifactor model (or two-tier model), in our concurrent calibration, we only need to constrain the loadings on the target factor, λ_j (or $\boldsymbol{a}_j$ from Equation 4, and of course item threshold parameters, $\boldsymbol{d}_j$), to be equal across groups and across time, without putting any equality constraints on nuisance factor loadings (i.e., γ_j) across groups. Similarly, in the three-stage calibration, only informative priors on $\boldsymbol{a}_j$ and $\boldsymbol{d}_j$ are needed. Second, ignoring nuisance factors will not bias the point estimation of θ, although its standard errors may inflate.

Remark II. Aside from the full information method, limited information weighted least square estimation (WLS) is a viable alternative for the two-tier model. WLS uses first-order and second-order marginal proportions obtained from response contingency tables to facilitate parameter estimation. The main idea is to find item threshold and loading parameter values such that they minimize the weighted deviations between the model-implied correlation matrix and

the sample tetrachoric correlation matrix. Because WLS usually requires a large sample to precisely estimate a full weight matrix (Flora & Curran, 2004; Muthén et al., 1997), researchers have suggested using only the diagonal elements of the weight matrix for estimation, leading to the diagonally weighted WLS estimators. Parameter estimates obtained from WLS can be translated to IRT parameters (Takane & de Leeuw, 1987). Because WLS works directly with the item response contingency table, it permits the fitting of high-dimensional models with much reduced computation time as compared to ML based methods. Moreover, because WLS is developed within the SEM framework, off-the-shelf absolute model fit indices such as the root mean squared error of approximation (RMSEA), the Tucker Lewis Index (TLI), and the confirmatory fit index (CFI) (Bentler, 1990) can be used to evaluate absolute model fit, and chi-squared difference tests can be used to evaluate relative fit of nested models. A widely recognized limitation of the WLS approach, however, is the difficulty of estimating tetrachoric (or polychoric) correlations, especially in the presence of missing data. This insufficiency of handling missing data makes WLS not suitable for the concurrent estimation of the MGTT model that may inherently have missingness by design. However, we still use WLS for single group two-tier model in the real data illustration to gather complementary model fit information.

Simulation Study

We conducted three simulation studies to compare the performance of MGTT via concurrent calibration and multi-stage estimation (denoted as "Concurrent" and "Multi-stage" respectively hereafter). A single unidimensional model (denoted as "Single" hereafter) that ignores the multi-cohort repeated measure structure, which is the current status quo in many psychology and health measurement studies, was also included in the study as a baseline. The main evaluation criterion is the recovery of individual latent change scores, along with the latent scores at both baseline

and follow up(s). In study I, we considered a two-cohort two-time points design, with eight manipulated conditions. In study II, we considered a three-cohort two-time points design, with four manipulated conditions, and in study III, we considered a two-cohort three-time points design, with two manipulated conditions. While study I was more comprehensive and it was consistent with the data design from our motivating real data example, study II and III were included to ensure generalizability of the findings.

Study I Design. Eight manipulated conditions were considered to imitate the real data collection design scenarios as closely as possible while allowing for generalization of the results. Assume there were two time points and two study cohorts and assume full measurement invariance across cohorts and longitudinally. Sample size was fixed at 800 per cohort throughout all three simulation studies to be consistent with the real data example. Two assessment designs were considered to follow ADNI EF and language assessment domain respectively. Item discrimination was simulated from U(0.70, 1.65) (Jiang et al., 2016), and thresholds were generated based on ADNI language domain analysis results. The same true item parameters were used in both item designs so that results from them are directly comparable. Table 1 presents the true item parameters for simulation studies.

Insert Table 1 here.

As shown in Table 1, item design I results in a complete overlap of items across two cohorts, establishing a strong link for a common scale across groups, whereas the percentage of overlapped items between two measurement occasions is only 50%. Hence, out of 9 items in total, only three items load on nuisance factors yielding non-zero γ's. In item design II, there is 50% overlap of items between two cohorts, whereas there is 100% sharing of items across two measurement occasions. As a result, all 9 items have non-zeros γ's. Item design I leads to a

simpler MGTT model with only 5 latent factors (2 main factors and 3 nuisance factors) whereas an 11-factor model is needed for item design II. Table 2 presents the design details for generating true θ's. Person designs I & II generate a bivariate normal distribution of θ's such that the θ correlations between baseline and follow up time points are manipulated via the covariance parameter. These first two designs differ in the level of correlation and the impact between the two cohorts. Person designs III & IV consider specific change pattern over time, which results in a high serial correlation between baseline and follow up traits at around 0.85.

Insert Table 2 here.

Estimation. We used MCMC implemented in M*plus* for model estimation and used default non-informative or weakly informative priors on all model parameters except in stage II of the three-stage estimation method described above. Specifically, for factor loadings and thresholds, a normal distribution N(0, 5) (where 5 is the prior variance) was used as prior. With MCMC estimation, factor loading reflection (i.e., sign change) can be a challenge as the iterative process constructs a bimodal posterior distribution by sampling positive and negative values for the loadings but converging on neither estimate (Bauer, et al., 2013). M*plus* unfortunately does not support truncated normal priors for loadings in its current version. This may not be a serious problem for loadings on main factors because there are more than three items loaded on a single factor, and only flipping all signs of the loadings simultaneously would result in an equivalent model (i.e., the item covariance matrix stays intact), which happens rarely. Loadings on nuisance factors, however, may be problematic because only two items load on each nuisance factor, and the loadings are constrained to be equal for model identification purposes. In this case, flipping the sign would happen much more often. We therefore decided to estimate the variance of the nuisance factors instead. We used inverse-Gamma (-1, 0) as priors for nuisance variances. For all

factor means that are freely estimable, we used a N(0, ∞) prior, and for factor covariance matrix, we used an inverse-Wishart (IW) prior $\left(\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}, 3\right)$. When a certain factor variance was fixed during estimation, say variance of θ_1 was fixed, then IW(0,3) and IW(1,3) were used as priors for cov(θ_1, θ_2) and var(θ_2) respectively.

The chain length was set at 50,000. Model convergence was assessed using the default Gelman-Rubin convergence criterion based on the potential scale reduction factor (PSR) for each parameter (Gelman & Rubin, 1992), and a cutoff of 1.1 was used (Muthén & Muthén, 1998-2017). Only a single chain was used, and discarding the first half as burn-in, the last half of the iterations was split into two quarters and the PSR factor was computed for these two quarters. By default, If PSR>1.1 by the end of the chain length, we increase the chain length to 100,000, and if PSR continues to be higher than 1.1, we conclude MCMC does not converge.[7] We set "THIN=1" by default, which implies every MCMC iteration is saved. Fifty replications were conducted per condition.

Results for Simulation Study I.

Table 3 presents the recovery of the latent traits at baseline and follow up and the recovery of change scores, in terms of bias, root mean squared error (RMSE), mean absolute bias (ABS), and mean standard error (SE) of the corresponding trait (or change score) estimates, under item design I. Table 4 presents the same results under item design II. In table 3, results from all three methods (i.e., single, concurrent, and 3-stage) are included whereas in table 4, results from only single and 3-stage methods are included because concurrent calibration did converge under item design II.

[7] We also tried to increase the chain length to 150,000 but it still failed to converge, with PSR consistently staying above 2. For the conditions where concurrent calibration did not converge, none of the replications converged.

Insert Table 3 here

In Table 3, the four person designs correspond to those defined in table 2. Several trends can be observed from the results. First, the three methods produce nearly comparable results under person design I, although the simplest single method generates slightly higher RMSE, mean ABS, and mean SE. Unsurprisingly, the results based on change score tend to magnify the difference among the three methods more than merely looking at the scores as either baseline or follow up. There is no appreciable difference between the results from cohort 1 versus cohort 2. Second, under three other person designs, the performance of the three methods begins to diverge. That is, both concurrent and 3-stage methods still generate almost the same results, but the single method produces much larger RMSE, mean ABS and mean SE, especially for the change scores. The mean bias is not sensitive to differentiate different methods. Although person designs II-IV all have unique features, the results and trend from them seem to be quite comparable. Therefore, one conclusion to make is, when there is high correlation between baseline and follow up data, ignoring such correlations would not bias the traits or change score estimates, but would make such estimates less efficient as reflected by the large standard error and hence large RMSE from the single method. Further, the difference between θ distributions across two cohorts (which is called impact in IRT DIF literature) does not seem to have much effect on trait recovery, as the results from design III and IV are close. Third, an interesting finding about mean SE is that, under person designs II-IV, the concurrent method generates slightly higher mean SE than the 3-stage method, and such difference is more salient when looking at the mean SE of change scores especially under person design II. This may be because the MGTT model contains many parameters in the concurrent calibration, such that estimation instability contributes to the larger SE.

Insert Table 4 here

Although concurrent method is not shown in Table 4, the same patterns show up in table 4 as well. That is, under person design I, both single and 3-stage method produce similar results across both cohorts, but the 3-stage method outperforms the single method by a large margin under person designs II-IV. Hence, combining results from both tables, we recommend the 3-stage estimation method for its robustness and estimation accuracy. If the model converges, concurrent calibration is also recommended as it is easier to implement than the 3-stage method, although the standard errors of latent traits and the change scores may be slightly inflated.

Study II Design and Results. Table 1 presents the true item parameters for the three-cohort two-time design. As in study I, we also considered two item designs, one with 50% item overlap overtime but 100% item overlap across cohorts, whereas the other one with 100% item overlap overtime but 50% item overlap across cohorts. As to person parameters, we followed the first two designs in Table 2 to generate true θ's for cohorts I and II, and for cohort III, we used $\theta \sim mvn([0.2, 0.4]', \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = [1, 0.4; 0.4,1]$ for design I and $\theta \sim mvn([0.25, 0.30]', \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = [1, 1.1; 1.1,1.3]$ for design II. The two person designs vary mainly by the level of correlation of main θ across two time points, i.e., median (0.4) and high (≈ 0.9). Fifty replications were conducted per condition, and the same estimation methods were considered. The only difference is for the multi-stage estimation, instead of performing three stages of estimation, we performed five stages of estimation as shown in Figure 2. Note that since the number of item parameters remain the same between study I and II (i.e., comparing Table 1 and 5), and because we assume measurement invariance across cohorts, including an additional cohort is like increasing sample size. Although the mean and variance of latent variables are freely estimated in cohort III, resulting in a slight increase in the total number of model

parameters, it is expected that different methods should not encounter extra convergence issues compared to study I.

Insert Tables 5, 6 here

Tables 5 and 6 present the latent trait and change score recovery for person designs I and II respectively. The same results pattern shown in simulation study I continue to hold here. That is, for item design I, all three methods converged properly across all replications. When the level of correlation of θ across two time points is medium, θ recovery at both baseline and follow up time points appear to be similarly precise across three methods, although the single method tends to produce slightly larger RMSE, mean absolute bias and mean SE, especially for change scores. When the level of correlation of θ across two time points is high, the improvement of MGTT over traditional single method is much more salient. Under item design II, again the concurrent calibration fails to converge even when the sample size increased by 50% (compared to study I due to the inclusion of cohort 3). Otherwise, the same conclusion holds as well. That is, the multi-stage and single method produce similar results for person design I, but the multi-stage method produces more accurate change score estimates than the single method for person design II. Across three cohorts, it appears that the bias is similar, but the RMSE and mean ABS are considerably smaller in cohort 3 than those in cohort 1 in person design II. This is merely due to the current data generation scheme for person design II, i.e., the correlation between θ's over time is .88, .95, and .96 for the three cohorts respectively. Higher correlation results in smaller range of change scores, and indeed, the "true" change scores across three cohorts have mean values of 0.054, 0.053, and 0.050 and standard deviation values of 0.540, 0.349, and 0.313, respectively. Because the true range of change scores are smaller in cohort 3, the RMSE and

mean ABS are also smaller. In addition, due to the high correlation in cohort 3, the improvement of two MGTT methods over the single method is also largest in cohort 3.

Study III Design and Results. In this study, a two-cohort three-time design was considered. Given that both study I and study II provide consistent evidence regarding the two item designs, we only focused on item design I here. That is, there is 50% item overlap between any two time points and 100% item overlap between two cohorts. Each item is shared across, at most, two time points. Note that if an item is shared across more than two time points, then instead of fixing the loadings on the nuisance factor to be 1 and let nuisance factor variance to be freely estimated, we can instead fix the nuisance factor variance to be 1 and let the loadings on the nuisance factor be freely estimated. This would ensure maximal model flexibility while still guaranteeing model identifiability. The true item parameters were presented in Table 1, and there were 12 items in total instead of 9 as in the previous studies. Two person designs were considered, and they differ by the level of correlation over time. Specifically, for design I, cohort I latent trait was generated from a multivariate normal distribution, $\boldsymbol{\theta} \sim mvn([0, 0.2, 0.3]', \boldsymbol{\Sigma})$, and cohort II latent trait was generated from $\theta \sim mvn([0.1, 0.3, 0.45]', \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ has diagonals 1 and off-diagonals 0.4 (Kuhfeld & Soland, 2022). For design II, cohort I latent trait was generated from $\theta \sim mvn([0, 0.05, 0.2]', \boldsymbol{\Sigma})$ and cohort I latent trait was generated from $\theta \sim mvn([0.5, 0.55, 0.3]', \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ has diagonals 1 and off-diagonals 0.85 (again following ADNI language data analysis results). Note for cohort II design II, we intentionally considered a non-monotone growth pattern just so that the results are not only restricted to monotone growth patterns. Also note that the number of parameters slightly increased compared to study I due to the inclusion of 3 more items, but the number of nuisance factors was 6, in between 3 and 9 in

the two designs in study I. In addition, adding a third time point also brought in additional data and hence, all three methods converged successfully across all replications.

Insert tables 7 here

Table 7 presents the latent trait and change score recovery from all three methods under the two person designs. Unsurprisingly, under person design I, there is not much difference among the three methods in terms of the precision of both latent traits and the change scores. Under person design II, both concurrent and multi-stage estimation outperform the single method by a large margin. This is reflected in the smaller RMSE, smaller mean ABS, and smaller mean SE for both latent trait and change scores across all three cohorts.

Real Data Illustration: ADNI

In this section, we used two-cohort repeated measure data from the Alzheimer's Disease Neuroimage Initiative (ADNI) to demonstrate different approaches that can be used to extract useful individual-level change scores. ADNI has had several funding cycles, with somewhat different enrollment goals, and we used data from ADNI 1 and ADNI 2 / ADNI GO cohorts. Specifically, ADNI 1 enrolled people with normal cognition, mild cognitive impairment (MCI), and AD in a 1:2:1 ratio during 2004 to 2010. In ADNI 2/ ADNI GO which occurred between 2009 and 2017, participants from ADNI 1 who met the enrollment criteria were carried forward for continued monitoring, while new participants were added to further investigate the evolution of AD. Overall, ADNI enrolled participants between the ages of 55 and 90 who were recruited at 57 sites in the United States and Canada. After obtaining informed consent, participants undertake a series of initial tests that are repeated at intervals over subsequent years, including a clinical evaluation, neuropsychological tests, MRI, and PET scans, among others. We focus on

the cognitive battery that is administered at each study visit of ADNI. The battery includes tests on four main domains (i.e., memory, executive functioning, language, and visuospatial functioning), and granular data are available from the LONI website (http://adni.loni.usc.edu/).

We aimed to extract individual learning effects from longitudinal ADNI data (i.e., both ADNI 1 and ADNI 2 / ADNI GO cohorts), which are quantified as change scores between baseline visit and 6-month follow up on the cognitive batteries. We hypothesized that individuals who were less able to learn as evidenced by either decrement in functioning between time points or small amounts of improvement compared to others would be at higher risk of conversion to AD (Jutten et al., 2020). Combining data from two study cohorts offers increased power due to larger sample size and enhanced external validity due to greater heterogeneity in samples. Only two time points were used because from the sponsor's perspective, it is cost-efficient to identify patients who are at highest risk of AD within the shortest time possible. We acknowledge that one follow-up point would enable enrollment after 6 months of contact, which may be feasible and desirable in fast-paced clinical trials, but necessarily limits the amount of longitudinal data available to characterize who could be at highest risk based on changes in cognition.

Because different sets of items (with some overlapping items) are used in different study cohorts, co-calibration, which is a valuable form of data harmonization, is needed to produce scores on the same metric. This is essential to facilitate combining data from different studies. Previous studies (Choi et al., 2020; Crane et al., 2012, 2021) have used data from ADNI 1 baseline visit to calibrate IRT models and fixed the item parameters in follow up visits and ADNI 2 / ADNI GO to extract factor scores at each time point and from each cohort all separately. In this section, we compared this status-quo approach versus the MGTT integrated approach, in terms of the raw individual latent change score, as well as the power of detecting

highest risk of conversion from MCI to AD using the extracted change score as predictors. Due to space limitations, only two domains were considered, Language and EF, as these two domains represent two different item designs that lead to different choices of estimation approaches: three-stage estimation versus concurrent calibration respectively. EF is also unique for illustration as we included additional nuisance factors to explain shared variances among items using the "clock method" (Gibbons et al., 2012).

Sample. For the language domain, the sample size for baseline and follow-up time points for ADNI 1 are 781 and 780 respectively, and for ADNI2/GO are 835 and 835 respectively. For the EF domain, the sample size for baseline and follow-up time points for ADNI 1 are 782 and 780 respectively, and for ADNI2/GO are 835 and 832 respectively. There is little to no data attrition during the 6-month follow up.

Item Design. Table 8 presents the item designs for ADNI1 and ADNI 2 / ADNI GO language and EF domains. As shown, there is a great overlap between the test batteries to establish common scales. For the language domain, there are six items from the Mini-Mental State Examination (MMSE), three language tasks in the AD Assessment Schedule - Cognition (ADAS-Cog), six language items from the Montreal Cognitive Assessment (MoCA), and three items that were not part of a global cognition composite. Four MMSE items were dropped for modeling because they were so easy that very few people got them wrong. The Category Fluency-Vegetable item was only used in ADNI1, and the four MoCA items, Animal Naming – Rhino, Sentence Repetition Task 1, Sentence Repetition Task 2, and Letter F Fluency, were only included in ADNI2 / ADNIGO. The baseline (initial visit) and the follow-up (six-month visit) in both phases used the same items. As to the EF domain, there were nine items used in both ADNI

1 and ADNI2 / ADNIGO cohorts at both visits. Five of these items shared the same common variance of "clock method."

Insert Table 8 here.

Analysis. The analysis procedure included the following: (1) variable recoding; (2) model selection; (3) extraction of individual change scores; and (4) uses of change scores for predicting conversion from MCI to AD. These four steps were conducted separately for each domain.

Collapsing response categories. In our preliminary check, we found that sparse responses (i.e., too few endorsements of certain response options of an item) would make the MCMC algorithm fail to converge, just because there is not enough information to properly estimate the corresponding threshold parameters. So, we collapsed categories to ensure proper convergence.

Specifically, for ADNI1, we merged score categories with less than 20 individuals into the adjacent category for items with more than two categories and dropped binary items which have less than 20 people in one response category. Specifically, for items with more than two response categories, the first and the last few categories in many cases had response counts less than 20, so the first few categories were merged one category up and the last few categories were merged one category down. For ADNI2 / ADNIGO, we kept the categories the same as ADNI1 for the common items because the item parameters were constrained to be equal across the two time points to establish a common longitudinal scale. For items not in ADNI1, we still used 20 as a cutoff for collapsing categories. We dropped Four MMSE and two animal naming items from the language domain since they had fewer than 20 people who got each item incorrect. Details of recoding for the two target domains are shown in the Supplemental Material.

Model Selection. Per each cohort and each domain, we first compared a longitudinal graded response model (GRM) with and without correlated residuals. The longitudinal GRM is a two-

dimensional GRM that contains two correlated main factors governing the responses from each time point. The model with correlated residuals contains nuisance factors for each of the common items between the two time points. Note that in the EF domain, we also included an additional secondary factor to account for the Clock Drawing methods effect as in Gibbons et al. (2012), see Figure 1 for details. The prior psychometric validation work conducted by the ADNI research team (e.g., Choi et al., 2020; Crane et al., 2012, 2021; Gibbons et al., 2007) allows us to make assumptions about measurement invariance over cohorts and time, which in turn allows for the focus on studying change in the latent variables over time and how best to associate and explain such change with external variables (i.e., time to conversion from MCI to AD). However, we still evaluated the invariance assumption at the item level through the posterior predictive p-(PPP) values produced as a byproduct of the MCMC algorithm. We used WLSMV and MCMC implemented in M*plus* for parameter estimation at this step. For MCMC, default non-informative priors described in the simulation study section were used on all unknown parameters. The only exception, as the method entails, is when we conducted multi-stage (in this case, 3-stage) estimation, in which we used the posterior mean and standard deviation of common item parameters from a preceding step in the next step.

For WLSMV, we used criteria of CFI≥0.95, TLI≥0.95, and RMSEA≤0.06 to evaluate whether a fit is acceptable (Hu & Bentler, 1999). As to MCMC, M*plus* did not provide Deviance Information Criterion (DIC; Spiegelhalter, et al., 2002) as there are different versions of DIC for hierarchical models and no census has been reached. Indeed, DIC will be computed differently depending on whether the joint likelihood or marginal likelihood is used. Furethermore, if the marginal likelihood is used, DIC will be computed differently depending on the level of the model at which the marginalization is computed (Zhang, et al., 2019). M*plus* reports PPP

value(Asparouhov & Muthén, 2021, Levy et al., 2009). However, we caution against using the overall PPP value as an evaluation of overall absolute model fit because it is computed based on the classical likelihood ratio chi-square test (see Equation 25, p. 30, *Mplus Bayesian technical manual*) that assumes multivariate normality of the data. Instead, PPP values reported at the item level provide much fine-grained information which is informative to make small adjustments of the model (i.e., relaxing certain measurement invariance constraints) to improve fit. PPP values below 0.05 indicate poor fit. Formally checking for measurement invariance was not the focus of the current study. Nevertheless, in the two-tier GRM, we constrained item parameters to be invariant over time, if there are items with PPP values less than 0.05, we could relax the longitudinal invariance assumption and re-evaluate whether PPP values become acceptable.

Insert Table 9 here

Table 9 presents the model fit results from both WLS and MCMC. When using WLS, we have CFI, TLI, and RMSEA as absolute fit indicators and chi-square difference test for comparing nested models. Results show that for both domains and cohorts, the longitudinal model with correlated residuals exhibits much better fit than the model without correlated residuals. The model with correlated residuals generated acceptable PPP-values (i.e., 0% of the items had extremely small PPP-values), although those from the model without correlated residuals also seem to be acceptable. PPP-values may be less sensitive as a model fit index in this application. Using WLS as an alternative for the purpose of evaluating model fit provided useful complementary evidence.

MCMC Estimation. For the EF domain, we fit the MGTT model on the combined data via concurrent calibration, but for the language domain, MGTT failed to converge. We thus used the three-stage estimation approach for the language domain. As described earlier, we achieved scale

determinacy in stage II estimation by estimating parameters of common items toward their prior means obtained in stage I. We used PPP values to monitor the degree of shrinkage towards the prior mean, which yielded scale determinacy through fixing the common items via the prior while still maintaining reasonable model fit. During the three-stage estimation, none of the item level PPP-values were smaller than 0.05. Figure 3 shows the common item parameter estimates in stage I and stage II for the language domain from the 3-stage estimation approach, just to illustrate that using informative priors can fix the scale at stage II as the common item parameters hardly differed, although the uncertainty of the parameter estimates in stage I was well considered in stage II.

Insert Figure 3 here.

We focused on the change score as possible indicators of risk of conversion from MCI to AD. Here the dependent variable is the time to conversion from MCI to AD. Results from the Single method were used as baseline. We first conducted K-means clustering of the change scores, aiming to find data-driven cutoff scores to bin participants into five categories: a lot better (in terms of cognitive domain at time 2), a little better, no change, a little worse, and a lot worse. Then we used the generated categorical indicator of change score as predictors in a Cox proportional hazard model to predict the time to conversion. We did not use raw change scores in the Cox model because that would impose a linear relationship between change score and rate of conversion. Table 10 presents the number of participants classified into each category per domain per method. As shown, the single method classified more individuals into the middle three categories whereas the spread of categories was broader with our recommended method.

Insert Table 10 here.

Table 11 presents the results from Cox PH models, using the third cluster (i.e., no change) as the reference group. Hazard ratio, which is the ratio of the hazard rate in the respective cluster relative to the control cluster (i.e., 3rd cluster of no change), implies how often the conversion from MCI to AD occurs in the target cluster versus the control cluster over time. For instance, 0.55 in Table 11 implies that, according to the traditional single method, the hazard of conversion for participants who have a lot better language scores in the follow up assessment is only 55% of the that of participants whose language scores do not change in the follow up assessment; whereas 2.42 implies that, according to the concurrent method, the hazard of conversion for participants who have a lot worse EF scores in the follow up assessment is 2.42 times higher than that of participants whose EF scores stay almost the same during the 6 month period. In sum, hazard ratios significantly different from 1.0 implies different hazard ratios of conversion in respective groups, and they are highlighted in Table 11. As shown, the clusters formed based on change scores estimated from the 3-stage or concurrent method predict conversion to AD much better than those from the single method. This real data example further supports our recommendation that the longitudinal model with correlated residuals (or MGTT if the model converges) should be used for multiple cohort repeated measure design.

Insert Table 11 and Table 12 here

Because a larger sample size in a cluster would yield higher statistical power, to further rule out the effect of cluster-level sample size differences from the two methods, we conducted two sensitivity analyses. In the first one, we recreated the 5 clusters from the "Single" method by ordering its change scores from smallest to largest and placed cutoffs such that the sample sizes per cluster were the same as those from the "3-stage/concurrent" method. This way, the confounding factor of sample size is eliminated from the comparison, whereas the actual patients

assigned to each cluster differ based on their change scores computed from respective methods. The results are presented in the first part of Table 12. The result is not much different from the "Single" method presented in Table 11, and still the results from the "3-stage/ concurrent" method show much stronger and useful signals. Similarly, we used the cluster size obtained from the "Single" method as a reference and adjusted the clusters from the "3-stage/concurrent" method accordingly. The results are presented in the second part of Table 12, and again there is not much difference except cluster 1 from the language domain. This sensitivity analysis further supports that the proposed method generates individual latent change scores that contain stronger signals as they are shown to be better predictors of conversion from MCI to AD.

Discussion

Tests and surveys are routinely used in education to track students' learning progress. Similarly, the advocacy of evidence-based practices in clinical psychology has also led to heavy reliance on tests and questionnaires (e.g., Garland et al., 2003) as regular outcome monitoring measures. Clinicians may use a patient's change score on tests or clinical scales to assess the degree to which the patient responds to the treatment and exhibits progress. Such information could also be used to influence decisions about subsequent treatment plans, health policies, and funding (Jacobson et al., 1984). Or cognitive neuropsychologists use lack of practice effects as a marker of cognitive decline, which may be a valuable input for a cost-effective strategy to select individuals who are at-risk for dementia for future interventions (Jutten et al., 2020). Tests used to measure individual change have been incorporated in large-scale research projects, such as the Patient-Reported Outcomes Measurement Information System (PROMIS), among others.

Many of the instruments used in education and psychology are psychometrically

validated, multi-item surveys/questionnaires. The calibrated item parameters provide crucial reference points with which comparability within and across studies can be achieved, therefore substantially improving cumulative science and replication (Cai & Houts, 2021). For instance, a typical clinical trial design has patient recruitment in multiple sites, use of randomization, and multiple follow ups. In such cases, using extant instruments with well-calibrated IRT parameters is convenient for clinical researchers because not all trials have adequate sample size needed for stable IRT model calibration (e.g., Jiang et al., 2016). Hence, using data from a single cohort and a single time point for IRT model calibration and then fixing the item parameters in all subsequent uses is a strategy that is used in many circumstances. In contrast, we aim to leverage the advanced multiple group two-tier model (Cai et al., 2016) and MCMC algorithm to demonstrate the precision gain that this added model complexity can bring in terms of latent trait estimation.

We study two complementary approaches: concurrent calibration of MGTT and multi-stage estimation. The former approach is statistically optimal in that it utilizes all available data in one integrated model estimation. Yet, due to model complexity, convergence may not be reached, especially when sample size is small. The latter approach provides a robust alternative. In the multi-stage approach, to preserve continuity between separate analyses for parameters that carried over from one cohort to another, and to establish a common scale, the posterior mean and standard deviation of the common item parameters from one cohort are used as prior distributions in the analysis for the next cohort. The simulation results reveal that there is no appreciable difference between these two approaches and in some cases, the concurrent calibration may generate slightly inflated standard errors of latent trait and latent change estimates, whereas the multi-stage method performs consistently well across all manipulated

conditions. Both approaches outperform the status quo single method, and the difference was considerable when data from repeated measures are highly correlated. In terms of computation time, when we fixed the Markov chain length to 50,000 across the board, the concurrent calibration (when it converges) is more efficient than multi-stage estimation because for the latter, computation time adds up across different stages. To optimize computation time, one may let Markov chain stop if PSR is smaller 1.1 instead of letting the chain length reach 50,000 universally because many stages in multi-stage estimation actually converge fast.

Our real data study provides additional, strong, supportive evidence. Researchers have previously developed composite IRT scores from the ADNI battery for memory (Crane et al., 2012), executive function (Gibbons et al., 2012), and language and visuospatial (S. E. Choi et al., 2020) using the single method. In each of these papers they evaluated comparative validity of composite scores by checking the strength of association between several imaging, fluid biomarker, and clinical comparisons with their composite scores and found that the composite scores outperform traditional scores such as total scores. In our real data example, we further show that the latent change scores derived from our recommended approaches provide an even stronger signal to predict risks of conversion from MCI to AD. This finding lends crucial external validity to the latent change scores derived from our recommended approaches.

In summary, the goal of the current study is to extract precise latent change scores for each individual, which will then be submitted to further analysis, such as clustering and Cox PH model used in our real data example. Certainly, obtaining reliable individual change scores is of practical importance in its own right if the focus is on evaluating intervention effect at an individual level, such as assessing psychometrically significant intra-individual change (e.g., Wang & Weiss, 2017, Wang et al., 2020). Further, this analysis protocol is operationally simple,

and it can entertain a full suite of secondary analyses. However, caution needs to be exercised in extrapolating individual level intervention effects to the group level, as Mislevy et al. (1992) noted that, even using individually optimal latent proficiency or change score estimates, population inference can still be severely biased. Instead, a marginal inference procedure such as a latent regression model with an IRT model on the outcome side of the regression equation would produce unbiased population inference. That said, the latent regression model is not without limitations either. For instance, careful thoughts need to go into the selection of covariates in the latent regression model to ensure congeniality of additional secondary analysis using the latent scores produced therefrom (Xie & Meng, 2017). Plus, if the secondary analysis does not fall within run-of-the-mill regression models, such as clustering analyses or survival analysis, then it would require considerable methodological advancements to integrate the respective models with IRT models. Clustering, for example, would add a layer of mixture components on top of the already complex MGTT model. Model convergence may be at risk unless the sample size is adequately large. Otherwise, alternative methods that handle measurement errors may be adopted, such as Wang et al. (2019) for regression models and Su et al. (2018) for clustering analysis.

Limitations and Future Research

A few limitations of the current study need to be emphasized. First, like in many simulation studies, we are limited in terms of the range of conditions, models, and assumptions we may use to evaluate and compare the methods. For instance, we did not use a fully crossed person design by crossing the magnitude of impact with the magnitude of correlations between repeated measures. Although from the current simulation results, we can infer that correlation contributes the most to the performance difference of the three methods, a fully crossed design

may have gotten the message across more straightforwardly. Further, although simulation study II and III provide further empirical evidence to support the uses of MGTT beyond just two cohorts and two time points, they are by no means comprehensive. If researchers decide to use MGTT on their real data, it may be advisable to conduct a simulation study that mimics their real data design just to gauge the discrepancies that could result when using different approaches.

Second, we assume the multivariate normality assumption holds, which could certainly be violated in practice. While past studies have shown that modest departure from multivariate normality do not deteriorate the results too much (e.g., Flora & Curran, 2004; Wang et al., 2019), to our knowledge, none of the studies have focused on individual latent change scores. Hence, this will be worth exploring in the future as change scores are affected by latent trait estimates from both time points in a compound fashion, and larger bias may result.

Third, in our real data example, although we evaluated model fit via WLS, we did not conduct a formal check on longitudinal measurement invariance (Widaman et al., 2010). The guiding philosophy of our real data analysis protocol is not to attempt to unearth the singular correct data-generating model, but rather to fit a model that produces the most stable and reliable inferences. While modern technology allows detailed features of single item response functions to be inspected, this should not lead to over-reporting of small details. At the minute level, all models are somewhat incorrect. Hence, the main question should be whether a model discrepancy seriously influences major conclusions. That said, a formal longitudinal invariance check may be needed if many items exhibit extremely small PPP-values. We are comforted here that our PPP findings were not very small. Further, in our analyses, we collapsed item response categories if the number of counts of endorsements is below 20. This cutoff was handpicked based on prior studies (e.g., Crane et al., 2012, 2021), whereas more systematic checks could be

applied in the future. In fact, some of our items in ADNI have up to 10 response categories, and in theory, treating those responses as continuous should work too, though that also makes linear assumptions on the relationship between item responses and the latent ability level which may not be correct. We treated these multi response category items as categorical so that the model retained flexibility to address this possible non-linearity.

Fourth, in our modeling approach, the latent change score is obtained by subtracting the latent trait estimates from two time points. In contrast, we can use an alternative parameterization first laid out in Embretson (1991) and McArdle (2009), wherein the latent trait from each subsequent occasion, θ_{it}, for person i at time t, is represented by $\theta_{i1} + \zeta_{it}$, $(t \geq 2)$. Here ζ_{it} is a latent change score that can be directly output from model estimation. Although this parametrization would be basically the same as our models in Equation 1 or 4, the advantage is one can directly regress ζ_{it} on external predictors such as treatment vs. placebo group. This way, the coefficient of the group variable has a clear meaning: how much the treatment changes the outcome of interest from time 1 to time t relative to the placebo group, which is essentially the same as the widely used difference-in-difference estimator.

Lastly, we used Bayesian MCMC for model estimation throughout the paper. However, as explained in earlier sections, the MGTT model can very well be estimated using EM algorithm with analytic dimension reduction. Because the general-purpose software such as M*plus* does not exploit dimension reduction for MGTT, future research would be to create customized R function or package that will implement this dimension reduction enabled EM algorithm. Moreover, the new function/package could have built-in flexibility to let users specify priors whenever needed, so that the multi-stage estimation approach can also be conducted in an EM framework.

Practical Implications

For the multi-cohort longitudinal design, we recommend using the MGTT model that not only takes advantage of integrated data from heterogeneous cohorts, but also handles nested data structure adequately. When a high correlation between main factors over time is anticipated, such as when the time gap between adjacent measurement waves is short, the MGTT model would produce much more precise latent trait and change score estimates. Given the MGTT model complexity and inherent missing data nature of pooling data from multiple studies (or cohorts), the Bayesian estimation approach is recommended. When there are not too many nuisance factors (i.e., the number of items administered repeatedly is small), a concurrent calibration is recommended, otherwise a multi-stage estimation is preferred. To implement Bayesian MCMC, researchers can use the non-informative priors as described in this paper (except the informative priors used in multi-stage estimation to transfer information between stages) and closely monitor chain convergence using the Gelman-Rubin statistic. If concurrent calibration fails to converge, the multi-stage estimation is a viable alternative, although its implementation may take a little more effort especially when there are more than two cohorts. Researchers need to ensure the transfer of information between stages is correct, and there are no viable shortcuts to careful and laborious attention to detail.

References

Albert, J. H. (1992). Bayesian estimation of normal ogive item response curves using Gibbs sampling. *Journal of Educational Statistics, 17*(3), 251-269.

Asparouhov, T., & Muthén, B. (2021). Advances in Bayesian model fit evaluation for structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal, 28*(1), 1-14.

Bauer, D. J., & Curran, P. J. (2016). The discrepancy between measurement and modeling in longitudinal data analysis. In J. R. Harring, L. M. Stapleton, & S. N. Beretvas (Eds.), *Advances in multilevel modeling for educational research* (pp. 3–38). Charlotte, NC: Information Age Publishing.

Bauer, D. J., Howard, A. L., Baldasaro, R. E., Curran, P. J., Hussong, A. M., Chassin, L., & Zucker, R. A. (2013). A trifactor model for integrating ratings across multiple informants. *Psychological Methods, 18*(4), 475-493.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological bulletin, 107*(2), 238.

Boker SM, Neale MC, Maes HH, Spiegel M, Brick TR, Estabrook R, Bates TC, Gore RJ, Hunter MD, Pritikin JN, Zahery M, Kirkpatrick RM (2023). *OpenMx: Extended Structural Equation Modelling*. R package version 2.21.8, https://CRAN.R-project.org/package=OpenMx.

Cai, L. (2008). A Metropolis-Hastings Robbins-Monro algorithm for maximum likelihood nonlinear latent structure analysis with a comprehensive measurement model (Doctoral dissertation, The University of North Carolina at Chapel Hill).

Cai, L. (2010a). High-dimensional exploratory item factor analysis by a Metropolis–Hastings Robbins–Monro algorithm. *Psychometrika, 75*(1), 33-57.

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

Cai, L., Choi, K., Hansen, M., & Harrell, L. (2016). Item response theory. *Annual Review of Statistics and Its Application, 3*, 297-321.

Cai, L., & Houts, C. R. (2021). Longitudinal analysis of patient-reported outcomes in clinical trials: Applications of multilevel and multidimensional item response theory. *Psychometrika, 86*(3), 754-777.

Cai, L., Yang, J. S., & Hansen, M. (2011). Generalized full-information item bifactor analysis. *Psychological Methods, 16*(3), 221.

Caprara, G. V., Vecchione, M., Alessandri, G., Gerbino, M., & Barbaranelli, C. (2011). The contribution of personality traits and self-efficacy beliefs to academic achievement: A longitudinal study. *British Journal of Educational Psychology, 81*(1), 78-96.

Chandler, R. K., Kahana, S. Y., Fletcher, B., Jones, D., Finger, M. S., Aklin, W. M., . . . Webb, C. (2015). Data collection and harmonization in HIV research: the seek, test, treat, and retain initiative at the National Institute on Drug Abuse. *American journal of public health, 105*(12), 2416-2422.

Choi, S. E., Mukherjee, S., Gibbons, L. E., Sanders, R. E., Jones, R. N., Tommet, D., . . . Lamar, M. (2020). Development and validation of language and visuospatial composite scores in ADNI. *Alzheimer's & Dementia: Translational Research & Clinical Interventions, 6*(1), e12072.

Crane, P. K., Carle, A., Gibbons, L. E., Insel, P., Mackin, R. S., Gross, A., . . . Harvey, D. (2012). Development and assessment of a composite score for memory in the Alzheimer's Disease Neuroimaging Initiative (ADNI). *Brain imaging and behavior, 6*(4), 502-516.

Crane, P. K., Choi, S. E., Gibbons, L. E., Mukherjee, S., Zhu, R., Scollard, P., . . . Mez, J. (2021). Cognitive assessments in ADNI: Lessons learned from the ADNI psychometrics project. *Alzheimer's & Dementia, 17*, e056474.

Curran, P. J., & Hussong, A. M. (2009). Integrative data analysis: the simultaneous analysis of multiple data sets. *Psychological Methods, 14*(2), 81.

Curran, P. J., Obeidat, K., & Losardo, D. (2010). Twelve frequently asked questions about growth curve modeling. *Journal of cognition and development, 11*(2), 121-136.

Edwards, M. C. (2010). A Markov chain Monte Carlo approach to confirmatory item factor analysis. *Psychometrika, 75*(3), 474-497.
Edwards, M. C., & Wirth, R. J. (2009). Measurement and the study of change. *Research in Human Development, 6*(2-3), 74–96.
Embretson, S. E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika, 56*(3), 495-515.
Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods, 9*(4), 466.
Fox, J.-P. (2010). *Bayesian item response modeling: theory and applications.* New York: Springer.
Fried, E. I., & Flake, J. K. (2018). Measurement matters. *APS Observer*. Retrieved from https://www.psychologicalscience.org/observer/measurement-matters
Garland, A. F., Kruse, M., & Aarons, G. A. (2003). Clinicians and outcome measurement: what's the use? *The journal of behavioral health services & research, 30*(4), 393-405.
Gelman, A., & Rubin, D. B. (1992). Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science, 7*(4), 457–472.
Gibbons, L. E., Carle, A. C., Mackin, R. S., Harvey, D., Mukherjee, S., Insel, P., . . . Crane, P. K. (2012). A composite score for executive functioning, validated in Alzheimer's Disease Neuroimaging Initiative (ADNI) participants with baseline mild cognitive impairment. *Brain imaging and behavior, 6*(4), 517-527.
Gibbons, R. D., Bock, R. D., Hedeker, D., Weiss, D. J., Segawa, E., Bhaumik, D. K., Kupfer, D.J., Frank, E., Grochocinski, V.J., & Stover, A. (2007). Full-information item bifactor analysis of graded response data. *Applied psychological measurement, 31*(1), 4-19.
Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika, 57*(3), 423-436.
Grimm, K. J., Kuhl, A. P., & Zhang, Z. (2013). Measurement models, estimation, and the study of change. *Structural Equation Modeling: A Multidisciplinary Journal, 20*(3), 504-517.
Hancock, G. R., Kuo, W.-L., & Lawrence, F. R. (2001). An illustration of second-order latent growth models. *Structural Equation Modeling, 8*(3), 470-489.
Hansen, M. (2013). *Hierarchical item response models for cognitive diagnosis.* UCLA,
Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. *Test validity*, 129-145.
Hsieh, C.-A., von Eye, A. A., & Maier, K. S. (2010). Using a multivariate multilevel polytomous item response theory model to study parallel processes of change: The dynamic association between adolescents' social isolation and engagement with delinquent peers in the National Youth Survey. *Multivariate behavioral research, 45*(3), 508-552.
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.
Huh, D., Mun, E. Y., Larimer, M. E., White, H. R., Ray, A. E., Rhew, I. C., . . . Atkins, D. C. (2015). Brief motivational interventions for college student drinking may not be as powerful as we think: An individual participant-level data meta-analysis. *Alcoholism: Clinical and Experimental Research, 39*(5), 919-931.
Isiordia, M., & Ferrer, E. (2018). Curve of factors model: A latent growth modeling approach for educational research. *Educational and Psychological Measurement, 78*(2), 203-231.

Jacobson, N. S., Follette, W. C., & Revenstorf, D. (1984). Psychotherapy outcome research: Methods for reporting variability and evaluating clinical significance. *Behavior therapy, 15*(4), 336-352.

Jiang, S., Wang, C., & Weiss, D. J. (2016). Sample size requirements for estimation of item parameters in the multidimensional graded response model. *Frontiers in psychology, 7*, 109.

Jutten, R. J., Grandoit, E., Foldi, N. S., Sikkes, S. A., Jones, R. N., Choi, S. E., . . . Tommet, D. (2020). Lower practice effects as a marker of cognitive performance and dementia risk: a literature review. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring, 12*(1), e12055.

Kamata, A., & Bauer, D. J. (2008). A note on the relation between factor analytic and item response theory models. *Structural Equation Modeling, 15*(1), 136-153.

Kemp, C. G., Lipira, L. L., David, H., Nevin, P. E., Turan, J., Simoni, J. M., . . . Andrasik, M. (2019). HIV stigma and viral load among African-American women receiving treatment for HIV: A longitudinal analysis. *AIDS (London, England), 33*(9), 1511.

Kim, S., & Kolen, M. (2016). *Multiple group IRT fixed-parameter estimation for maintainingan established ability scale* (Center for Advanced Studies in Measurement and Assessment Report No. 49). Retrieved from https://education.uiowa.edu/sites/education.uiowa.edu/files/documents/centers/casma/casma-research-report-49.pdf

Kuhfeld, M., & Soland, J. (2022). Avoiding bias from sum scores in growth estimates: an examination of IRT-based approaches to scoring longitudinal survey responses. *Psychological Methods*, 27, 234-260.

Levy, R., Mislevy, R. J., & Sinharay, S. (2009). Posterior predictive model checking for multidimensionality in item response theory. *Applied psychological measurement, 33*(7), 519-537.

McArdle, J. J. (2009). Latent variable modeling of differences and changes with longitudinal data. *Annual review of psychology, 60*, 577-605.

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.

Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement, 29*(2), 133-161.

Muthén, B., & Asparouhov, T. (2012). Bayesian structural equation modeling: a more flexible representation of substantive theory. *Psychological methods, 17*(3), 313.

Muthén, B. O., du Toit, S. H. C., & Spisic, D. (1997). Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes. Unpublished manuscript. Available at https://www.statmodel.com/down load/Article_075.pdf

Muthén, L. K., & Muthén, B. O. (1998-2017). Mplus user's guide (8th ed.). Los Angeles, CA: Muthén & Muthén

Nance, R. M., Delaney, J. C., Golin, C. E., Wechsberg, W. M., Cunningham, C., Altice, F., . . . Gordon, M. S. (2017). Co-calibration of two self-reported measures of adherence to antiretroviral therapy. *AIDS care, 29*(4), 464-468.

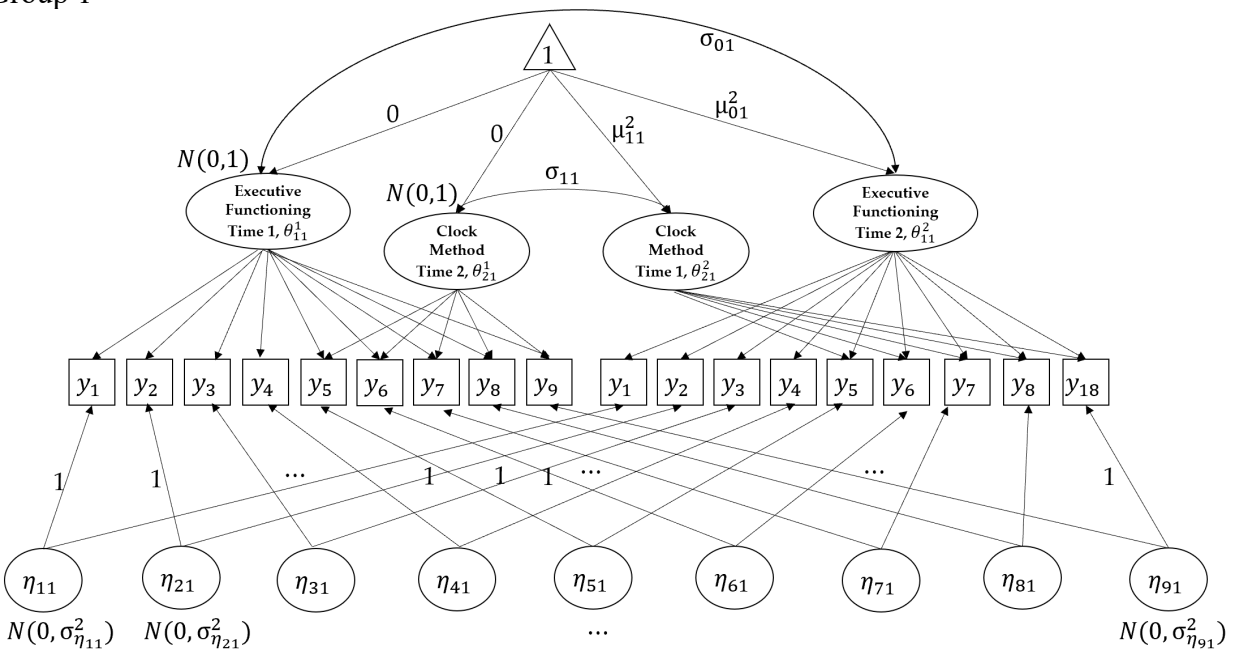
Paek, I., Li, Z., & Park, H.-J. (2016). Specifying Ability Growth Models Using a Multidimensional Item Response Model for Repeated Measures Categorical Ordinal Item Response Data. *Multivariate behavioral research, 51*(4), 569-580.
Paek, I., Park, H.-J., Cai, L., & Chi, E. (2014). A comparison of three IRT approaches to examinee ability change modeling in a single-group anchor test design. *Educational and Psychological Measurement, 74*(4), 659-676.
Patz, R. J., & Junker, B. W. (1999). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics, 24*(4), 342-366.
Penfield, R. D., & Camilli, G. (2006). 5 Differential Item Functioning and Item Bias. *Handbook of statistics, 26*, 125-167.
Reckase, M. (2009). *Multidimensional item response theory* (Vol. 150): Springer.
Rijmen, F. (2010). Formal relations and an empirical comparison among the bi-factor, the testlet, and a second-order multidimensional IRT model. *Journal of Educational Measurement, 47*(3), 361-372.
Rijmen, F., Vansteelandt, K., & De Boeck, P. (2008). Latent class models for diary method data: Parameter estimation by local computations. *Psychometrika, 73*(2), 167.
Robert, C.P., Casella, G. (1999). The Metropolis—Hastings Algorithm. In: Monte Carlo Statistical Methods. *Springer Texts in Statistics. Springer*, New York, NY.
Rosseel Y (2012). "lavaan: An R Package for Structural Equation Modeling." *Journal of Statistical Software*, **48**(2), 1–36. doi:10.18637/jss.v048.i02.
Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika monograph supplement*.
Schober, P., & Vetter, T. R. (2018). Repeated measures designs and analysis of longitudinal data: If at first you do not succeed—try, try again. *Anesthesia and analgesia, 127*(2), 569.
Silins, E., Fergusson, D. M., Patton, G. C., Horwood, L. J., Olsson, C. A., Hutchinson, D. M., . . . Coffey, C. (2015). Adolescent substance use and educational attainment: an integrative data analysis comparing cannabis and alcohol from three Australasian cohorts. *Drug and alcohol dependence, 156*, 90-96.
Soland, J., Jensen, N., Keys, T. D., Bi, S. Z., & Wolk, E. (2019). Are test and academic disengagement related? Implications for measurement and practice. *Educational Assessment, 24*(2), 119-134.
Soland, J., & Kuhfeld, M. (2019). Do students rapidly guess repeatedly over time? A longitudinal analysis of student test disengagement, background, and attitudes. *Educational Assessment, 24*(4), 327-342.
Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 64*(4), 583-639.
Su, Y., Reedy, J., & Carroll, R. J. (2018). Clustering in general measurement error models. *Statistica Sinica, 28*(4), 2337.
Takane, Y., & De Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52*(3), 393-408.
Vonk, J. M., Gross, A. L., Zammit, A. R., Bertola, L., Avila, J. F., Jutten, R. J., . . . O'Connell, M. E. (2022). Cross-national harmonization of cognitive measures across HRS HCAP (USA) and LASI-DAD (India). *PloS one, 17*(2), e0264166.
Wang, C., Chang, H. H., & Douglas, J. A. (2013). The linear transformation model with frailties

for the analysis of item response times. *British Journal of Mathematical and Statistical Psychology, 66*(1), 144-168.
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, C., & Nydick, S. W. (2020). On longitudinal item response theory models: A didactic. *Journal of Educational and Behavioral Statistics, 45*(3), 339-368.
Wang, C., & Nydick, S. W. (2015). Comparing two algorithms for calibrating the restricted non-compensatory multidimensional IRT model. *Applied psychological measurement, 39*(2), 119-134.
Wang, C., Su, S., & Weiss, D. J. (2018). Robustness of Parameter Estimation to Assumptions of Normality in the Multidimensional Graded Response Model. *Multivariate behavioral research, 53*(3), 403-418.
Wang, C., & Weiss, D. J. (2017). Multivariate Hypothesis Testing Methods for Evaluating Significant Individual Change. *Applied psychological measurement*, 0146621617726787.
Wang, C., Weiss, D. J., & Suen, K. Y. (2020). Hypothesis Testing Methods for Multivariate Multi-Occasion Intra-Individual Change. *Multivariate behavioral research*, 1-17.
Wang, C., Xu, G., & Zhang, X. (2019). Correction for Item Response Theory Latent Trait Measurement Error in Linear Mixed Effects Models. *Psychometrika*, 1-28.
Widaman, K. F., Ferrer, E., & Conger, R. D. (2010). Factorial invariance within longitudinal structural equation models: Measuring the same construct across time. *Child development perspectives, 4*(1), 10-18.
Wirth, R., & Edwards, M. C. (2007). Item factor analysis: current approaches and future directions. *Psychological Methods, 12*(1), 58.
Witkiewitz, K., Hallgren, K. A., O'Sickey, A. J., Roos, C. R., & Maisto, S. A. (2016). Reproducibility and differential item functioning of the alcohol dependence syndrome construct across four alcohol treatment studies: An integrative data analysis. *Drug and alcohol dependence, 158*, 86-93.
Woods, C. M. (2009). Evaluation of MIMIC-model methods for DIF testing with comparison to two-group analysis. *Multivariate behavioral research, 44*(1), 1-27.
Woods, C. M., & Grimm, K. J. (2011). Testing for nonuniform differential item functioning with multiple indicator multiple cause models. *Applied psychological measurement, 35*(5), 339-361.
Xie, X., & Meng, X.-L. (2017). Dissecting multiple imputation from a multi-phase inference perspective: what happens when God's, imputer's and analyst's models are uncongenial? *Statistica Sinica*, 1485-1545.
Zhang, X., Tao, J., Wang, C., & Shi, N. Z. (2019). Bayesian model selection methods for multilevel IRT models: A comparison of five DIC-Based indices. *Journal of Educational Measurement, 56*(1), 3-27.

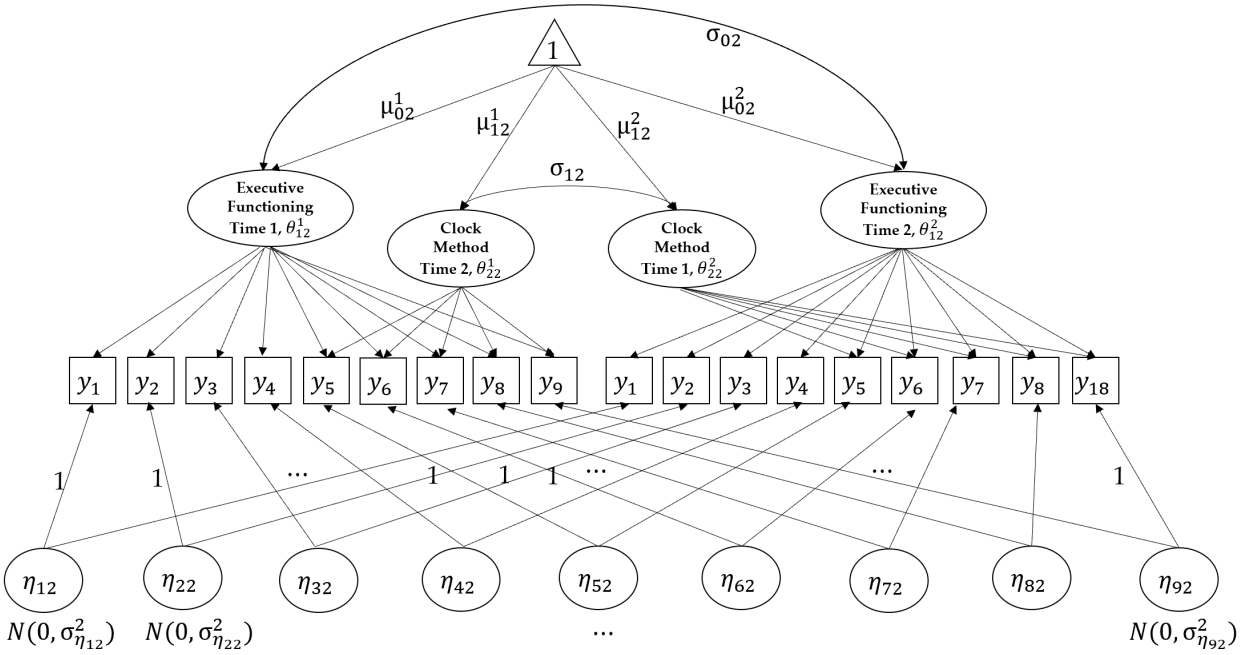
Figure 1

An illustration of MGTT model for the ADNI-Executive Functioning (EF) scale

Group 1



Group 2



Note. Notation wise, the superscript denotes time point (1 or 2), and subscripts denote factors and groups. That is, θ_{11}^{1} and θ_{11}^{2} denote the main factor (i.e., EF) for group 1 at time 1 (baseline) and

time 2 (6-month follow-up) respectively, whereas θ_{12}^{1} and θ_{12}^{2} denote the main factor for group 2 at both time points. Similarly, θ_{21}^{1} and θ_{21}^{2} denote the nuisance factor (i.e., clock method) for group 1 at time 1 (baseline) and time 2 (6-month follow-up) respectively. The second subscript is used to denote group membership (1 or 2). For instance, μ_{01}^{2} and μ_{11}^{2} denote the mean of EF and clock method factors in group 1 at time 2, whereas μ_{02}^{1} and μ_{12}^{1} denote the mean of EF and clock method factors in group 2 at time 1. As shown, the means and variances of EF and clock method factors at time 1 group 1 are fixed at 0 and 1 respectively, whereas their means and variances in all remaining time points and groups are freely estimated. The covariances among the same factor over time are freely estimated and differ per group. Similarly, the variances of the other nuisance factors (i.e., η's) are freely estimated per group.

Figure 2

An illustration of the multi-stage estimation approach with a three-cohort design

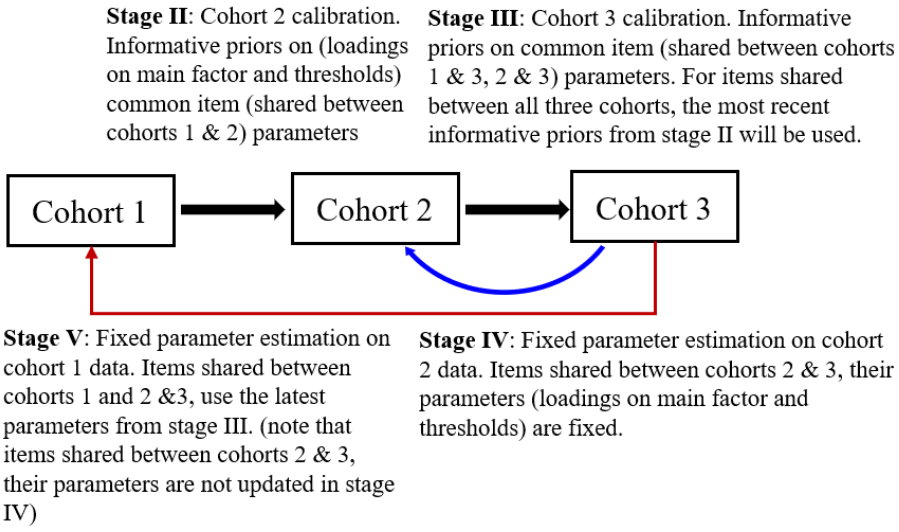


Figure 3

Discrimination and threshold parameter estimates of 7 common items (see Table 8) from ADNI language domain in stage I and stage II

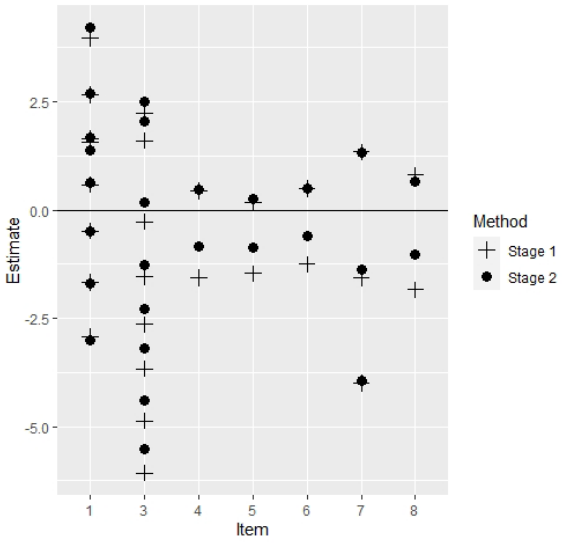


Table 1

True item parameters for simulation studies

Cohort 1			Cohort 2			Cohort 3		Item	a	b			γ
Baseline	Follow I	Follow II	Baseline	Follow I	Follow II	Baseline	Follow I						
Item Design I													
√	√		√	√		√	√	1	0.912	-2.469	-0.876	1.760	0.955
√	√		√	√		√	√	2	1.019	-2.256	-0.588	2.555	0.905
√	√		√	√		√	√	3	1.219	-1.854	-0.647	2.869	1.199
√		√	√		√	√		4	1.555	-1.184	0.374	1.424	(0.703)
√		√	√		√	√		5	0.849	-2.597	-0.232	2.303	(1.116)
√		√	√		√	√		6	1.545	-1.203	0.540	1.251	(1.131)
	√			√			√	7	1.592	-1.111	-0.005	1.534	
	√			√			√	8	1.308	-1.678	0.435	1.772	
	√			√			√	9	1.276	-1.742	0.984	1.027	
		√			√			10	0.709	-2.876	-0.576	1.372	
		√			√			11	0.853	-2.588	0.303	2.655	
		√			√			12	0.824	-2.647	-0.749	2.337	
Item Design II													
√	√		√	√				1	0.912	-2.469	-0.876	1.760	0.955
√	√		√	√				2	1.019	-2.256	-0.588	2.555	0.905
√	√		√	√				3	1.219	-1.854	-0.647	2.869	1.199
√	√					√	√	4	1.555	-1.184	0.374	1.424	0.703
√	√					√	√	5	0.849	-2.597	-0.232	2.303	1.116
√	√					√	√	6	1.545	-1.203	0.540	1.251	1.131
			√	√		√	√	7	1.592	-1.111	-0.005	1.534	1.459
			√	√		√	√	8	1.308	-1.678	0.435	1.772	1.017
			√	√		√	√	9	1.276	-1.742	0.984	1.027	1.194

Note. Simulation study I used cohort I and cohort II baseline and follow up I data, for both item design I and design II. Note that under item design I, only the first nine items were relevant. The three γ parameters in the parenthesis were not used. Simulation study II used cohort I, II, and III baseline and follow up I data, for both item design I and design II. Again, under item design I, only the first nine items were relevant. The three γ parameters in the parenthesis were not used. Simulation study III used cohort I and cohort II baseline, follow up I and II data, for only item design I. All 12 items were used and all six γ parameters were used.

Table 2

True generating person parameters for the simulation study I

Person Design	Cohort I		Cohort II		Remarks
I	$\theta \sim mvn([0, 0.2]', \mathbf{\Sigma})$, where $\mathbf{\Sigma} = [1, 0.4; 0.4,1]$ (Kuhfeld & Soland, 2022)		$\theta \sim mvn([0.1, 0.3]', \mathbf{\Sigma})$, where $\mathbf{\Sigma} = [1, 0.4; 0.4,1]$		*Moderate* correlation between baseline and follow up traits, *small* mean change of θ (0.2) over time, and *little* impact between two cohorts (mean difference =0.1, no difference of $\mathbf{\Sigma}$)
II	$\theta \sim mvn([0, 0.05]', \mathbf{\Sigma})$, where $\mathbf{\Sigma} = [1, 1.04; 1.04,1.37]$ (again following ADNI language data analysis results)		$\theta \sim mvn([0.5, 0.55]', \mathbf{\Sigma})$, where $\mathbf{\Sigma} = [1.12, 1.15; 1.15,1.3]$		*High* correlation between baseline and follow up traits, small mean change of θ (0.05) over time, and *moderate* impact between two cohorts (mean difference =0.5, a little difference of $\mathbf{\Sigma}$)
III	$\theta_1 \sim N(0,1)$	$\theta_2 = \theta_1 - 0.75 + r$ for the first 20% sample, $\theta_2 = \theta_1 + 0.75 + r$ for the last 20% sample, and $\theta_2 = \theta_1 + r$ for the middle 60% sample, where $r \sim U(-0.2,0.2)$	$\theta_1 \sim N(0.1,1)$	Same as design III & IV	*High* correlation between baseline and follow up traits, low impact.
IV	$\theta_1 \sim N(0,1)$		$\theta_1 \sim N(0.5, 1)$		*High* correlation between baseline and follow up traits, moderate impact.

Table 3

Latent trait and latent change score recovery for item design I in simulation study I

Person Design			Baseline			Follow up			Change		
			Single	Concurrent	3-stage	Single	Concurrent	3-stage	Single	Concurrent	3-stage
I	Cohort 1	Bias	0.015	0.001	-0.010	-0.016	-0.019	-0.031	-0.031	-0.020	-0.021
		RMSE	0.435	0.424	0.425	0.430	0.421	0.423	0.562	0.544	0.544
		Mean ABS	0.344	0.336	0.337	0.340	0.333	0.334	0.447	0.433	0.433
		Mean SE	0.414	0.413	0.405	0.406	0.414	0.390	0.582	0.586	0.564
	Cohort 2	Bias	0.006	0.006	-0.009	-0.021	-0.020	-0.048	-0.027	-0.027	-0.039
		RMSE	0.438	0.430	0.431	0.431	0.422	0.425	0.566	0.548	0.550
		Mean ABS	0.345	0.339	0.340	0.341	0.334	0.336	0.449	0.435	0.437
		Mean SE	0.412	0.413	0.406	0.403	0.418	0.395	0.578	0.589	0.568
II	Cohort 1	Bias	0.008	-0.006	-0.012	-0.022	-0.035	-0.036	-0.030	-0.03	-0.023

		RMSE	0.435	0.380	0.380	0.456	0.420	0.415	0.548	0.407	0.404
		Mean ABS	0.343	0.299	0.299	0.359	0.331	0.326	0.435	0.324	0.321
		Mean SE	0.416	0.369	0.361	0.432	0.420	0.386	0.602	0.560	0.529
	Cohort 2	Bias	0.008	0.009	-0.015	-0.014	-0.012	-0.047	-0.006	-0.021	-0.032
		RMSE	0.453	0.380	0.378	0.464	0.402	0.400	0.464	0.306	0.307
		Mean ABS	0.356	0.299	0.297	0.364	0.317	0.314	0.429	0.244	0.244
		Mean SE	0.429	0.374	0.360	0.434	0.399	0.375	0.611	0.547	0.521
III	Cohort 1	Bias	-0.001	-0.009	-0.029	-0.029	0.006	-0.050	-0.028	0.015	-0.023
		RMSE	0.435	0.383	0.384	0.450	0.405	0.410	0.544	0.404	0.406
		Mean ABS	0.344	0.303	0.304	0.355	0.321	0.325	0.433	0.326	0.329
		Mean SE	0.415	0.368	0.364	0.428	0.397	0.382	0.598	0.542	0.528
	Cohort 2	Bias	-0.016	-0.016	-0.032	-0.021	0.009	-0.063	-0.005	0.025	-0.031
		RMSE	0.438	0.386	0.388	0.450	0.410	0.414	0.536	0.406	0.406
		Mean ABS	0.346	0.305	0.307	0.355	0.324	0.328	0.426	0.327	0.327
		Mean SE	0.414	0.365	0.369	0.427	0.396	0.388	0.596	0.539	0.536
IV	Cohort 1	Bias	0.007	0.007	-0.018	-0.020	0.027	-0.040	-0.027	0.020	-0.022
		RMSE	0.435	0.383	0.384	0.450	0.409	0.409	0.545	0.408	0.409
		Mean ABS	0.344	0.303	0.303	0.354	0.324	0.324	0.434	0.328	0.330
		Mean SE	0.413	0.368	0.363	0.426	0.398	0.381	0.596	0.543	0.527
	Cohort 2	Bias	-0.003	0.005	-0.024	-0.012	0.036	-0.055	-0.009	0.031	-0.031
		RMSE	0.441	0.389	0.391	0.463	0.421	0.421	0.543	0.411	0.411
		Mean ABS	0.348	0.307	0.308	0.365	0.332	0.332	0.432	0.332	0.332
		Mean SE	0.420	0.372	0.371	0.433	0.403	0.394	0.605	0.550	0.542

Table 4

Latent trait and latent change recovery for item design II in simulation study I

Person Design		**Cohort I**						**Cohort II**					
		Baseline		Follow up		Change		Baseline		Follow up		Change	
		Single	3-stage	Single	3-stage	Single	3-stage	Single	3-stage	Single	3-stage	Single	3-stage
I	Bias	-0.001	-0.006	0.012	-0.001	0.012	0.004	-0.002	-0.030	-0.019	-0.040	-0.016	-0.010
	RMSE	0.493	0.490	0.499	0.496	0.565	0.558	0.522	0.522	0.522	0.522	0.570	0.565
	Mean ABS	0.392	0.389	0.396	0.393	0.448	0.442	0.416	0.415	0.417	0.416	0.453	0.448

	Mean SE	0.483	0.480	0.476	0.477	0.680	0.678	0.506	0.479	0.504	0.485	0.716	0.682
II	Bias	-0.005	-0.008	0.010	0.001	0.014	0.009	-0.004	-0.029	-0.018	-0.031	-0.014	-0.002
	RMSE	0.492	0.451	0.526	0.499	0.525	0.406	0.538	0.482	0.550	0.510	0.519	0.306
	Mean ABS	0.391	0.360	0.417	0.396	0.417	0.323	0.427	0.384	0.438	0.406	0.412	0.244
	Mean SE	0.483	0.447	0.499	0.504	0.696	0.674	0.523	0.461	0.531	0.494	0.747	0.676
III	Bias	-0.006	-0.013	0.000	-0.021	0.006	-0.008	-0.008	-0.026	-0.013	-0.021	-0.005	0.005
	RMSE	0.495	0.456	0.522	0.494	0.520	0.407	0.523	0.489	0.546	0.525	0.514	0.410
	Mean ABS	0.393	0.362	0.413	0.393	0.414	0.328	0.417	0.390	0.436	0.420	0.409	0.331
	Mean SE	0.484	0.450	0.497	0.491	0.696	0.667	0.503	0.455	0.521	0.496	0.725	0.674
IV	Bias	-0.010	-0.009	0.009	-0.010	0.020	-0.000	0.005	-0.032	-0.006	-0.030	-0.010	0.002
	RMSE	0.493	0.454	0.518	0.492	0.522	0.407	0.523	0.488	0.550	0.526	0.526	0.410
	Mean ABS	0.391	0.361	0.412	0.392	0.415	0.328	0.416	0.389	0.437	0.419	0.418	0.332
	Mean SE	0.485	0.451	0.496	0.496	0.695	0.671	0.518	0.460	0.534	0.502	0.745	0.684

Table 5

Latent trait and latent change recovery for item design I in simulation study II

Person			**Baseline**			**Follow up**			**Change**		
Design			Single	Concurrent	Multi-stage	Single	Concurrent	Multi-stage	Single	Concurrent	Multi-stage
I	Cohort 1	Bias	0.008	0.037	-0.011	-0.014	0.016	-0.031	-0.023	-0.020	-0.020
		RMSE	0.436	0.428	0.426	0.433	0.422	0.425	0.565	0.547	0.548
		Mean ABS	0.342	0.337	0.336	0.343	0.334	0.336	0.449	0.434	0.435
		Mean SE	0.416	0.407	0.409	0.411	0.402	0.396	0.587	0.574	0.571
	Cohort 2	Bias	0.005	0.041	-0.016	-0.006	0.019	-0.032	-0.012	-0.022	-0.016
		RMSE	0.438	0.431	0.431	0.434	0.424	0.427	0.561	0.548	0.550
		Mean ABS	0.345	0.340	0.339	0.343	0.335	0.337	0.445	0.436	0.437
		Mean SE	0.415	0.405	0.405	0.41	0.405	0.396	0.585	0.574	0.568
	Cohort 3	Bias	0.004	0.043	-0.011	-0.010	0.019	-0.048	-0.014	-0.024	-0.036
		RMSE	0.453	0.429	0.427	0.434	0.422	0.426	0.464	0.306	0.307
		Mean ABS	0.344	0.339	0.337	0.342	0.333	0.336	0.559	0.547	0.549
		Mean SE	0.416	0.403	0.41	0.411	0.404	0.397	0.586	0.573	0.572
II	Cohort 1	Bias	0.011	0.040	-0.016	-0.027	0.022	-0.042	-0.038	-0.018	-0.026
		RMSE	0.433	0.379	0.377	0.456	0.421	0.414	0.545	0.408	0.401
		Mean ABS	0.341	0.299	0.298	0.359	0.332	0.327	0.433	0.324	0.318
		Mean SE	0.415	0.37	0.365	0.432	0.425	0.393	0.602	0.565	0.537

	Cohort 2	Bias	-0.001	0.067	-0.024	-0.030	0.053	-0.052	-0.029	-0.014	-0.028
		RMSE	0.457	0.388	0.380	0.467	0.414	0.405	0.542	0.311	0.306
		Mean ABS	0.359	0.305	0.298	0.366	0.325	0.317	0.431	0.248	0.244
		Mean SE	0.427	0.379	0.355	0.435	0.41	0.374	0.611	0.559	0.517
	Cohort 3	Bias	0.001	0.059	-0.014	-0.021	0.049	-0.056	-0.022	-0.011	-0.041
		RMSE	0.440	0.360	0.354	0.455	0.398	0.394	0.526	0.272	0.271
		Mean ABS	0.348	0.285	0.279	0.357	0.314	0.310	0.418	0.217	0.216
		Mean SE	0.414	0.357	0.343	0.427	0.399	0.371	0.596	0.535	0.506

Table 6

Latent trait and latent change recovery for item design II in simulation study II

Person Design			**Cohort I**				**Cohort II**				**Cohort III**			
			Bias	RMSE	Mean ABS	Mean SE	Bias	RMSE	Mean ABS	Mean SE	Bias	RMSE	Mean ABS	Mean SE
I	Baseline	Single	0.001	0.496	0.395	0.481	-0.008	0.520	0.414	0.509	-0.016	0.473	0.375	0.451
		M-S	0.002	0.492	0.392	0.472	-0.019	0.516	0.410	0.497	-0.027	0.470	0.371	0.431
	Follow up	Single	0.003	0.499	0.396	0.473	-0.024	0.525	0.418	0.508	-0.018	0.479	0.379	0.451
		M-S	0.000	0.496	0.394	0.474	-0.025	0.521	0.415	0.501	-0.031	0.475	0.375	0.435
II	Baseline	Single	-0.004	0.495	0.393	0.483	-0.007	0.539	0.428	0.528	-0.012	0.473	0.375	0.454
		M-S	0.001	0.453	0.360	0.446	0.002	0.485	0.385	0.495	-0.011	0.415	0.328	0.405
	Follow up	Single	0.005	0.524	0.415	0.499	-0.023	0.553	0.440	0.534	-0.014	0.504	0.397	0.476
		M-S	0.008	0.501	0.397	0.503	0.000	0.514	0.409	0.523	-0.009	0.462	0.364	0.448

Note. Here "M-S" denotes multi-stage estimation.

Table 7

Latent trait and latent change recovery for simulation study III

Person design			Baseline			Follow up 1			Follow up 2		
			concurrent	three-stage	single	concurrent	three-stage	single	concurrent	three-stage	single
I	Cohort 1	Bias	0.033	-0.003	0.004	0.039	-0.004	-0.011	0.042	-0.011	-0.008
		RMSE	0.471	0.469	0.488	0.417	0.415	0.426	0.484	0.482	0.498
		Mean ABS	0.373	0.372	0.388	0.331	0.330	0.338	0.385	0.383	0.395
		Mean SE	0.459	0.453	0.474	0.392	0.387	0.404	0.471	0.458	0.481
			Baseline – Follow up 1			Follow up 1– Follow up 2			Baseline – Follow up 2		
			concurrent	three-stage	single	concurrent	three-stage	single	concurrent	three-stage	single
		Bias	0.005	-0.001	-0.016	0.003	-0.007	0.003	0.009	-0.009	-0.013
		RMSE	0.564	0.564	0.588	0.598	0.599	0.619	0.585	0.585	0.605
		Mean ABS	0.449	0.450	0.468	0.475	0.476	0.492	0.465	0.465	0.482
		Mean SE	0.605	0.597	0.625	0.614	0.601	0.629	0.659	0.645	0.677
	Cohort 2		Baseline			Follow up 1			Follow up 2		
			concurrent	three-stage	single	concurrent	three-stage	single	concurrent	three-stage	single
		Bias	0.044	0.008	-0.005	0.036	-0.005	-0.004	0.039	-0.013	-0.008
		RMSE	0.477	0.476	0.494	0.416	0.414	0.424	0.486	0.486	0.500
		Mean ABS	0.379	0.378	0.393	0.330	0.328	0.336	0.387	0.387	0.398
		Mean SE	0.461	0.442	0.474	0.393	0.389	0.399	0.473	0.45	0.475
			Baseline – Follow up 1			Follow up 1– Follow up 2			Baseline – Follow up 2		
			concurrent	three-stage	single	concurrent	three-stage	single	concurrent	three-stage	single
		Bias	0.044	0.008	-0.005	0.036	-0.005	-0.004	0.039	-0.013	-0.008
		RMSE	0.477	0.476	0.494	0.416	0.414	0.424	0.486	0.486	0.500
		Mean ABS	0.379	0.378	0.393	0.330	0.328	0.336	0.387	0.387	0.398
		Mean SE	0.461	0.442	0.474	0.393	0.389	0.399	0.473	0.45	0.475
II	Cohort 1		Baseline			Follow up 1			Follow up 2		
			concurrent	three-stage	single	concurrent	three-stage	single	concurrent	three-stage	single
		Bias	0.034	-0.002	0.006	0.042	-0.005	-0.010	0.052	-0.006	-0.011
		RMSE	0.415	0.412	0.488	0.381	0.374	0.425	0.421	0.414	0.492
		Mean ABS	0.331	0.328	0.388	0.302	0.297	0.336	0.335	0.329	0.391
		Mean SE	0.407	0.403	0.474	0.376	0.357	0.407	0.426	0.401	0.479
			Baseline – Follow up 1			Follow up 1– Follow up 2			Baseline – Follow up 2		
			concurrent	three-stage	single	concurrent	three-stage	single	concurrent	three-stage	single

		Bias	0.008	-0.003	-0.016	0.009	-0.001	0.000	0.018	-0.004	-0.017
		RMSE	0.425	0.424	0.552	0.439	0.438	0.580	0.431	0.430	0.550
		Mean ABS	0.339	0.338	0.440	0.350	0.349	0.461	0.344	0.342	0.437
		Mean SE	0.555	0.539	0.627	0.569	0.538	0.63	0.59	0.57	0.676
	Cohort 2		Baseline			Follow up 1			Follow up 2		
			concurrent	three-stage	single	concurrent	three-stage	single	concurrent	three-stage	single
		Bias	0.066	0.004	-0.010	0.061	0.006	-0.009	0.056	-0.001	-0.006
		RMSE	0.427	0.419	0.499	0.387	0.378	0.428	0.423	0.418	0.499
		Mean ABS	0.340	0.333	0.395	0.307	0.299	0.338	0.338	0.333	0.396
		Mean SE	0.425	0.402	0.479	0.379	0.367	0.406	0.426	0.394	0.473
			Baseline – Follow up 1			Follow up 1– Follow up 2			Baseline – Follow up 2		
			concurrent	three-stage	single	concurrent	three-stage	single	concurrent	three-stage	single
		Bias	-0.004	0.002	0.001	-0.005	-0.007	0.003	-0.010	-0.005	0.004
		RMSE	0.428	0.428	0.555	0.442	0.442	0.584	0.435	0.433	0.556
		Mean ABS	0.340	0.341	0.441	0.352	0.352	0.464	0.347	0.346	0.442
		Mean SE	0.57	0.545	0.629	0.572	0.539	0.624	0.603	0.564	0.674

Table 8

Cognitive battery design for ADNI 1 and ADNI 2/GO

Domain	Item		ADNI 1	ADNI 2/GO
Language	Neuropsychological Battery	Category Fluency-Animal	✓	✓
		Category Fluency-Vegetable	✓	
		Boston Naming (Total)	✓	✓
	MMSE	MMSE Repeating a sentence	✓	✓
		MMSE Following a Series of Instructions (hand)	✓	✓
	ADAS-Cognitive Behavior	ADAS-Cog Following Commands	✓	✓
		ADAS-Cog Object Naming	✓	✓
		ADAS-Cog Ideational Practice	✓	✓
	MoCA	Animal Naming –Rhino		✓
		Sentence Repetition Task 1		✓
		Sentence Repetition Task 2		✓
		Letter F Fluency		✓

EF	Neuropsychological Battery	Clock copy – Circle	✓	✓
		Clock copy – Symmetry	✓	✓
		Clock copy – Numbers	✓	✓
		Clock copy – Hands	✓	✓
		Clock copy – Time	✓	✓
		WAIS-R Digit Symbol	✓	
		Digit Span Backwards	✓	
		Trails A	✓	✓
		Trails B	✓	✓

Note: The same set of items are used in both baseline and follow up in the respective cohorts.

Table 9

Absolute and comparative model fit results from WLS and MCMC

Domain	**Data**	**Model (GRM)**	**Absolute fit indices for WLS**			**Chi-Square Test for Difference Testing**			**% of PPP-value<0.05**
			CFI	TLI	RMSEA	Value	Degrees of Freedom	P-value	Item level
Language	ADNI 1	Without correlated residuals	0.954	0.960	0.081	544.406	8	0.000	0.0625
		With correlated residuals	0.991	0.992	0.037				0
	ADNI 2 /GO	Without correlated residuals	0.884	0.891	0.077	931.981	11	0.000	0
		With correlated residuals	0.977	0.977	0.035				0
Executive Function	ADNI 1	Without correlated residuals	0.965	0.970	0.076	675.120	9	0.000	0
		With correlated residuals	0.995	0.996	0.029				0
	ADNI 2 /GO	Without correlated residuals	0.981	0.983	0.046	127.623	7	0.000	0
		With correlated residuals	0.997	0.997	0.020				0

Table 10

Descriptive statistics of counts of participants in each category of change from both domains and two methods

	Language		**Executive functioning**	
Cluster	Single	3-stage	Single	Concurrent
1 (A lot better)	55 (7%)	81 (11%)	68 (9%)	73 (10%)
2 (A little better)	204 (27%)	221 (29%)	216 (28%)	181 (24%)
3 (No change)	226 (30%)	245 (32%)	237 (31%)	240 (31%)
4 (A little worse)	198 (26%)	148 (19%)	193 (25%)	188 (25%)
5 (A lot worse)	81 (11%)	69 (9%)	50 (7%)	82 (11%)

Table 11

Hazard ratio, its confidence interval and p-value from Cox PH model

		Single			**3-stage/concurrent**		
	Cluster	HR	95% CI	p-value	HR	95% CI	p-value
Language	1 (A lot better)	0.55	[0.31, 0.97]	0.037	0.5	[0.30, 0.83]	0.007
	2 (A little better)	0.63	[0.45, 0.87]	0.005	0.79	[0.58, 1.09]	0.2
	4 (A little worse)	1.14	[0.85, 1.53]	0.4	1.98	[1.47, 2.67]	<0.001
	5 (A lot worse)	1.20	[0.81, 1.77]	0.4	1.95	[1.29, 2.93]	0.001
EF	1 (A lot better)	0.58	[0.36, 0.92]	0.022	0.51	[0.30, 0.84]	0.008
	2 (A little better)	0.71	[0.53, 0.96]	0.027	0.46	[0.32, 0.66]	<0.001
	4 (A little worse)	1.00	[0.75, 1.35]	>0.9	1.58	[1.18, 2.10]	0.002
	5 (A lot worse)	0.95	[0.58, 1.55]	0.8	2.42	[1.68, 3.48]	<0.001

Note. HR denotes hazard ratio.

Table 12

Sensitivity analyses: Hazard ratio, its confidence interval and p-value from Cox PH model

		Single (frequencies match Bayesian categories)			**3-stage/concurrent (frequencies match traditional categories)**		
	Cluster	HR	95% CI	p-value	HR	95% CI	p-value
Language	1 (A lot better)	0.49	[0.30, 0.80]	0.004	0.67	[0.37, 1.22]	0.2
	2 (A little better)	0.65	[0.48, 0.88]	0.006	0.93	[0.66, 1.31]	0.7
	4 (A little worse)	1.10	[0.81, 1.50]	0.5	2.27	[1.68, 3.07]	<0.001
	5 (A lot worse)	1.23	[0.82, 1.84]	0.3	2.34	[1.58, 3.48]	0.001
EF	1 (A lot better)	0.56	[0.35, 0.88]	0.013	0.45	[0.27, 0.77]	0.003
	2 (A little better)	0.68	[0.50, 0.93]	0.017	0.48	[0.34, 0.67]	<0.001
	4 (A little worse)	0.89	[0.66, 1.20]	0.4	1.62	[1.22, 2.14]	<0.001
	5 (A lot worse)	0.92	[0.61, 1.37]	0.7	2.52	[1.61, 3.93]	<0.001