# The Observed Dependency of Longitudinal Data

**John J. McArdle**

**Abstract** It is well known that longitudinal data can deal with different concepts than cross-sectional data (see Baltes & Nesselroade, 1979; McArdle & Nesselroade, 2014). The key is in the observed dependency—that allows us to examine individual changes. Thus, all of the individual changes that can be examined are due to the longitudinal models (see McArdle, 2008) allowing dependencies among the observed scores at various time points. It is demonstrated here that the statistical power to detect changes is an explicit function of the positive dependencies and the timing of the observations. A lot of time is spent on the move to the *latent curve model* (LCM) from the basic regression structural model and the repeated measures model (RANOVA) because the latter seems standard in the field now. This LCM is introduced in this chapter as a principle that does have power to detect many more changes than the usual regression analysis but it comes along with several (to be discussed) assumptions.

The four articles to follow in this volume are reviewed with longitudinal dependency in mind, and the highlights of each chapter are brought out. The chapter "Nonlinear Growth Curve Models" extends the LCM to handle serious forms of nonlinearity, and this is clearly prevalent in Psychology. The chapter "Stage-Sequential Growth Mixture Modeling" extends this work to include multistage models, Poisson relations, all in the context of a multiple mixture model. This is a fairly complex example. The chapter "General Growth Mixture Modeling: The Study of Developmental Pathways of Externalizing Behavior from Preschool Age to Adolescence" is a real-life example that includes LCMs for five mixture groups. The chapter "A Generalization of Nagin's Finite Mixture Model" extends the mixture models further, mainly by adding a slope component.

But what is also important in this regard is "measurement invariance" and how this can be crucial to understanding changes. Some elaboration of the early work

A contribution for a book on "Dependent Data in Social Science Research" Edited by M. Stemmler, A. von Eye and W. Wiedermann.

J.J. McArdle (✉)
Department of Psychology, University of Southern California, Los Angeles, CA, USA
e-mail: jmcardle@usc.edu

M. Stemmler et al. (eds.), *Dependent Data in Social Sciences Research*, Springer Proceedings in Mathematics & Statistics 145, DOI 10.1007/978-3-319-20585-4_1

on scales is further developed for selected items. The data to be considered here for LCM are a subset of the full set of data collected in the Cognition in the USA (CogUSA survey; McArdle & Fisher, 2015). These scales were chosen in a way that would be consistent with the principles of *multiple factorial invariance over time* (MFIT) but the result of the age-related changes over two waves was largely unknown and in need of establishment. Basically, we first try to establish MFIT over the two waves and then look for latent changes in these scales over age. Thus there are only eight scales to consider here (four cross-sectional scales by two longitudinal occasions), so there is still a lot of work to do!

It is well known that longitudinal data can deal with different concepts than cross-sectional data (see Baltes & Nesselroade 1979; McArdle & Nesselroade 2014). That is, cross-sectional data has many good opportunities for "between person differences" but it cannot deal with "within a person changes." The first dependency that is created and observed is that the same person is used at multiple occasions. This dependency has been used in multivariate modeling a great deal. Because the same person has multiple inputs and outcomes we can deal with this in different ways. All of the individual changes that can be examined are due to the longitudinal models (see McArdle 2008) allowing dependencies among the observed scores at various time points. This dependency is also responsible for the popularity of multi-level modeling (see Bryk & Raudenbush, 1987, 1992). It is demonstrated here that the statistical power to detect changes is an explicit function of the positive dependencies and the timing of the observations.

The typical lack of dependency is monitored in statistics by a careful assessment of the original scores, typically using linear regression with an outcome score ($\mathrm{Y_n}$) and a predictor ($\mathrm{X_n}$) score and usually written as

$$\mathrm{Y_n} = \beta_0 + \beta_1 \mathrm{X_n} + e_\mathrm{n}, \tag{1}$$

where the regression terms $\beta_0$ and $\beta_1$ are thought to apply to everyone, and the residual term ($e_\mathrm{n}$) is an individual characteristic that is unmeasured and supposedly follows a normal distribution. This is an effort to find the relationships between some outcome Y and the input variable X. If X is a group then this model provides a way to determine group differences on the outcome (the usual ANOVA as a between groups *t*-test). But this is not an effort to deal with observed dependency in traditional regression analysis (see Fox 1999).

But some people noticed that having an individual measured more than once created a statistical virtue. Indeed this was the stimulus for progressively repeated measures. One classical representation of longitudinal data can be found in the *repeated measures model for the analysis of variance* (*RANOVA*; see Fisher 1925). In this first model the individual score at any time point ($\mathrm{Y[t]_n}$) is assumed to be decomposed as

$$\mathrm{Y[t]_n} = \beta_{0\mathrm{n}} + \beta_1 \mathrm{X_n} + e\mathrm{[t]_n} \tag{2}$$

where the individual (n = 1 to N) is allowed to differ at all throughout the time series (t = 1 to T) in two ways: (1) Individuals are different from one another at all times, and (2) there are random normal fluctuations at each time point ($e\mathrm{[t]_n}$). The use of the X weighted function is an adjustment in the mean of the scores for group differences in the trends over time. This model can give correct statistics for the mean of the individuals and the effect of X (assuming it is the same over all occasions) as long as the contrast questions are "spherical" in shape (among others, see Davidson 1972; Huynh & Feldt 1976).

The repeated measures model permits the power to detect differences between treatment groups in means (or over time) as a function of the standard deviations of the scores (as usual, with the sample size included as the square root of N at the end). But in repeated measures, the variance at the second occasion is also based on the correlation of the observed scores over time:

$$\mu_\mathrm{d} = (\mathrm{m}\,[1] - \mathrm{m}\,[2]) \,/\, \left(\mathrm{s}[1]^2 + \mathrm{s}[2]^2\right) - 2\ \ ((\mathrm{s}\,[1] + \mathrm{s}\,[2])\,\mathrm{r}\,[1,2]) \tag{3}$$

where we have symbolized the estimated mean difference as $\mu_\mathrm{d}$, using the two observed means as m[1] and m[2], the two observed variances as $\mathrm{s}[1]^2$ and $\mathrm{s}[2]^2$, and the observed correlation over time as r[1,2]. This is nothing more than the mean difference over the standard deviation, but the correlation is for the same measure at two occasions. So for the same mean difference (m[1] − m[2]) as found in a cross section we can say we have found a significant different from zero if the correlation of the two measures is positive (which it typically is; see Bonate 2000; Cribbie & Jamieson 2004). For this reason, it is typically far better (depending on the sign of the correlation) to measure a person twice than to measure twice as many people just once. That is, *the longitudinal case is far more powerful than the cross-sectional case*. This is not the only issue of statistical power (see Tu et al. 2005) that could be considered, but it is relevant here. Of course, there are more than two time points over which change is to be measured, and this typically increases our power.

## The Move to a Latent Curve Model

A straightforward generalization of this *RANOVA* model allows the move to a *latent curve model* (*LCM)* and makes it not very hard to understand. This LCM was first used by Tucker (1958, 1960 1966) and Rao (1958), and later Meredith and Tisak (1990) gave it a *structural equation model* (*SEM*) interpretation (also see McArdle 1986 and McArdle & Epstein 1987) to determine the best fitting curve to the observed data. Basically, the slope can vary along with any way the individual changes. Each individual is assumed to have three latent variables, defined as

$$\mathrm{Y[t]_n} = L_\mathrm{n} + S_\mathrm{n}\Omega\,[\mathrm{t}] + u[\mathrm{t}]_\mathrm{n} \qquad (4)$$

so the three sources of variation in any response are: (1) A constant change for the individual over all times (the latent level = $L$), (2) a systematic change (based on a slope score = $S$, which is systematic with the set of basis coefficients = $\Omega$[t]), and (3) a unique change = $u$[t], which is essentially random with respect to the other changes. We can examine that the set of basis coefficients ($\Omega$[t] is not necessarily linear) to determine the slope of the best fitting line or trajectory of the data, but this line supposedly has the same coefficients for everyone.

All sources of individual differences are indexed by variance ($\phi_\mathrm{L}{}^2$, $\phi_\mathrm{S}{}^2$, and $\psi^2$). In addition, the constant change is allowed to have covariance ($\phi_\mathrm{LS}$) or be correlated ($\rho_\mathrm{LS}$) with the systematic changes. The variance that remains (the uniquenesses, $\psi^2$) is assumed to be uncorrelated with the changes or the starting point and is furthermore assumed to be equal over time.

We can also have the observed group effects on these individual coefficients, and we can do what we want with them. What is usually done follows the usual regression logic with two of the latent variables as new outcomes:

$$L_\mathrm{n} = \alpha_0 + \alpha_1\mathrm{X_n} + e_\mathrm{Ln} \quad \text{and} \quad S_\mathrm{n} = \beta_0 + \beta_1\mathrm{X_n} + e_\mathrm{Sn} \qquad (5)$$

in which case the $e_\mathrm{L}$ and $e_\mathrm{S}$ account for the residual variance and covariance. This kind of mixed model function, including both fixed ($\alpha_{0,}$ $\alpha_{1,}$ $\beta_0$, $\beta_1$, and $\Omega$[t]) and random ($\phi_\mathrm{L}{}^2$, $\phi_\mathrm{S}{}^2$, $\psi^2$, and $\phi_\mathrm{LS}$) effects, can be evaluated for goodness of fit using the standard SEM statistical logic (see Meredith & Tisak 1990; McArdle 1986). If the model fits the data of means and covariances we assume that the score model (of [4] and [5]) is reasonable.

The kind of change we will test is dependent largely on the set of basis coefficients we employ. We can force the systematic change to be linear with the time simply by fixing the coefficients $\Omega$[t] = [0,1,2,3 . . . T]. This is often done, but it is only one option, and there are many others. We can even estimate some of the coefficients (T-2 in the one factor case) so that they form an optimal curve for the data. This is basically what the earliest pioneers (Tucker, Rao, Meredith, etc.) did. But there are many more ways to examine the curves and a lot can be done here. Using the basic logic, we can also consider more than one curve for these data (as done in later chapters).

The LCM is considered useful now because it can describe both, group (i.e., fixed) and individual (i.e., random). For this reason it is popular in psychology where we often are interested in group effects but individual differences from the same perspective. We should note that it is not widely used in other areas of science (e.g., Econometrics) where the dominant paradigm uses time as a causal hinge, so which measure came last in time is regressed on all the prior instances. The same longitudinal data can be used in this way (see McArdle 2008; McArdle & Nesselroade 2014).

We note immediately that the LCM does not try to explain how the prior time points (if measured) impact the subsequent events. This makes the procedures of LCM more descriptive than inferential. But all is not lost because there is some savings in the number of parameters used to define these differences.

## Model Fit and Model Selection

A good question can be asked about "Does the model fit the data?" This question can be answered in a number of ways. But what we want is a model that has easy to understand parameters and fits as well or better than others of its kind. The approach, known by the *Bayesian Information Criteria* (BIC) is used throughout this book so it is useful to investigate it further now, according to Raftery (1996) and Nagin (2005, p.64) the formula for BIC can be written as

$$\text{BIC} = log\,(L)\text{–}1/2p\;log(n) \tag{6}$$

where the *log* is the natural logarithm, and *L* is the model's maximum likelihood, and this is penalized (lowered) by *p*, the effective number of parameters used, and *n*, the sample size of individuals used. "If one is comparing several models we should prefer the one the lowest BIC values." (Raftery 1996, p. 145). In this way, the BIC "counterbalances" a good fitting model by the number of parameters and the sample size used. So, although it does not seem to be the fit of the model, it can help choose one model among many others. What we hope to obtain is a model where the BIC is as negative as possible, although there are several ways to use this information. Several keen insights into how this BIC behaves are given in Nagin (2005), and these will not be repeated here, but the use of Bayes factors is illustrated. The use of the BIC is obviously Nagin's favored device for model selection with groups, but he does conclude that:

Such debate is important for advancing the theoretical foundations of model selection. However, disagreement about the technical merits of alternative criteria may obscure a fundamental point—there is no correct model. Statistical models are just approximations. The strengths and weaknesses of alternative model specifications depend upon the substantive questions being asked and the data available for addressing these questions. Thus the choice of the best model specification cannot be reduced to the application of a single test statistic. To be sure, the application of formal statistical criteria to the model selection process serves to discipline and constrain subjective judgment with objective measures and standards. However, there is no escaping the need for judgment; otherwise insight and discovery will fall victim to the mechanical application of method. In the end the objective of model selection is not the maximization of some statistic of model fit. Rather it is to summarize the distinctive features of the data in as parsimonious a fashion as possible (Nagin 2005, p.77).

I can easily say I am in complete agreement about these model-fitting issues.

## Potential Biases

Thus, the collection of longitudinal data is useful because: (1) They allow the study of the natural history of the development of problem behavior, such as externalizing behavior, its onset and termination. (2) They allow the study of trajectories or pathways. A pathway is defined as “when a group of individuals experience a behavioral development that is distinct from the behavioral development of another group of individuals” (Loeber & Farrington, 1994, p. 890). Trajectories or pathways provide information of processes of continuity and discontinuity and on inter-individual differences. In addition, Loeber and Farrington (1994) postulate that the best studies now rely on multiple informants. The chapter by Stemmler and Lösel (Chapter 4) meets all of these criteria and this chapter should be considered carefully.

But we need to be clear about the difference between a repeated measures design and a multivariate design because both allow correlation over time. For both, sample members are measured on several occasions, or trials. But in the repeated measures design, each trial represents the measurement of the same characteristic, in the same way, at a different time. In contrast, for the multivariate design, each trial represents the measurement of a different characteristic. It is generally inappropriate to test for mean differences between disparate measurements, so the difference score is useful (in contrast to what is stated in Cronbach & Furby 1970).

But the longitudinal method is not without some well-reasoned detractors (see Rogosa 1988). Among many critiques of the longitudinal method: (1) It is hard to get the representative sample to come back to a second testing, and the people who do come back have done very well at the first time (see McArdle 2012); (2) if they do come back, they have seen the measures before, so it is difficult to measure exactly the same constructs at a second time, without retest or practice effects; and (3) the construct or thing that we want to measure may have changed, and we will not know it by simply looking at the variance or taking the difference between measures. These are some of the many potential confounds of the longitudinal method.

The results of these problems lead us to think that a cross-sectional study had less potential confounds than a longitudinal study. This is hardly ever true because these conditions can occur in cross sections as well, and we may not know it.

## Assumption 1: In the LCM, the Latent Scores Used Are Related to Latent Change Scores

It seems that all the prior work has focused on the “change” at the individual and group levels but very few researchers are willing to say so. Instead, words like “curve” or “slope” or “trajectory” are used. But there turns out to be an easy way to represent these basic change ideas and we will usually do so here.

We can define the basic model of change to isolate the functions as

$$\mathrm{Y[t]_n} = L_\mathrm{n} + \sum \mathrm{i} = 1, \mathrm{t}\ \{\Delta \mathrm{y[i]_n}\} + u\mathrm{[t]_n} \tag{7}$$

so the changes are just accumulated up to that time (i = 1 to t). This is not intended to be a controversial statement and it leads to the same fit as the prior linear models, but it is really another way to consider have the outcome at time t (after McArdle, 2008).

The change as an outcome can be strictly defined at that latent variable level (after McArdle & Nesselroade 2014) as

$$\Delta \mathrm{y[t]_n} = \mathrm{y[t-1]_n} - \mathrm{y[t]_n} \quad \text{or} \quad \mathrm{y[t]_n} = \mathrm{y[t-1]_n} + \Delta \mathrm{y[t]_n}, \tag{8}$$

so the latent score is the source of all inquiry. This can be useful in a number of interpretations, especially for the regression of latent changes. For example, we now can fit

$$\Delta \mathrm{y[t]_n} = \beta_0 + \beta_1 \mathrm{X_n} + e_{\Delta \mathrm{n}} \tag{9}$$

so the latent change score is modeled directly, and has a residual ($e_{\Delta \mathrm{n}}$). But the LCS approach is entirely consistent with the LGM approach, as stated by McArdle (2008) and this is why the same values emerge for various estimates. The LCS model is largely a clearer change-based re-interpretation of the LCM, and the LCS model can be programmed and used efficiently (see McArdle 2008; McArdle & Nesselroade 2014).

Latent changes are apparent in this model. Much more could be said about this approach, but this is all that will be needed here.

## Assumption 2: In the LCM, the Model Parameters Have the Same Shape for Everyone

This assumption is also true of all regression models (see Eq. (1)) but it is most clearly not appropriate here. That is, we can control the size and sign of some parameters of the trajectory with the means and the variances of the latent variables, but the shape of the latent change is a combination that is beyond the usual reach.

The chapters listed here do distinguish between these shapes using an unobserved difference between people. That is, this clear difference between individuals is recast at the main reason they are members of a latent grouping—a mixture of different distributions. This was evidenced in the brilliant early work of Tucker (1960 1966, also see Tucker 1992), and the subsequent maximum-likelihood formalizations of Nagin (1999 2005) and Muthén and Sheddon (1999).

This logic using multiple groups is indeed a good idea, because it is focused on different kinds of changes within the person. But Tucker (1960 1966 1992) seems to have found a way to differentiate people with standard methods of factor-cluster analysis. Perhaps the first time this procedure was used in real questions and stated

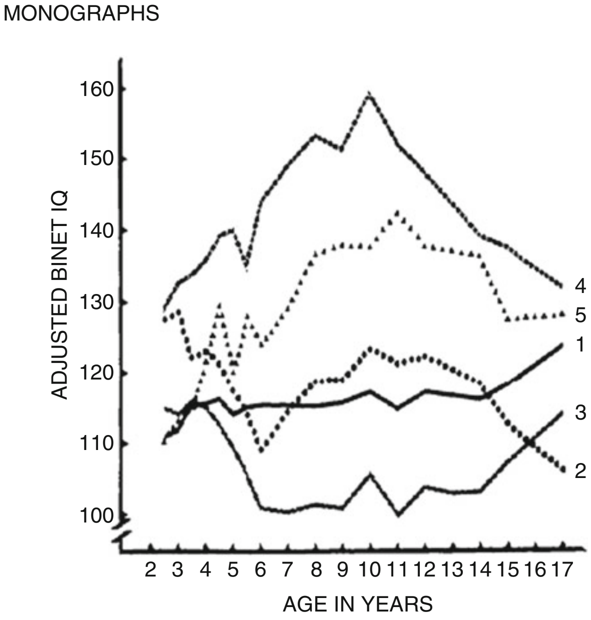


**Fig. 1** From McCall, Applebaum & Hogarty (1973, p. 48)

clearly was by McCall, Applebaum, and Hogarty (1973, pp. 44–48) who suggest that there are five clusters of people based on their changes over age in IQ tests over age (see Fig. 1).

Now it is clear that Tucker (1958 1960 1966) did not have all the statistical tests (or MLE) to support these choices, nor did he have or did develop the mixture model as the possibility of a person belonging to multiple clusters (this allowing for a much better mixture), but he did distinguish large group of persons on their trajectory using multiple factors and he resolved multiple clusters, so we will generally consider Tucker's (1958 1966) work as pre-dating the more recent work of Nagin (1999 2005) and Muthén and Sheddon (1999).

## Assumption 3: In the LCM, the Residuals Are Equal and Uncorrelated, and the Model Fits

There is much more that could be said about the equality of the unique variance (for details, see Grimm & Widaman 2010) but the basic idea is on must have an a priori theory about why these kind of unique but uncorrelated changes are needed. If we do have such ideas we can remove the variance terms at each time and achieve a much better fit to the data. We will not deal with these issues too much here. In this regard this is an unchallenged assumption that deserves much more scrutiny.

The simple fact that "everything else" is supposedly uncorrelated is actually never met and yet this is what is tested by the model fit. The test of goodness of fit is supposed to test whether or not the LCM can be considered viable. But the way we typically test any hypothesis is to remove all other features until all that are left are random variables. This is primarily because we do know how to test for random events (usually with the $\chi^2$ goodness-of-fit test; but see Raftery 1996).

## Assumption 4: In the LCM, the Model Has the Properties of Invariant Measurement

In all cases, it is also necessary to illustrate the loss of fit due to "multiple factorial invariance over time," (MFIT) and how this invariance can be crucial to understanding changes. That is, some things may not change while others will. Here we will only use common factor analysis in a simple example. This is a second dependency because the measures are somewhat the same within a time. Some elaboration of the early work on any scale is further developed for items. This is related to both "test bias" and "harmony." That is, if we assume that a test is a good measurement of a construct, it should behave the same way at all waves.

I do not view MFIT as a "testable hypothesis" as many others do (e.g., Meredith 1993) but I view this as a necessary feature of longitudinal data. That is, in the absence of MFIT it is not clear that we can take differences between successive occasions, and this is critical to most any accumulation model. Thus, this test would be a useful foil against a measure, and we can use it to evaluate an existing measure. But to create one, we must be accumulating something, and that something is strictly defined as the object of our MFIT. Perhaps it is best to say we can evaluate the part of the MFIT that works the way we intended. At least our intentions for MFIT are clarified in this way.

## Assumption 5: In the LCM, the Model Variables All Have Normal Properties

Another kind of dependency is that due to items that are miscalculated as normal. That is, we typically assume all variables are normally distributed, even when they are highly skewed. This is also the case of a variable that can reach an upper or lower limit and should be considered censored (see Wang et al. 2008). As we do not illustrate here, but could have, this can pose a major problem for our understanding of the changes (but for an example, see Hishinuma et al. 2012; McArdle et al. 2014).

## Assumption 6: In the LCM, the Individuals Have All Been Measured at Exactly the Same Developmental Time Periods

This is also probably never true in epidemiological and psychological studies. The problem comes only because the model assumes this is true. In fact, the age-at-measurement is usually not told to the analyst. This means people can be "measured on their birthdays" or at approximate yearly intervals of time, but we just never know. The word "approximate" is used here frequently, and many see this as a natural feature of longitudinal data. But it is not. The big problem that this creates is that the correlations over time, if they are not in a sequential proper timing, can yield some haphazard results. The timing is important to future studies and not enough is done about this issue yet.

The further assumption that we know the true developmental timing is quite absurd. We do not know this and we do not track it very well either. It could be age or it could be something else like puberty (see McArdle 2011), but we need to know it to state how the individuals form groups of people (see Nagin 2005). We often just use whatever longitudinal data we are given, because we are very happy to get some, and we assume we can do something with it, as is. But we cannot.

## The Studies of the First Section of This Book

The studies of the first section of this book seem to criticize some of the basic assumptions of the standard LCM. This should be considered fair as a target because it is loaded with assumptions and the linear LCM was designed to be just a starting point for future work. The concepts of simultaneous estimation are also critical here to distinguish what is being done.

The first study by Paolo Ghisletta, Eva Cantoni, and Nadège Jacot as presented here is an examination of more than linear relationships in psychological research, which they term an NGCM (for nonlinear growth curve model). That is, they do not stop at the quadratic form of the prior LCM, and they do not consider the linear model to capture all the relevant variation in their outcomes (in their example, four blocks of 20 trials of time on task in a pursuit rotor task). Instead, they consider other terms (see their Eq. (6)) that are not a usual part of this basic model (our Eq. (4)).

These author(s) do fit a wide variety of nonlinear models to these data, and this is notable, and they compare each, and this is also notable. But they do drop linearity quickly as a possibility and I think this is a mistake. That is, before we deal with how nonlinear a model can be I think we ought to first see how linearity works, in terms of explained variance at each time point ($\eta^2[t]$) at least.

So I also think these claims can be made from a different perspective. That is, the LCM with a different curve may capture some of these individual changes. The curve could obviously be defined using the last 18 measurements, but an exponential curve could be fitted with less parameters. Nevertheless, the model with the best fit for least parameters is an obvious choice. This, at least, is how I could deal with all the nonlinearity that seems to be present here. I would like to see LCM and the quadratic model as a comparison in their tables.

The second application titled "Stage-sequential growth mixture modeling with criminological panel data" is by Jost Reinecke, Maike Meyer, and Klaus Boers does exactly what this title suggests. However, it uses *General Growth Mixture Modeling* (GMM, from Muthén & Sheddon 1999) within a LCM framework to empirically distinguish between people. Expanding upon the prior work of Kim and Kim (2012) they consider three distinctive types of stage sequences: (1) stage-sequential (and linear) growth mixture models, (2) traditional piecewise GMM, and (3) discontinuous piecewise GMM and sequential process GMM. These three models are applied to a range of adolescence and young adulthood using data from the German panel study termed, *Crime in the modern City* (CrimoC, Boers et al., 2014). In the case of count variables a Poisson or negative binomial distributions (following the work of Hilbe, 2011, not Nagin 2005) can be considered which give a better model representation of the data. With the count data that criminologists seem to have, the Poisson model for measurement is used because it is more appropriate. That is, a regular regression model (but not evaluated) may still work, but the Poisson model that is used here as a measurement device because is sensitive to the use of a probability of an event. The zero-inflated Poisson (or ZIP; see Nagin 2005) model may even be a better choice because it essentially proposes that the reason for the zero counts (no criminal acts) is possibly different than the reasons for the rest of the counts (one, and so on). This can always be compared to the assumption of a continuous distribution of the LCM. And this all can be combined sequentially in a program like Mplus (Muthén & Muthén 2012).

This chapter is notable in a number of ways. First the author(s) use a three-part curve model, with knot points that are notable in terms of substance. This is a distinction that is worthwhile to make and it could be pursued further. I do not see this as quite as different as the typical LCM, so I would compare the fit of both of them. Second, they simultaneously use a measurement model based on a Poisson distribution for the scores. This is decidedly different and is most appropriate for data that comes in the form of counts. But their justification for the use in real data is not presented clearly. Third, they simultaneously use a mixture model to examine for the German Crime data. This use of multiple groups is based on the trajectory differences and they assume these cannot be accounted for otherwise. I would very much like to hear what Nagin (2005, p. 54) says about this part of the analysis. But in any case, any one of these three concerns would be a challenge to fit but they proceed as if this is all standard. This is not standard, and what they do here is quite amazing, partly because it can be done at all.

The differences between the current versions of Mplus (Muthén & Muthén 2012) and SAS PROC TRAJ (Nagin 2005) are important here. Currently, in Mplus, we can ask if any parameter is invariant over groups, and we do not need to define the group membership in advance. This can be in terms of any mean, regression, or covariance component. But in this same sense the analysis is entirely exploratory.

If we further assume that the factor loadings Ω[t], for at least t = 3, T, are different we can have different curves. This can be written with different means and variance terms so the entire placement within groups can differ. This is somewhat different than assuming different linear or polynomial coefficients for the same data. Much more could be said here (see Nagin 2005, p. 54) but Mplus 7 (now used by almost everyone here) seems much more flexible to me now. But I fully expect the debate about "groupings" will go on, and this is productive.

The third application by Mark Stemmler and Fredrich Lösel is titled, "Developmental pathways of externalizing behavior from preschool age to adolescence," and also uses general growth mixture modeling (GMM) with BIC this time to separate five categories of persons among their total sample size of $n = 541$. The goal of this study is to analyze the data of the Erlangen-Nuremberg Development and Prevention Study (ENDPS; Lösel et al., 2009) for the first time with regard to different trajectories for externalizing behavior. ENDPS is a normative sample and is a combined experimental and longitudinal study on antisocial child behavior covering a time period of nearly ten years. Social behavior was rated by multiple informants such as self, mothers, kindergarten educators, and school teachers. Using this longitudinal data, they seem to have found (1) the "*high chronics*" (2.4 %; $n = 13$), who are receiving the highest values for externalizing behavior from childhood on up to adolescence; (2) the "*low-chronics*" (58.8 %; $n = 317$) who are low on externalizing behavior throughout the years; (3) the "*high-reducers*" (7.9 %; $n = 43$) who start out high in childhood, but who reduce their externalizing behavior monotonically over time; the (4) "*late-starters-medium*" (8.7 %; $n = 47$); and the (5) "*medium-reducers*" (22.4 %; $n = 121$). The results stress the idea of a life course perspective, which enable the study of the natural history of the development of externalizing behavior, its onset, and termination.

In all, these authors give an excellent history of the GMM, and demonstrate how it has been used before in many criminological samples. They seem to show that most studies report between three and five groups (with a total range of two to seven groups), and they use the BIC. Most studies show the group of life-course persistent or chronic offenders, and one group that does not exhibit violent, aggressive, or delinquent behavior; in addition, there are existing groups of late onset or desisting. Jennings and Reingle (2012) claim that the number and shape of the groups depend on the nature of the sample (high risk versus normative sample), the life course captured, the length of the observation, and the geographical context. Among the author(s) conclusions, they postulate that further research should be based on multiple observations and across multi-informants (e.g., child/youth reports, parents and teacher report) to ensure the best results. Since this result requires expertise in criminology, we must leave it up to the reader to make sense of these trajectories.

The fourth application by Jang Schiltz is proposal for the potential extension of "the Nagin model" of multiple groups. This can be a quite useful technology because in this representation we do not have to think everyone has the same general nonlinear slope of their trajectory. The problem with Nagin's original formulation is that he only determined trajectories for the mean level and a quadratic slope, and less effort was put into the variance terms or other forms of the slope (see Nagin 2005,

p. 54). These changes are made and the basic model is extended here to include group differences in the slopes and the error terms.

Since we all believe that there will be substantial heterogeneity in real data—different change patterns for different groups—and the LCM will not be capable of dealing with these based on two means and covariances alone, it is clear that this model is more correct. This and other examples on the use of the mixture model is certainly a powerful latent variable modeling approach. But this latent variable model is not the only way to explore the groups—they can even be formed out of measured variables too (see Brandmaier et al. 2013).

The exploratory use of measured rather than latent variables is attractive on a number of counts. First, there are usually many extra ancillary variables that are measured and used as covariates for no particular reason other than they exist. As we will demonstrate, this typical usage can tell us something about their impact on mean differences or between group effects. But what we are interested in is putting them into the analysis is to see if they impact the variances and covariances also. Second, there are always extra ancillary variables that are measured and these could be selected for this exploration. That this is any mixture model is an exploration that is obvious to anyone who uses them and the selection of a group is complicated. So we do not try to handle all these assumptions at once but instead we refer to Nagin (2005) for details on this issue.

## Our Cognition in the USA (CogUSA) Study

Our CogUSA study (see McArdle & Fisher 2015) was designed to do something different than those in this section—that is, the most notable feature of the design of this particular longitudinal study is the variation of age at the initial time, and the variation between time intervals for different waves of testing. As stated earlier in our last Assumption 6, this is a feature of many psychological measurements although it is hardly ever dealt with on a formal basis.

Our ability to measure similar constructs in an in-person *face-to-face* (FTF) interview and over the *telephone* (TEL) is not the key issue here, but it is important. In prior surveys (including the HRS; see Juster & Suzman, 1995; Heeringa, Berglund, & Khan 2011) the only human abilities measured over the phone (say, using the *Telephone Interview of Cognitive Status*; TICS; Fisher et al. 2013) were the very simplest ones (*Episodic Memory* and *Mental Status*; see McArdle, Fisher, & Kadlec 2007). It is not too surprising that these simple variables could be measured in the same way in either modality (FTF or TEL) and still retain MFIT (see McArdle 2010; McArdle & Nesselroade 2014).

But when we consider measuring something as important in aging research as *fluid intelligence* (*Gf*) in a survey, we remain perplexed (see Lachman & Spiro 2002). This variable needs to measure “reasoning in novel situations” and this is fairly hard to do. One of the ways this can be done in surveys is with indices that supposedly measure *numerical reasoning* (*NR*), a decided subset of all reasoning and thinking, and the measure of *numerosity* (*NU*) from the HRS is a good indicator of this. Another way to consider *NR* this is to measure *Serial Seven's* (*S7*) from the HRS, because this takes some *NR* as well as holding specific but complex ideas in memory (see Blair 2006). Still another way to indicate *NR* is to measure something like *Number Series* (*NS*) because these are intended to be small puzzles in numerical form.

One adaptation is that we initially reasoned that people, especially older people, would not take all test items necessary for a reliable score on anything, so the items administered had to be cut down. In the case of both *Immediate Recall* (*IR*) and *Delayed Recall* (*DR*) and *Numeracy* (*NU*) and *Serial 7's* (*S7*) the work had already been done by the HRS staff. These were properly considered as short forms due to the required telephone constraints on time.

The final telephone definitions follow on Table 1. They were all administered over the telephone and this is a limitation because we do not really know what the respondent is doing. These include definitions of IR, and DR to measure a *general memory or general retrieval* (*Gr*) factor, and NU, and NS to measure a *general fluid* (*Gf*) factor at each time ([1] or [3]). We will see if the fit of this specific two factor model is different than a one *general intelligence* (*G*) factor, but we will examine the factor loadings. Clearly, McArdle et al., (2007) found the first two scales (*IR* and *DR*) to be highly correlated (r 0.80) and suggested they be added up and calculated as a single score termed *episodic memory* (*EM*) to distinguish it from another scale of cognitive measurement from the TICS, *mental status* (*MS*; $\{BC + S7 + NA + DA\}/4$), but the second factor here is much different. And we hope it is clear that several other cognitive measures obtained in CogUSA were not yet used here (see McArdle & Fisher 2015).

For common factors to retain their meaning over time, we required them to have "strict" invariance (Meredith 1993). In this case, this implies the factor loadings ($\Lambda$), unique variable intercepts ($\mathbf{I}$), and unique variable variances ($\Psi^2$) are all assumed to be invariant over time (for each measure). We also brought all means differences to the factor score level. This is typically tested but it is clear that any differences or changes over time must go through the common factors or they are not worth using and summarizing at this level. This is basic or, indeed, fundamental to our definition of the latent variables. This does imply that the way we measure the common factors can change from time to time, but for now we assume they are identical at both occasions of measurement.

Many other researchers search for different forms of invariance (e.g., see Byrne, Shavelson, & Muthén 1989; Reise, Widaman, & Pugh, 1993; McArdle, Petway, & Hishinuma 2014), and now this is an evaluation of configural, metric, strong, or strict invariance constraints. We will not partake in this quest again here. This is primarily because we only want the number of factors (*K*) to be determined by what is comparable over time in measurement (as in McArdle & Cattell 1994; McArdle 2007) not by a lack of invariance. There is a prominent thought that the search for the type of invariance of a measure is crucial (see Byrne et al. 1989), but if this is not met then the number (or type) of common factors (can be) needs to be altered to meet this criterion. That is, the criterion of invariance should always be met before

**Table 1** Selected Telephone Measures used in CogUSA (McArdle & Fisher, 2015)

All HRS/AHEAD cognitive measures were selected to satisfy the following considerations: (a) provide descriptive information on a comprehensive range of cognitive functions; (b) span all difficulty levels from competent cognitive functioning to cognitive impairment; (c) be sensitive to change over time; (d) be administrable in a survey environment with lay interviewers, over the telephone, in a short time; and (e) be valid and reliable (from the HRS documentation Report by Ofstedal, Fisher and Herzog. 2005; DR-006). As always, the IWER is asked a series of questions about the incorrect responses. In addition, several other clearly cognitive measures (BC, S7, RF, CESD) are obtained at both waves were not used in these analyses

*IR* = or immediate recall (IR)—One set of 20 stimulus word (from four lists) are read aloud, and the respondent (R) needs to restate these words (no credit is given for errors of any kind). The observed score is from 0 to 10. At W3 they are administered a different list of ten words (from the four lists)

*DR* = or delayed recall (DR)—after about 5 min (depending upon how long it took to do the eight CESD items), the R is asked if they recall any of the words from the IR. They are then asked to restate these words (no credit is given for errors of any kind). The observed score is from 0 to 10

*NU* = "Numeracy"—Since HS 2002, the R is asked to answer up to three numerical questions: (1) "Next I would like to ask you some questions which assess how people use numbers in everyday life. If the chance of getting a disease is 10 %, how many people out of 1,000 would be expected to get the disease?"(2) "If 5 people all have the winning numbers in the lottery and the prize is two million dollars, how much will each of them get?" (3) "Let's say you have $200 in a savings account. The account earns ten percent interest per year. How much would you have in the account at the end of two years?" The observed score is from 1 to 3

*NS* = Even though we wanted to, the Woodcock-Johnson "*Number Series*" items was far too long to be included in CogUSA so we cut it down from about 42 items to about 6 adaptive items. A modification of "which six" items was tried in each of the two occasions, *Wave 1* (W1) and *Wave 3* (W), but both testings supposedly yielded a W-ability estimate of NS. In the W1 testing the plan was to administer a first item of medium difficulty (for their level) and (0) if they got it incorrect an easier item about half way down the scale (based on the known difficulty of the WJ item) was presented, but (1) if the R got the item correct a harder item, about half way up the W-scale, was presented. All testing ending at six items and a WJ score was estimated from this pattern of responses. In the W3 testing s similar items were administered in a block adaptive fashion. The key idea here is to only administer six items, but the same three items are given first, spread out in difficulty, and the second set of three items are supposedly centered around the persons' ability level. In this case a W-score can be formed. Thus we assume, but do not test, MFIT

we evaluate the latent changes (as in McArdle & Cattell 1994). This is only our belief system, and we use this belief at all occasions, but we should point out that it is not one used by many others.

## Methods

### *Available Data*

The data to be analyzed are a small subset (4) of scales from recent tests of *Cognition in the USA* (CogUSA; see McArdle & Fisher 2015). These scales were chosen in a way that would be consistent with the principles of MFIT but the result of the changes over two time points (W1 and W3 here) is unknown. Basically, we first try to establish MFIT over all ages and then look for changes in these scales over ages. We now only present eight scales in all to consider (four cross sections at two longitudinal occasions).

At each occasion, the people who took the HRS (for details, see Fisher et al. 2013) were asked to fill out the forms for all scales. Most specifically, they were asked each time to fill out a questionnaire about their own health and well-being and the full CES-D was included. We did not use sessions at Wave 2, we only use Sessions 1 and 3 primarily because this time-lag did not offer enough Age differences for Age changes to be picked up. We also do not include all items in analyses here, but we will only include eight scales from the full set (of many). We will state that 13/20 items (from the Center for Epidemiological Studies-Depression scale; CES-D) were previously analyzed by McArdle et al. (2014) who seemed to find MFIT in 13 items from the full set (of 20 items). But here there are several differences: (1) We deal with scales not items; (2) the confusion of the usual testing of MFIT was emphasized in McArdle et al. (2014); (3) in CogUSA the ages-at-testing at each occasion are substantially different.

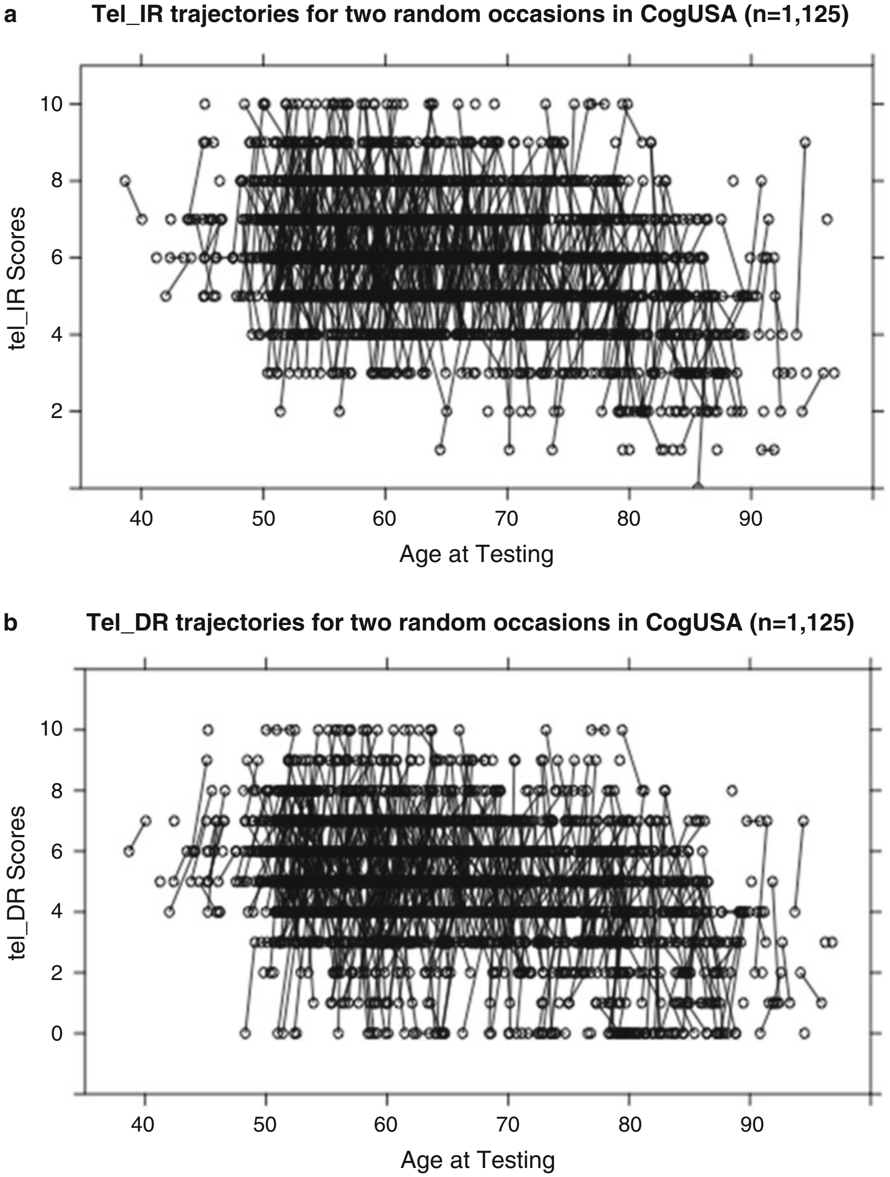
The plots of Figs. 2a–d illustrate what we are trying to examine in the model. These are plots of the four manifest variables (*IR*, *DR*, and *NS*, *NU*) against the ages-at-testing for each person separately (i.e., joined by a line), and these illustrate lots of variability and only one kind of dependency among persons (that is, the people are largely the same ages when they are measured but the scores do change over age). They could change for a number of other reasons (such as errors of measurement or practice effects; see McArdle & Woodcock, 1997).

## Models

Figure 3a is an elaboration of a latent curve model with Age differences as a double cross-sectional variable. The only variable used here is the *NS* measured at two occasions (1 and 3 for comparability) and the age-at-testing is also measured at each of these waves. The model here uses the two occasions in a double cross-sectional mode in an effort to capture the means and covariation of the *NS*-age relationship. That is

$$\begin{gathered}\mathrm{NS}[1]_{\mathrm{n}} = \beta_{01} + \beta_{11}\,(= fun\{\mathrm{Age}\,[1]\}_{\mathrm{n}}) + e_{1\mathrm{n}} \\ \text{and} \\ \mathrm{NS}[3]_{\mathrm{n}} = \beta_{03} + \beta_{13}\,(= fun\{\mathrm{Age}\,[3]\}_{\mathrm{n}}) + e_{3\mathrm{n}}\end{gathered} \tag{10}$$

where some fixed function of age is used as a linear predictor (e.g., such as $fun\{\mathrm{AGE[t]}\} = (\mathrm{Age[t]} - 65)/10$—so the intercept is at age 65 and the difference in score is for each 10 years of Age). But using SEM we can also test whether the equations ($\beta_{01} = \beta_{03}$, $\beta_{11} = \beta_{13,}$ and the respective residual variances, $\psi 1 = \psi 3$) are



a Tel_IR trajectories for two random occasions in CogUSA (n=1,125)

b Tel_DR trajectories for two random occasions in CogUSA (n=1,125)

**Fig. 2** (**a**) Immediate recall (IR). (**b**) Delayed recall (DR). (**c**) Number series (NS). (**d**) Numeracy (NU)

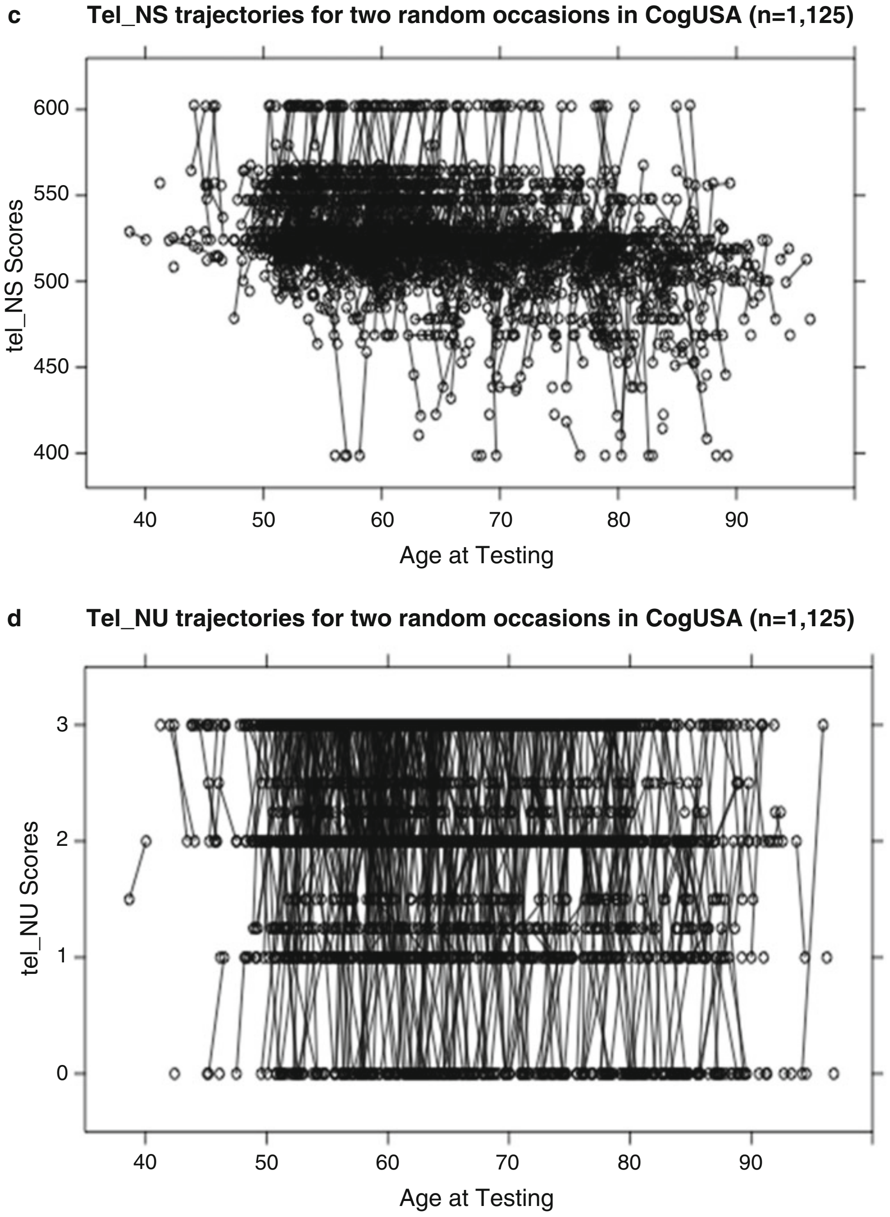


**Fig. 2** (continued)

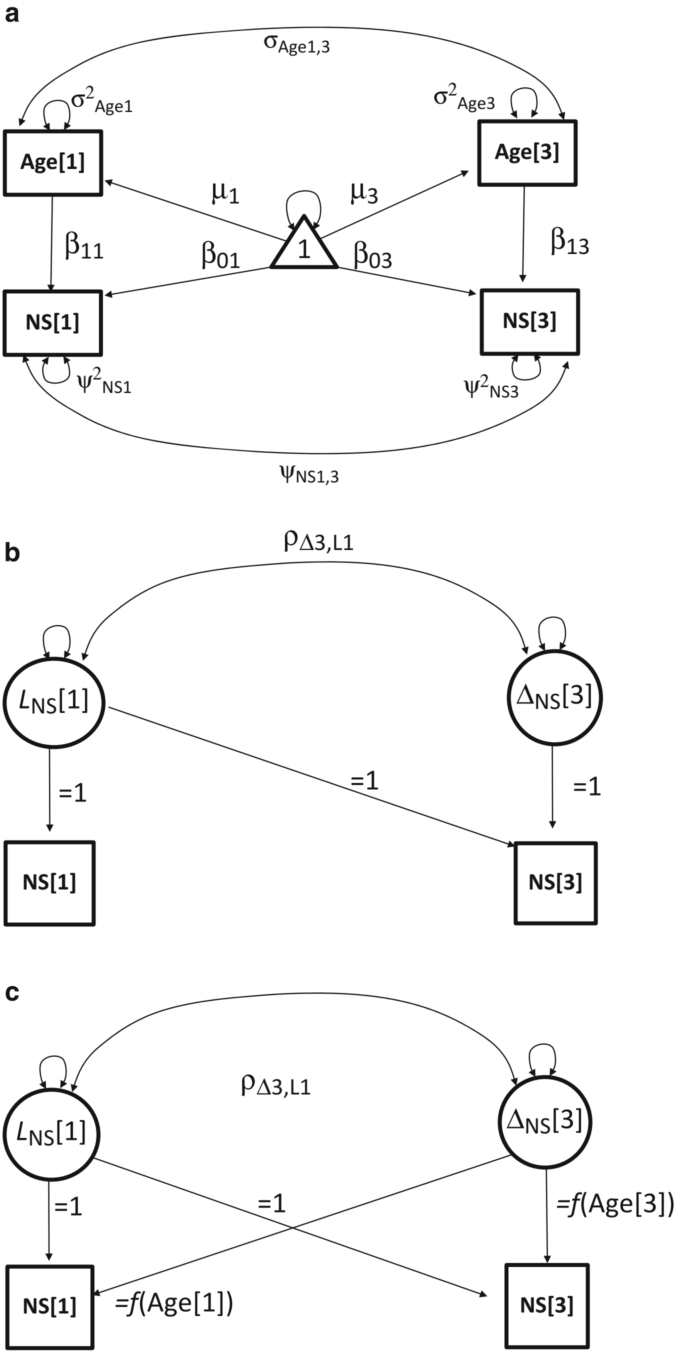


**Fig. 3** (**a**) A path diagram of a one-variable model for multiple waves of measurement (W1 and W3) but as usual treated as a dual cross-section. (**b**) A path diagram of a one-variable LCM model for multiple waves of measurement (W1 and W3). (**c**) A path diagram of a one-variable LCM model for multiple waves of measurement (W1 and W3) with different ages of measurement

supposedly the same at each time of measurement. This could be useful because we may find it does not work the same way at Wave 1 and Wave 3, primarily because of the exposure at Wave 1. At very least, the prespecified *fun*{Age} becomes a testable hypothesis.

Figure 3b is a path diagram of a LCM in cases of only one change. The variable here is the *NS* measured at two occasions ([1] and [3] for comparability). But the model here is an effort to capture the mean and covariation of the *NS* test. We notice that this uses the leftover variation as the difference (or slope) and this simple representation can b credited to Joreskog (1974).

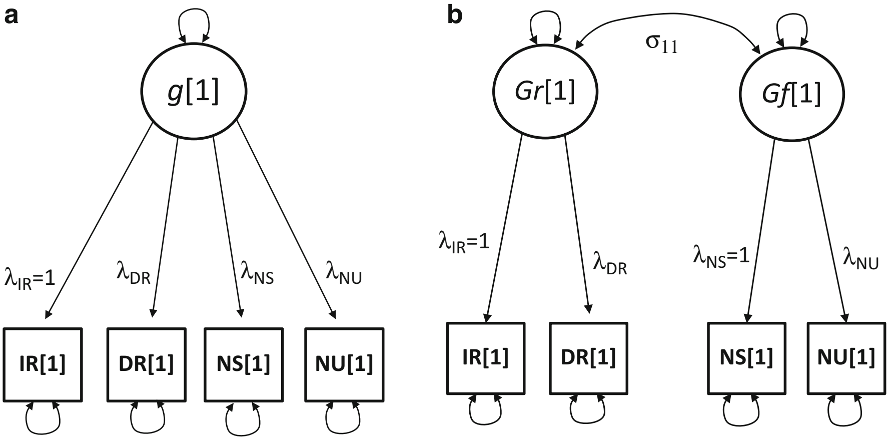
Figure 3c answers a different question about where we would add age variation to this model. Recall in CogUSA (Figs. 2, 3, 4, and 5) there is a lot of age variation at the beginning (Wave 1) and they are not measured over the same age interval over time. This variation in age was considered a random source of variation (and it was done on purpose) because we did not really know how to break up ages into groups. This is an expression of the work of the primary author of this paper (see McArdle & Woodcock, 1998). For these model to work, some predefined fixed function of age (e.g., $=f(\text{Age})$; it does not need to be linear, but it must be pre-specified) needs to be designated as a regression (or as a factor loading) that must be able to change over the individual case (because of the different ages-of-measurement). This precise feature of varying factor loadings can be used in many current computer programs (see Appendices 1 and 2 here for Mplus code). The concept of the individual loading was used by McArdle (1998, pp. 390–406) fitted together with the concept of individual likelihoods (primarily to check on individual fit). This examination of age-variation is an important concept here, but we would use this representation for any departure from the average timing that is measured (see LCM Assumption 6). This is the same concept that was subsequently used by Mehta and West (2000) and Mehta and Neale (2005) in their description of “definition” variables.

## Adding a Latent Variable Measurement Model

Needless to say, these are common factor models where we assume a factor score for each person ($f_n$) is indicative of multiple measures at multiple occasions. This is an important addition and it can be done with SEM. Following McArdle (2007), every variable ($m = 1$ to M) we measure at each time ($t = 1$ to T) can be decomposed as

$$\mathrm{Y[t]_{m,n}} = \lambda_\mathrm{m} f[\mathrm{t}]_\mathrm{n} + u_\mathrm{m} \tag{11}$$

into a common part (the time related common factor score $f_n$ multiplied by a time invariant factor loading $\lambda_m$) and a unique part (the random or unique factor score $u_m$). We can think of the variable having an intercept of mean ($\iota_m$) too, but this could just as well be a property of the unique factor score. This leads to a common factor model hypothesized for each time point.



**Note:**
**Using only Time 1 latent variables**
***G* [1], and Time 1 observed variables**
**IR[1], DR[1], NS[1], NU[1];**

**Note:**
**Using only Time 1 latent variables**
***Gm*[1], *Gf*[1] and Time 1 observed variables**
**IR[1], DR[1], NS[1], NU[1];**

**Fig. 4** (**a**) A latent variable path diagram of the one-factor model at Wave 1. (**b**) A latent variable path diagram of the two-factor model at Wave 1

The specific models fitted to the one time point data are equivalent to many others, so we will not belabor the process. Needless to say, these are factor models where we assume a factor score for each person ($f_n$) is indicative of multiple measures at each occasions ($t = 1$ to T). This is presented in Fig. 4a, b for both one and two factors at one wave. We do notice that the one factor (*G*) also has several demographic influences, including scaled versions of age (and education, sex, and race). That is, in addition to the requirement that this factor account for the covariation of all the internal variables, this *G* must also account for all the covariation of all demographic influences with these measured variables. In our Fig. 4b, two common factors (of *Gf* and *Gr*) are expected, and these two factors are allowed to be correlated above and beyond the external (demographic) influences. This relaxation of the factor pattern is not the only way two common factors could be fit here, but it should fit better (see McArdle & Prescott, 1992).

In the next model we consider a single latent variable, perhaps termed a *G* factor for general intelligence. This is a very popular model for a number of good reasons (see McArdle 2012) and it can be fitted here to the four variables. In this context, the model makes the additional assumption that all four variables have a common part and a unique part. The common part is not necessarily the same for each score, and this size is as indexed by a factor loading ($\lambda_m$) or by the size of its' unique variance ($\psi_m^2$). The size is only relative here and this is made clear by the requirement that one of the factor loadings (or the common variance) needs to be fixed at some positive value (usually $\lambda_1 = 1$).

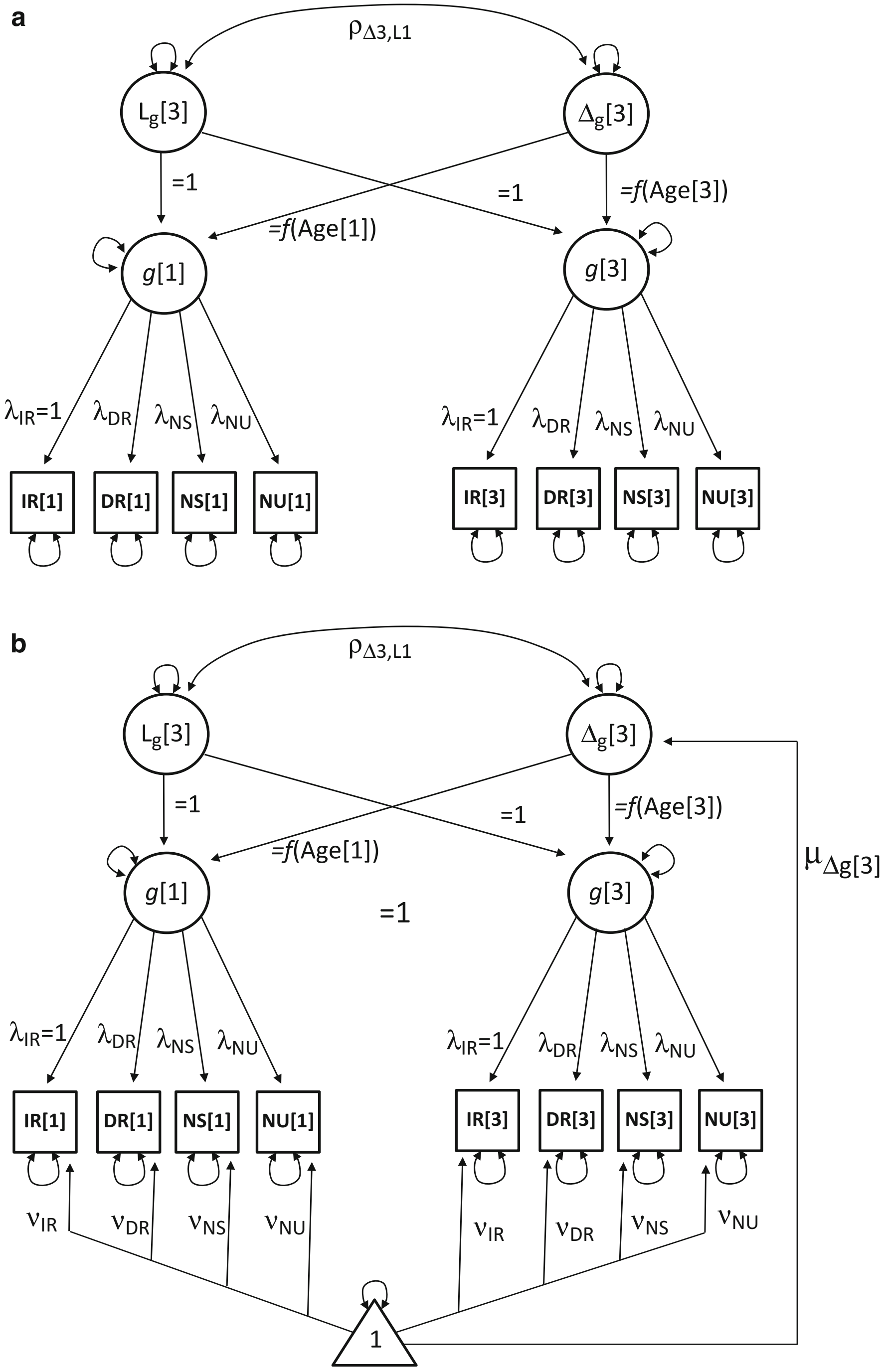


**Fig. 5** (**a**) A latent variable path diagram of the invariant one-factor model for multiple waves of measurement (W1 and W3). (**b**) A full (means and covariances) latent variable path diagram of the invariant one-factor model for Wave 1 and Wave 3

As an alternative, in the next model we consider multiple latent variables termed *Gf,* for *general fluid reasoning*, and *Gr*, for a *general memory* or *general retrieval* function of memory. This multiple factor model is a very popular model for a number of good reasons (see McArdle 2012) and it can be tested (fitted) here to the four variables. This is not necessarily the best fitting model to these data (that is, other factor loadings can be estimated instead). The model requires that each factor loading have a fixed positive value (usually $\lambda_1 = 1$, and $\lambda_3 = 1$) but this is an arbitrary choice that can be altered but must be made by the investigator. Most critically, this is the same as the prior model if the correlation among the two separate factors is unity.

In this context, the model makes the additional assumption that all four variables have a common part and a unique part. The common part is not necessarily the same for each score, and this size is as indexed by a factor loading ($\lambda_m$) or by the size of its unique variance ($\psi_m{}^2$). The size is only relative here and this is made clear by the requirement that one of the factor loadings (or the common variance) needs to be fixed at some positive value (usually $\lambda_1 = 1$, and for the second factor, $\lambda_3 = 1$).

The same comparison of these two models can be examined over time and models for this type of data are drawn in Fig. 5a, b. Here the factor or factors have to do two related things: (1) Define the internal features of the covariation of the measures within a time point, and (2) account for the changes over time in the measures. Since we want the same factor at time 1 and time 2 (or W3 here), and since we define the factors by their factor loadings, we do force the factor loadings to be exactly the same over time. Although it is not necessary for this problem, to simplify our presentation here, we assign the unique variances to be the same over time as well.

Of most importance here is change over age, and the common factor part can further be decomposed as

$$f[\mathrm{t}]_\mathrm{n} = f0_\mathrm{n} + f1_\mathrm{n}\Omega[\mathrm{t}]_\mathrm{n} + e_\mathrm{n} \tag{12}$$

where $f0_\mathrm{n}$ is the unobserved level or intercept of the factor score, $f1_\mathrm{n}$ is the unobserved slope of the changes due to a one-unit shift in the $\Omega[\mathrm{t}]_{\mathrm{n},}$ and $e_\mathrm{n}$ is the random noise or disturbance that is thought to be randomly distributed around the predicted value of the first two parts. In this way, the factor score can change and this creates change in the observed variable even with an invariant measurement model.

The differences between this and other formulations of the more standard *LCM* (Meredith & Tisak 1990; McArdle 1986) are that (1) this is a *curve of factors* model (CUFFS; after McArdle, 1988) and (2) here we explicitly assume the assignment of a factor loading that varies across the individual (McArdle & Hamagami 1996, pp. 106–112; especially p. 108). Of course, individual fitting of likelihoods is a common feature of many fitting functions now (but see McArdle 1998, pp. 390–406), so we use the program M+ here. The consistency assumption for the individual to look like the group is used to form the basic test statistics—this use of an individually measured score as a model parameter is sometimes call "adding definition variables" (from Mx manual; Neale et al., 1993). Indeed, this kind of raw data procedure was only available in Mx in the past, and it was based on the statistical concepts of *unbalanced pedigree analysis* (from Lange et al. 1976).

If aging impacts the latent score alone we would think that it impacts both the levels and slopes in some consistent fashion. To this we add that there can be age differences at Wave 1 and Wave 3, and these are summarized for each person as

$$\begin{gathered} f_{0n} = \beta_{01} + \beta_{11}\,(= fun\{\text{Age}\,[1]\}_n) + e_{0n} \\ \text{and} \\ f_{1n} = \beta_{03} + \beta_{13}\,(= fun\{\text{Age}\,[3]\}_n) + e_{1n} \end{gathered} \tag{13}$$

so the $f_0$ is a level over both occasions, and $f_1$ is a slope that is at two particular time points determined by the age of measurement. The terms are indicative of a level (or when the $fun\{\text{Age}\} = 0$; so at 65 here) and a slope (for each unit—or decade of age—of $fun\{\text{Age}\}$) of the prespecified age function. That at is, each person's unique contribution to the two ages is built up in this way. Each person has a level and a slope score and under the assumption that it is the same information about age changes is present in each variable. We can look further at this function. In other words, we have essentially taken the Age model to the latent variable level.

Figure 5a is a path diagram of this one common factor model and Fig. 5b is a more compete version (including means and unique covariances). We fit the latter one here.

Figure 6a extends this logic to having two common factors at each measurement occasion, and Fig. 6b is a full mean and covariance path diagram of this extension. Here the variable slopes and levels are all correlated, and the factor of levels does not assume a mean difference, due to lack of identification, so we do not add one.

In the final model (Fig. 6b) we use a two factor solution, but we also include: (1) the means in the diagram (as the regression from a constant triangle), and (2) the covariance of any unique features of the data ($\psi^2[1,3]$). This is simply a more complete picture of the model we will fit.

The same principles hold when we move to multiple occasions of data. The common factors are supposedly the same, but the age changes in the factor scores is examined. Here the models of 5a and 6a will be compared.

The key thing we will note about CogUSA is the staggered time lag of this longitudinal study. This is unusual for a longitudinal study (see McArdle & Woodcock, 1997), but we put in time lag as a variable because we wanted to study. That, in most cases of experimental design we vary all the things that are important to us and leave the rest as fixed quantities.

## Results

The summary statistics appear in Table 2 for $n = 1{,}125$ people who supposedly took all four scales at both Wave 1 and Wave 3. These are *full information maximum likelihood* (FIML) estimates because only about 98 % participated at all times. In

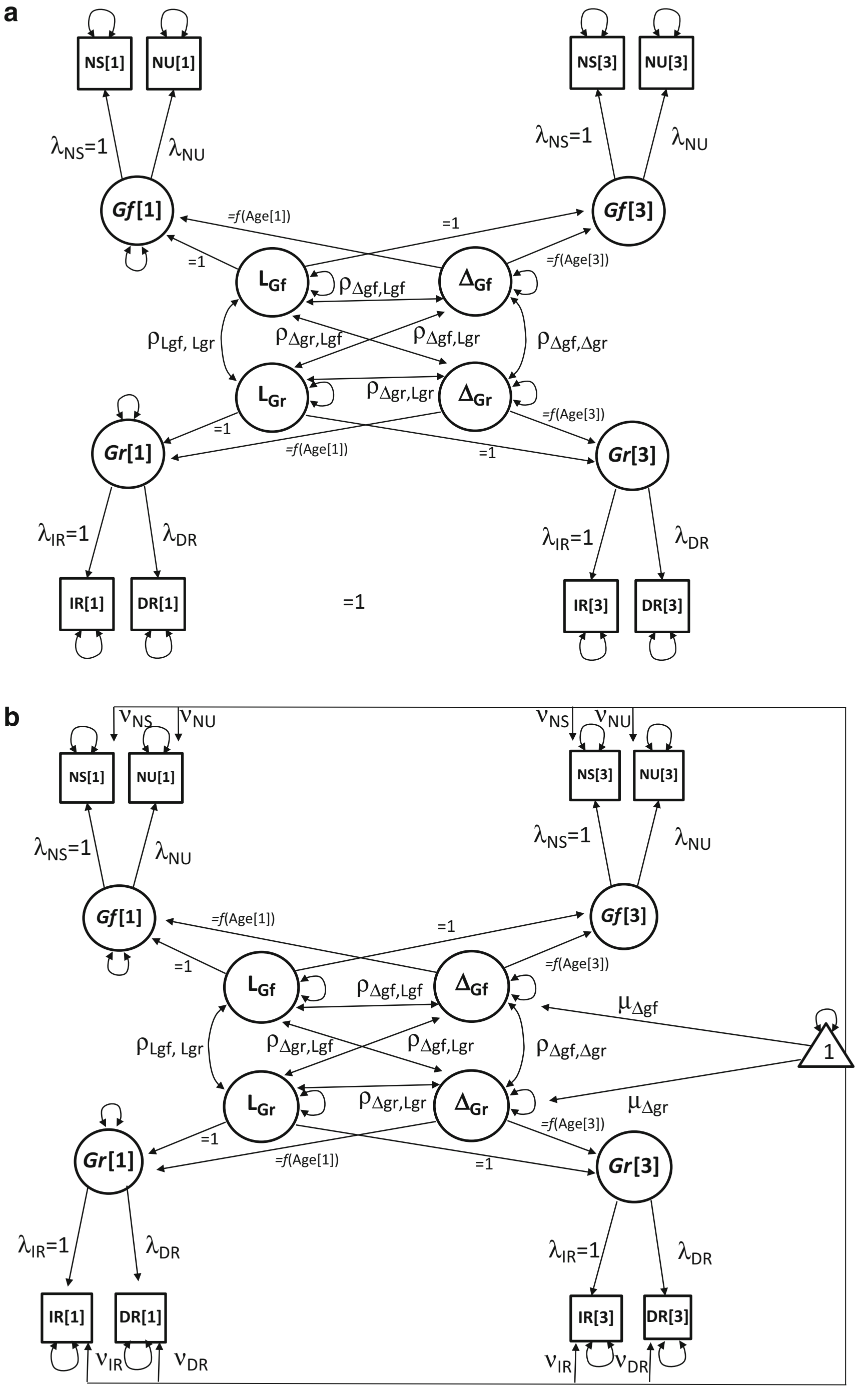


**Fig. 6** (**a**) A latent variable path diagram of the two-factor invariant model for Wave 1 and Wave 3. (**b**) A full (means and covariances) latent variable path diagram of the two-factor invariant model for Wave 1 and Wave 3

**Table 2** Summary statistics for $n = 1125$ participants who were all considered on four scores at both Waves 1 and 3 (and using FIML)

(2a) Means (& Variances)

Means

| | NS[1]* | NS[3]* | NU[1] | NU[3] | IR[1] |
|---|---|---|---|---|---|
| Means | 2.383 | 2.685 | 2.029 | 2.059 | 5.951 |
| (Variances) | (7.790) | (9.444) | (0.969) | (0.912) | (2.807) |

Means

| | IR[3] | DR[1] | DR[3] | AGE[1]* | AGE[3]* |
|---|---|---|---|---|---|
| Means | 6.069 | 4.865 | 5.152 | -0.060 | 0.061 |
| (Variances) | (2.647) | (4.099) | (3.941) | (1.103) | (1.107) |

(2b) Correlations

| | NS[1]* | NS[3]* | NU[1] | NU[3] | IR[1] |
|---|---|---|---|---|---|
| NS[1] | 1.000 | | | | |
| NS[3] | 0.519 | 1.000 | | | |
| NU[1] | 0.456 | 0.433 | 1.000 | | |
| NU[3] | 0.383 | 0.502 | 0.504 | 1.000 | |
| IR[1] | 0.338 | 0.330 | 0.268 | 0.258 | 1.000 |
| IR[3] | 0.218 | 0.295 | 0.177 | 0.245 | 0.425 |
| DR[1] | 0.301 | 0.282 | 0.218 | 0.215 | 0.762 |
| DR[3] | 0.244 | 0.344 | 0.199 | 0.265 | 0.424 |
| AGE[1] | -0.304 | -0.330 | -0.214 | -0.192 | -0.346 |
| AGE[3] | -0.304 | -0.332 | -0.215 | -0.193 | -0.344 |

| | IR[3] | DR[1] | DR[3] | AGE[1] | AGE[3] |
|---|---|---|---|---|---|
| IR[3] | 1.000 | | | | |
| DR[1] | 0.389 | 1.000 | | | |
| DR[3] | 0.759 | 0.443 | 1.000 | | |
| AGE[1] | -0.352 | -0.310 | -0.366 | 1.000 | |
| AGE]3] | -0.354 | -0.308 | -0.366 | 0.999 | 1.000 |

Notes: W1_NS and W3_NS are scaled by mean = 500 and SD = 10

W1_Age and W3_Age are scaled by mean = 65 and SD = 10

To get back to the original scaling of each score we can simply multiply by the SD and add the mean

using FIML we basically assume that there is nothing special about those who did not participate again, and we use their time 1 data assuming they also follow the same general pattern as we observe in those that did come back. But, for example, NS is listed with a mean of 2.38 and a variance of 7.79 at Wave 1, and this can be recast (by the usual W Rasch-scale transformation, following McArdle & Woodcock, 1997; Here this is a transformation that basically raises the score to a power of about 9, and then adds 500) into a raw W-scale score of 524, while the same (or similar) test is listed at a mean of 2.83 with a variance of 9.46 at Wave 3, and this is a raw W-score of 528. It is thought (by many others) that the W-score will have a linear relation with other scales while the raw score will not. The correlation (in Table 2b) is r∼0.520. Thus, the W-scores go up over time in a pattern that is related to time 1 (those that are high to begin with seem to get high scores here). The high scores can be accounted for by the ceiling of 600 on both test forms (this is the highest we would go) and it appears (as we can see in Fig. 2a) people seemed to get this score with unusually high frequency.

The resulting age-based model #1 is based on $n = 1{,}125$ who answered both scales, starting ages with a mean of 68.83 (with a variance of 106.91) and a time lag of 1.21 years (with variance 0.23). These are first fitted to *Number Series* (*NS*) at two waves (Wave 1 and Wave 3) although any one of the four scales could be used. The results are highlighted in the first column of Table 3 and all of Table 4. In addition the computer script used to assess these variables is in Appendix 1.

The numerical results of model #1 shows *NS* has a mean at age 65 of −1.36 and a slope per decade of age of 3.79. These are both significantly different from zero given their respective z-values, so we can talk about the W-scores of 484 at age 65 and with a positive increase of +3.87 (or +38.7 in W units) points per decade from that point onwards (and backwards). This should usually be contrasted with model #0 where no slope is assumed, but the equal variance assumption of the residual was assumed (in fact an equal and fixed variance assumption had to be used, so this model has only two estimated parameters; the mean and variance of the level; see McArdle, 1998). The extra parameters estimated (from the 2 in #0) are the slope mean, the slope variance, and the covariance of levels and slopes. The fact that all variables have such large variance estimates (in #1) and that increasing scores go with increasing ages (this is positive) is a surprise. The fact that model #0 has a much larger BIC is also a question that requires an answer. Perhaps the age changes are too small to count here, but the fact that they are positive is a definite difference from prior results (see McArdle et al. 2007).

The results for the one factor *G* model of behavior is listed in the next two models of Table 3 (#2 and #3). In model #3 we fit a level and slope model with one common factor (as in Fig. 4b). In model #2 we fix this changing score assumption and did not fit a slope to the *G* factor and lost substantially in fit (on $df = 4$). The single factor with factor loadings set $\lambda = 1$ for *NS*, but the factor loading is estimated as 0.77 for *NU*, 2.25 for *IR*, and 1.91 for *DR* seem to fit these data very well (with $df = 3$). The three extra parameters estimated (from the 10 in #2) are the slope mean, the slope variance, and the covariance of levels and slopes. The invariant loadings do not add anything to the misfit here. But the mean intercept of this age slope factor is 2.68 (indicating a raw W-score of 527 at age 65), and the mean of this slope factor is −0.24, and it is significant, indicating a −0.24 downhill slide in this factor for every decade of age that is increased.

This is a decidedly different result. It is not at all what the observed data seem to show (for any subscale) but this does not take the age-of-measurement into account, it uses only the common (not unique) variance, and it is a closer to what we expected. In fact, it suggests that (1) the common factor is *episodic memory* (because the loadings of 2.25 for *IR* dominates the factor), and (2) there is some decline over age. This one factor model is listed as #3 and this is given more completely in Table 5

**Table 3** Numerical results for selected univariate models of NS (fitted to the data of Table 2, with the M+ computer scripts of Appendices 1 and 2)

| Univariate parameters | Model 0: zero age effects at both Waves | Model 1: free age effects at both Waves | Model 2: equal age effects at both Waves | Model 3: using only NS with level only | Model 4: using only NS with level and slope |
|---|---|---|---|---|---|
| Intercept (at age 65) for NS (z) | =0 | =0 | =0 | 2.53 (33.0) | −1.36 (0.75) |
| Intercept (at age 65) for NS[1] (z) | 2.39 (38.7) | 2.35 (29.4) | 2.76 (33.2) | =0 | =0 |
| Intercept (at age 65) for NS[3] (z) | 2.71 (29.5) | 2.76 (31.9) | 2.76 (33.2) | =0 | =0 |
| Mean slope (per decade of age) for NS (z) | =0 | =0 | =0 | =0 | 3.79 (3.8) |
| Mean slope (per decade of age) for NS[1] (z) | =0 | −0.81 (10.7) | −0.89 (13.2) | =0 | =0 |
| Mean slope (per decade of age) for NS[3] (z) | =0 | −0.98 (11.8) | −0.89 (13.2) | =0 | =0 |
| −ll/free p | −5369/5 | −5469/7 | −5469/4 | −230243/2 with $\psi^2$=.01 fixed | −8539/5 |
| BIC | 10737 | 10589 | 10587 | 4602300 | 17113 |

Note: The complete baseline model of no level and no changes with all four measured variables has $\chi^2/df = 3034/28$ with $-2\text{ll} = -14327$

**Table 4** Numerical results for selected multivariate models (fitted to the data of Table 2, with the M+ computer scripts of Appendices 3 and 4)

| Multivariate parameters | Model 0: using only G with age level only at each Wave | Model 1: using only G with equal age level and slope at each Wave | Model 2: one common factor of G with level only | Model 3: one common factor of G with levels and slopes | Model 4: two common factors of *GF* and *Gr* with both levels and slopes |
|---|---|---|---|---|---|
| Mean level (at age 65) for G[1] (z) | 2.66 (44.7) | 2.67 (45.4) | =0 | =0 | =0 |
| Mean level (at age 65) for G[2] (z) | 2.66 (44.7) | 2.67 (45.4) | =0 | =0 | =0 |
| NS | =0 | =0 | =0 | =0 | =0 |
| NU | =0 | =0 | =0 | =0 | =0 |
| IR | =0 | =0 | =0 | =0 | =0 |
| DR | =0 | =0 | =0 | =0 | =0 |
| *G (z)* | =0 | =0 | 2.55 (42.0) | 2.68 (43.0) | =0 |
| *Gf (z)* | =0 | =0 | =0 | =0 | 2.75 (46.0) |
| *Gr (z)* | =0 | =0 | =0 | =0 | 6.04 (155.4) |
| Mean slope (per decade of age) for G[1] (z) | =0 (=0) | −0.26 (15.0) | =0 | =0 | =0 |
| Mean slope (per decade of age) for G[3] (z) | =0 (=0) | −0.26 (15.0) | =0 | =0 | =0 |
| NS | =0 | =0 | =0 | =0 | =0 |
| NU | =0 | =0 | =0 | =0 | =0 |
| IR | =0 | =0 | =0 | =0 | =0 |
| DR | =0 | =0 | =0 | =0 | =0 |
| *G (z)* | =0 | =0 | =0 | −0.25 (13.0) | =0 |
| *Gf (z)* | =0 | =0 | =0 | =0 | −0.28 (9.3) |
| *Gr (z)* | =0 | =0 | =0 | =0 | −0.50 (12.3) |
| Factor loading for NS | =1.00 | =1.00 | =1.00 | =1.00 | =1.00 |
| NU | 0.75 | 0.75 | 0.77 | 0.75 | 0.74 |
| IR | 2.26 | 2.25 | 2.33 | 2.25 | =1.00 |
| DR | 1.91 | 1.91 | 1.97 | 1.91 | 0.84 |
| −ll/free p | −15288//10 | −15288/11 | −18287/10 | −15973/13 | −15560/22 |
| BIC | 32263 | 32046 | 38060 | 32038 | 31275 |

**Table 5** Parameter estimates for the single-variable (number series) model with age changes

```
TESTS OF MODEL FIT

Loglikelihood

          H0 Value                       -8539.069

Information Criteria

          Number of Free Parameters              5
          Akaike (AIC)                   17088.137
          Bayesian (BIC)                 17113.265
          Sample-Size Adjusted BIC       17097.384
            (n* = (n + 2) / 24)


MODEL RESULTS

                                                    Two-Tailed
                    Estimate       S.E.  Est./S.E.    P-Value

 L        BY
    NS[1]              1.000      0.000    999.000    999.000
    NS[3]              1.000      0.000    999.000    999.000

 S        BY
    NS[1]            999.000      0.000    999.000    999.000
    NS[3]            999.000      0.000    999.000    999.000

 L        WITH
    S              -1618.371     76.563    -21.138      0.000

 Means
    L                 -1.362      1.827     -0.746      0.456
    S                  3.789      0.977      3.879      0.000

 Intercepts
    NS[1]              0.000      0.000    999.000    999.000
    NS[3]              0.000      0.000    999.000    999.000

 Variances
    L               3694.429    156.884     23.549      0.000
    S               1055.135     44.750     23.579      0.000

 Residual Variances
    NS[1]              0.000      0.000    999.000    999.000
    NS[3]              0.000      0.000    999.000    999.000
```

Note: See text for details. The use of 999.000 is used for empty locations and when the individual varies. This is Mplus 7.11 output from the Appendix 2 program and the Mplus computer script that was used is in Appendix 2. We can see the BIC is a lot larger for the no-changes model (#2), but we do not assume that Age changes modeled in this way are important. This is only one latent variable after all.

A specially selected (and highly restricted) two-factor model alternative is listed as #4. This is a more complex two-factor model that allows for more covariation among the measures but also requires "strict" invariance of the loadings and unique variances over time. Similarly, all variable intercepts are set to zero here, so the mean changes have to go through the common factors (as in McArdle & Nesselroade 2014). In this approach the model has two common factors with $\lambda = 1$ the required fixed loading for *NS* and *IR* and estimated loadings of 0.74 (for *NU*) and 0.84 (for *DR*). This model has 22 parameters, largely due to the extra common factor covariances, and these extra parameters are penalized heavily by the BIC, including the level intercepts of 2.82 (or $\mathrm{W} = 528$) and 6.04 (for *IR*). This still seems to have the smallest BIC value, so it could be chosen on this basis.

It appears that *Gr* as measured by *IR* and *DR* declines the most over age, with −0.49 per age decade. The function termed *Gf*, indicated by *NS* and *NU*, decline significantly but only at −0.27 per decade. In contrast to the one-factor version, the BIC for this model is smaller, and needless to say, this is far less decline than we initially expected from a normal aging population, so maybe we do not have the right factor yet. This has the smallest BIC of all used, so it could be chosen as the best model for the data. But this BIC (about 31,275) is not much smaller than the prior BIC (about 32,038), and the two-factor model has a lot of extra parameters so this substantive model is not yet considered a large improvement.

## Discussion

The final model chosen did not have the smallest BIC but it seemed to fit the data the best. So there is much more to be done here. We fully realize the two-factor model did not fit as well as we would have liked. We would have liked to separate apart the aspects of $G_F$ and $G_R$ but this proved difficult with only 4 measures. But this is a complicated choice (made here only by BIC) and because the number of measurements is small and we want to say a lot about aging. That is, we had a difficult distinction because we were only working with four variables. But this SEM is a useful starting point even if it only deals with LCM Assumption 6.

We tried to use the standard HRS procedures (e.g., Genesys surveys) to contact households with some persons in the HRS age range (over 50). We succeeded in reaching over 3,000 people, but not everyone agreed to test further. We can only suggest the reader look carefully at Heeringa et al. (2011) and McArdle and Fisher (2015) for details. At Wave 1 we actually had measured over 1,500 people we have measured age and other demographics (like respondent education, sex, minority status, health, dyad, nursing home, currently employed), as well as administration of a telephone versions of standard cognitive tests (the TICS; defined here as *BC*, *S7*, *IR*, *DR*), as well as some additional tests HRS (*NU*) and some new telephone

**Table 6** Parameter estimates for the current invariant measurement model with two common factor of changes

```
THE MODEL ESTIMATION TERMINATED NORMALLY

TESTS OF MODEL FIT

Loglikelihood

          H0 Value                      -15973.231

Information Criteria

          Number of Free Parameters             13
          Akaike (AIC)                   31972.461
          Bayesian (BIC)                 32037.793
          Sample-Size Adjusted BIC       31996.502
            (n* = (n + 2) / 24)

MODEL RESULTS
                                                    Two-Tailed
                    Estimate       S.E.  Est./S.E.    P-Value

 G[1]     BY
    NS[1]              1.000      0.000    999.000    999.000
    NU[1]              0.748      0.017     43.113      0.000
    IR[1]              2.250      0.048     46.838      0.000
    DR[1]              1.907      0.042     45.848      0.000

 G[3]     BY
    NS[3]              1.000      0.000    999.000    999.000
    NU[3]              0.748      0.017     43.113      0.000
    IR[3]              2.250      0.048     46.838      0.000
    DR[3]              1.907      0.042     45.848      0.000

 L_G      BY
    G[1]               1.000      0.000    999.000    999.000
    G[3]               1.000      0.000    999.000    999.000
    G[1]               1.000      0.000    999.000    999.000
    G[3]               1.000      0.000    999.000    999.000

 S_G      BY
    G[1]             999.000      0.000    999.000    999.000
    G[3]             999.000      0.000    999.000    999.000
    G[1]             999.000      0.000    999.000    999.000
    G[3]             999.000      0.000    999.000    999.000

 L_G      WITH
    S_G                0.026      0.008      3.283      0.001

 Means
    L_G                2.673      0.059     45.338      0.000
    S_G               -0.245      0.019    -13.001      0.000
```

(continued)

(continued)

```
 Intercepts
    NS[1]                0.000       0.000     999.000     999.000
    NS[3]                0.000       0.000     999.000     999.000
    NU[1]                0.000       0.000     999.000     999.000
    NU[3]                0.000       0.000     999.000     999.000
    IR[1]                0.000       0.000     999.000     999.000
    IR[3]                0.000       0.000     999.000     999.000
    DR[1]                0.000       0.000     999.000     999.000
    DR[3]                0.000       0.000     999.000     999.000

 Variances
    L_G                  0.151       0.022       6.929       0.000
    S_G                  0.030       0.015       1.960       0.050

 Residual Variances
    NS[1]                7.509       0.228      32.890       0.000
    NS[3]                7.509       0.228      32.890       0.000
    NU[1]                0.872       0.028      31.166       0.000
    NU[3]                0.872       0.028      31.166       0.000
    IR[1]                0.378       0.045       8.469       0.000
    IR[3]                0.378       0.045       8.469       0.000
    DR[1]                1.556       0.055      28.213       0.000
    DR[3]                1.556       0.055      28.213       0.000
    G[1]                 0.228       0.017      13.260       0.000
    G[3]                 0.228       0.017      13.260       0.000
```

Note: This is Mplus 7.11 program output for the input in Appendix 4

administered WJ-based adaptive tests (*NS*, *RF*) and the depression scale (*CES-D*). Our main goal here was to see if we could measure the same constructs as before, but the time over the telephone, and we basically found we could. But we are also involved in a number of selection issues (see McArdle 2013; Heeringa et al. 2011), and a small sample of 200 HRS respondents who were at the top and the bottom of the HRS Cognition scores (in 2008) were also re-measured on our instruments.

Perhaps with more measured variables we can also make finer distinctions among multiple factors. Or perhaps we can take into account the non-normality of the data (see Fig. 2a–d). In this same sense, we were evaluating only part of the measured scale and not all of it. For another example, take our previous analytic work in McArdle et al. (2015). In this research we evaluated 13 items from the CES-D. If we had been evaluating the CES-D for use in these grades (9th and 10th) we would use 20 items as listed in the typical CES-D. But we were most interested in evaluation the concept of depression, the latent factor that it represented, and we thought we could test this idea with only four items. But we looked at 13 items in this chapter only, mainly because we were trying to indicate one factor only.

Indeed an expansion or delineation of the purported common factors can be achieved with measured variables such as *S7* (a telephone administration of HRS Serial Sevens) and *RF* (a telephone administration of WJ-Retrieval Fluency). Each of these could expand the factor space in important ways and may lead to some stability. We can also add several occasions of measurement, before and after Wave 3. This use of Wave 2 was a few weeks later than Wave 1 but we gave them a full battery of tests, about 1,200 people for 3 h in the home, including the WJ-R, and the WASI, plus some *personality scales* (e.g., *BFI*, *NCS*) and some *dispositional measures* (*RISK*). This was mainly included so we could verify the telephone measures against telephone adapted tests and several were administered here. There are many measures here and they can be useful too.

At Wave 3 we went back to the original telephone forms after an average of about 1.21 years (with 0.21 as a standard deviation). This was supposedly different for people of different ages in some time lag but other conditions need to be stated up-front as well. For example, Rodgers et al. (2003) and McArdle and Fisher (2015) make it very clear that there we naturally had eight groups of respondents by time-lag (this was not designed) and it was hard to do a second telephone test in a short time if the people had not completed the *FTF* of Wave 2. This confound does not apply to any people selected but it is there. We started this testing by verifying the Birth Date of the person being tested (the interviewee, or IV). We repeatedly tested the original people again on the same battery used in Wave 1 with, as stated, about a year and a half delay (e.g., Wave 3).

During the last 5 years we have measured the same people again as part of a second 5-year study. During these times (2009–2014) we were mainly interested in the difference (if there are any) between Telephone testing and Internet testing. This issue will not be raised or resolved here. There are currently (as of 2015) no plans by us to re-measure these same people again but we have let the CogECON and HRS teams at the University of Michigan contact them. In any case, our experiment is complex, as are our multiple assumptions, but this is the basic model of aging and invariance we will use in further analyses of these data, so comments are welcome.

## Appendix 1: An example of an M+ 7.11 Computer Program for Two Time Point Data with Age Changes

```
TITLE: EX_MOD1 -- Dynamic Impact of constraints NS at
Two Times
Using Age at Testing as a cross-section
Run of NS with No Slope ALL CogUSA scores (McArdle,
2014)
DATA: FILE = CogUSA_Repeated3.dat;
!LISTWISE=ON;
VARIABLE: NAMES =
id
w1_age w2_age w3_age
educ female black
ldays12 ldays23 ldays13
```

```
W1_IR W2_IR W3_IR
W1_DR W2_DR W3_DR
W1_BC W2_BC W3_BC
W1_S7 W2_S7 W3_S7
W1CESDp W2CESDp W3CESDp
W1_NU W2_NU W3_NU
W1_NS W2_NS W3_NS
W1_RF W2_RF W3_RF
;
USEVAR = W1_Age W3_Age W1_NS W3_NS;
MISSING=.;
DEFINE: W1_NS = (W1_NS - 500)/10;
W3_NS = (W3_NS - 500)/10;
W1_age = (W1_age - 50)/10;
W3_age = (W3_age - 50)/10;
ANALYSIS: TYPE=MEANSTRUCTURE;
MODEL:
W1_NS on W1_age (B1);
W3_NS On W3_age (B3);
!equating error variances
! W1_NS W3_NS (V_U);
OUTPUT: SAMPSTAT STANDARDIZED;
```

## Appendix 2: An example of an M+ 7.11 Computer Program for Two Time Point Data with Age Changes as a Loading Constraint

```
TITLE: EX_Mod: Table 3.2 -- Dynamic Impact of
constraints NS by Two Times
Run of All TEL Time CogUSA scores (McArdle, 2014)
DATA: FILE = CogUSA_Repeated3.dat;
!LISTWISE=ON;
VARIABLE: NAMES =
id
w1_age w2_age w3_age
educ female black
ldays12 ldays23 ldays13
W1_IR W2_IR W3_IR
```

```
W1_DR W2_DR W3_DR
W1_BC W2_BC W3_BC
W1_S7 W2_S7 W3_S7
W1CESDp W2CESDp W3CESDp
W1_NU W2_NU W3_NU
W1_NS W2_NS W3_NS
W1_RF W2_RF W3_RF;
CONSTRAINT = W1_Age W3_Age;
USEVAR = W1_NS W3_NS;
MISSING=.;
DEFINE: W1_NS = (W1_NS - 500)/10;
W3_NS = (W3_NS - 500)/10;
ANALYSIS: TYPE=MEANSTRUCTURE;
MODEL:
!frst get a level and a slope for NS;
l BY W1_NS@1 W3_NS@1;
s BY W1_NS * (LNS1)
W3_NS * (LNS3);
[l s];
l s;
l WITH s;
!eliminating original variables
[W1_NS@0 W3_NS@0];
W1_NS@0 W3_NS@0;
!equating error variances
! W1_NS W3_NS (V_U);
MODEL CONSTRAINT: ! To get at individual loadings;
LNS1 = (W1_age - 50)/10;
LNS3 = (W3_age - 50)/10;
OUTPUT: SAMPSTAT STANDARDIZED;
```

## Appendix 3: M+ 7.11 Computer Program for Four Variables at Two Time Points of Data based on a Model of One Common Factor with Age Changes

```
EX_MOD: Table 4.2 -- Multivariate Dynamic Impact of
Equal Age
```

```
For four variables NS, NU, IR,DR, with and MFIT for G
(McArdle, 2015)
DATA: FILE = CogUSA_Repeated3.dat; !LISTWISE=ON;
VARIABLE: NAMES = id
w1_age w2_age w3_age
educ female black
ldays12 ldays23 ldays13
W1_IR W2_IR W3_IR W1_DR W2_DR W3_DR
W1_BC W2_BC W3_BC W1_S7 W2_S7 W3_S7
W1CESDp W2CESDp W3CESDp
W1_NU W2_NU W3_NU W1_NS W2_NS W3_NS
W1_RF W2_RF W3_RF
;
USEVAR = W1_Age W3_Age
W1_NS W3_NS W1_NU W3_NU
W1_IR W3_IR W1_DR W3_DR
;
MISSING=.;
DEFINE:
W1_NS = (W1_NS - 500)/10;
W3_NS = (W3_NS - 500)/10;
W1_age = (W1_age - 65)/10;
W3_age = (W3_age - 65)/10;
ANALYSIS: TYPE=MEANSTRUCTURE;
MODEL:
W1_G BY W1_NS (L_NS);
W1_G BY W1_NU (L_NU);
W1_G BY W1_IR (L_IR);
w1_G BY W1_DR (L_DR);
W3_G BY W3_NS (L_NS);
W3_G BY W3_NU (L_NU);
W3_G BY W3_IR (L_IR);
W3_G BY W3_DR (L_DR);
W1_G On W1_Age (b11); [W1_G] (b01);
W3_G ON W3_Age (b13); [W3_G] (b03);
!Equal uniquenesses at the factor level
W1_G W3_G (U2_G);
!Equal Uniqueness at the Variable Level
W1_NS W3_NS (U_NS); W1_NU W3_NU (U_NU):
```

```
W1_IR W3_IR (U_IR); W1_DR W3_DR (U_DR);
!original means of variables not used
[W1_NS@0 W3_NS@0 W1_NU@0 W3_NU@0 W1_IR@0 W3_IR@0
W1_DR@0 W3_DR@0];
OUTPUT: SAMPSTAT STANDARDIZED;
```

## Appendix 4: M+ 7.11 Computer Program for Four Variables at Two Time Points of Data Based on a Model of Two Common Factors with Age Changes as a Loading Constraint

```
TITLE: EX_MOD: TABLE 4.4 -- Dynamic Impact of
constraints on
four variables NS,NU, IR,DR, with am MFIT-G
Run of 4 TEL Time CogUSA scores (McArdle, 2015)
DATA: FILE = CogUSA_Repeated3.dat;
!LISTWISE=ON;
VARIABLE: NAMES =
id
w1_age w2_age w3_age
educ female black
ldays12 ldays23 ldays13
W1_IR W2_IR W3_IR
W1_DR W2_DR W3_DR
W1_BC W2_BC W3_BC
W1_S7 W2_S7 W3_S7
W1CESDp W2CESDp W3CESDp
W1_NU W2_NU W3_NU
W1_NS W2_NS W3_NS
W1_RF W2_RF W3_RF
;
CONSTRAINT = W1_Age W3_Age;
USEVAR = W1_NS W3_NS W1_NU W3_NU W1_IR W3_IR W1_DR
W3_DR;
MISSING=.;
DEFINE:
!Imputation just to keep the time 1 people in at all
occasions...
IF (W3_age LT 0) THEN W3_age=999;
```

```
W1_NS = (W1_NS - 500)/10;
W3_NS = (W3_NS - 500)/10;
ANALYSIS: TYPE=MEANSTRUCTURE;
MODEL:
W1_G BY W1_NS (L_NS);
W1_G BY W1_NU (L_NU);
W1_G BY W1_IR (L_IR);
W1_G BY W1_DR (L_DR);
W3_G BY W3_NS (L_NS);
W3_G BY W3_NU (L_NU);
W3_G BY W3_IR (L_IR);
W3_G BY W3_DR (L_DR);
!get a level and a slope for G;
l_G BY W1_G@1 W3_G@1;
s_G BY W1_G * (L1);
s_G BY W3_G * (L3);
[l_G s_G]; l_G s_G; l_G WITH s_G;
!But with equal error variances
W1_G W3_G (U_G);
!original variables not used
[W1_NS@0 W3_NS@0];
W1_NS@0 W3_NS@0;
!Except equal error variances
W1_NS W3_NS (U_NS);
!original variables not used
[W1_NU@0 W3_NU@0];
W1_NU@0 W3_NU@0;
!Except equal error variances
W1_NU W3_NU (U_NU);
!original variables not used
[W1_IR@0 W3_IR@0];
W1_IR@0 W3_IR@0;
!Except Equal error variances
W1_IR W3_IR (U_IR):
!original variables not used
[W1_DR@0 W3_DR@0];
W1_DR@0 W3_DR@0;
!Except equal error variances
W1_DR W3_DR (U_DR);
```

```
MODEL CONSTRAINT: ! To get at individual loadings;
L1 = (W1_age - 65)/10;
L3 = (W3_age - 65)/10;
OUTPUT: SAMPSTAT STANDARDIZED;
```

## References

Baltes, P. B., & Nesselroade, J. R. (1979). History and rationale of longitudinal research. In J. R. Nesseloade & P. B. Baltes (Eds.), *Longitudinal research in the study of behavior and development* (pp. 1–39). New York: Academic.

Blair, C. (2006). How similar are fluid cognition and general intelligence? A developmental neuroscience perspective on fluid cognition as an aspect of human cognitive ability. *Behavioral and Brain Sciences, 29*(02), 109–125.

Brandmaier, A. M., von Oertzen, T., McArdle, J. J., & Lindenberger, U. (2012). Exploratory data mining with structural equation model trees (Chapter 4). In J. J. McArdle & G. Ritschard (Eds.), *Contemporary issues in exploratory data mining*. New York: Taylor & Frances.

Brandmaier, A.M., von Oertzen, T., McArdle, J.J. & Lindenberger, U. (2013). Structural Equation Model Trees. *Psychological Methods, 18*, 71–86 (doi: 10.1037/a0030001).

Bryk, A.S. & Raudenbush, S.W. (1992). *Hierarchical linear models: Applications and data analysis methods.* Newbury Park, CA: Sage Pub. Inc.

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105*(3), 456.

Bonate, P. L. (2000). *Analysis of pretest-posttest designs*. Boca Raton, FL: Chapman & Hall.

Cribbie, R. A., & Jamieson, J. (2004). Decreases in posttest variance and the measurement of change. *Methods of Psychological Research Online, 9*(1), 37–55.

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

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Boston: Houghton-Mifflin.

Cromwell, J. B., Labys, W. C., & Terraza, M. (1994). *Univariate tests for time series models*. Thousand Oaks, CA: Sage.

Davidson, M. L. (1972). Univariate versus multivariate tests in repeated-measures experiments. *Psychological Bulletin, 77*, 446–452.

Fisher, G. G., Hassan, H., Rodgers, W. L., & Weir, D. R. (2013). *Health and retirement study imputation of cognitive functioning measures: 1992-2010 (Final Release Version) data description*. Ann Arbor: University of Michigan, Survey Research Center.

Fisher, R. A. (1925). *Statistical methods for research workers*. Edinburgh: Oliver & Boyd.

Fox, J. (1999). *Applied regression analysis, linear models, and related methods*. Thousand Oaks, CA: Sage.

Grimm, K. J., & Widaman, K. F. (2010). Residual structures in latent growth curve analysis. *Structural Equation Modeling, 17*, 424–442.

Heeringa, S. G., Berglund, P. A., & Khan, A. (2011). *Sampling error estimation in design-based analysis of the PSID data.* Institute for Social Research: University of Michigan Survey Research Center. Retrieved from http://psidonline.isr.umich.edu/Publications/Papers/tsp/2011-05_Heeringa_Berglung_Khan.pdf

Hilbe, J. M. (2011). *Negative binomial regression.* Cambridge, MA: Cambridge University Press.

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

Hishinuma, E., Chang, J., McArdle, J. J., & Hamagami, A. (2012). Potential causal relationship between depressive symptoms and academic achievement in the Hawaiian High Schools Health Survey using contemporary longitudinal latent variable change models. *Developmental Psychology, 48*(5), 1327–1342.

Huynh, H., & Feldt, L. S. (1976). Estimation of the box correction from degrees of freedom from sample data in the randomized block and split plot design. *Journal of Educational Statistics, 1*, 69–82.

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika, 36*, 409–426.

Jöreskog, K. G. (1974). Analyzing psychological data by structural analysis of covariance matrices. In D. Krantz et al. (Eds.), *Contemporary developments in mathematical psychology—Volume II*. San Francisco, CA: W.H. Freeman Co.

Juster, F.T., & Suzman, R. (1995). The health and retirement study: An overview. HRS Working Papers Series #94-1001. Journal of Human Resources, 1995 Supplement (JHR 30-S).

Lachman, M. E., & Spiro, A. (2002). Critique of cognitive measures in the Health Retirement Study (HRS) and the Asset and Health Dynamics among the Oldest Old (AHEAD) study. *HRS Data Monitoring Board Report,* November 2002.

Lange, K., Westlake, J., & Spence, M. A. (1976). Extensions to pedigree analysis: III. Variance component by the scoring method. *Annals of Human Genetics, 39*, 485–491.

Loeber, R., & Farrington, D. P. (1994). Problems and solutions in longitudinal and experimental treatment studies of child psychopathology and delinquency. *Journal of consulting and clinical psychology, 62*(5), 887.

McArdle, J. J. (1986). Latent variable growth within behavior genetic models. *Behavior Genetics, 16*(1), 163–200.

McArdle, J.J. (1988). Dynamic but structural equation modeling of repeated measures data. In J.R. Nesselroade & R.B. Cattell (Eds.), *The Handbook of Multivariate Experimental Psychology,* New York, Plenum Press, *2*, 561–614.

McArdle, J. J. (1998). Modeling longitudinal data by latent growth curve methods. In G. Marcoulides (Ed.), *Modern methods for business research* (pp. 359–406). Mahwah, NJ: Lawrence Erlbaum Associates.

McArdle, J. J. (2007). Five steps in the structural factor analysis of longitudinal data. In R. Cudeck & R. MacCallum (Eds.), *Factor analysis at 100 years* (pp. 99–130). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

McArdle, J. J. (2008). Latent variable modeling of longitudinal data. *Annual Review of Psychology, 60*, 577–605.

McArdle, J. J. (2010). Contemporary challenges of longitudinal measurement using HRS data. In G. Walford, E. Tucker, & M. Viswanathan (Eds.), *The SAGE handbook of measurement* (pp. 509–536). London: SAGE Press.

McArdle, J. J. (2011). Latent curve modeling. In R. Hoyle (Ed.), *Handbook of structural equation modeling*. NY: Oxford.

McArdle, J. J. (2012). Testing the idea of general intelligence. *F&M Scientist, 1*, 27–66.

McArdle, J. J. (2013). Dealing with longitudinal attrition using logistic regression and decision tree analyses. In J. J. McArdle & G. Ritschard (Eds.), *Contemporary issues in exploratory data mining in the behavioral sciences* (pp. 282–311). New York: Taylor & Frances.

McArdle, J. J., & Cattell, R. B. (1994). Structural equation models of factorial invariance in parallel proportional profiles and oblique contactor problems. *Multivariate Behavioral Research, 29*(1), 63–113.

McArdle, J. J., & Epstein, D. B. (1987). Latent growth curves within developmental structural equation models. *Child Development, 58*(1), 110–133.

McArdle, J. J., & Fisher, G.G. (2015). *New analyses of longitudinal cognitive data from The CogUSA survey.*

McArdle, J. J., Fisher, G. G., & Kadlec, K. M. (2007). Latent variable analysis of age trends in tests of cognitive ability in the health and retirement survey, 1992-2004. *Psychology and Aging, 22*(3), 525–545.

McArdle, J. J., & Hamagami, F. (1996). Multilevel models from a multiple group structural equation perspective. In G. Marcoulides & R. Schumacker (Eds.), *Advanced structural equation modeling techniques* (pp. 89–124). Hillsdale, NJ: Erlbaum.

McArdle, J. J., Hishinuma, E., Chang, J., & Hamagami, A. (2014). Longitudinal analyses of the dynamic relationships among depression and academic achievement from the Hawaiian High Schools Health Survey. *Structural Equation Modeling, 21*(4), 608–629.

McArdle, J. J., & Nesselroade, J. R. (2014). *Longitudinal data analysis using structural equation models*. Washington, DC: APA Books.

McArdle, J. J., Petway, K. T., & Hishiuma, E. S. (2015). IRT for growth and change (chap. 20, p. 435–456. In S. P. Reise & D. A. Revicki (Eds.), *Handbook of item response theory modeling*. New York: Routledge.

McArdle, J.J. & Woodcock, J.R. (1997). Expanding test-rest designs to include developmental time-lag components. *Psychological Methods, 2* (4), 403–435.

McArdle, J.J. & Prescott, C.A. (1992). Age-based construct validation using structural equation modeling. *Experimental Aging Research, 18* (3), 87–115.

McCall, R. B., Appelbaum, M. I., & Hogarty, P. S. (1973). Developmental changes in mental performance. *Monographs of the Society for Research in Child Development, 38*, 1–84.

McGaw, B., & Joreskog, K. G. (1971). Factorial invariance of ability measures in groups differing in intelligence and socioeconomic status. *British Journal of Mathematical and Statistical Psychology, 24*, 154–168.

Meredith, W. (1964). Notes on factorial invariance. *Psychometrika, 29*(2), 177–185.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*(4), 525–543.

Meredith, W., & Tisak, J. (1990). Latent curve analysis. *Psychometrica, 55*, 107–122.

Mehta, P. D., & Neale, M. C. (2005). People are variables too: multilevel structural equations modeling. *Psychological Methods, 10*(3), 259.

Mehta, P. D., & West, S. G. (2000). Putting the individual back into individual growth curves. *Psychological Methods, 5*(1), 23.

Muthén, B., & Christoffeson, A. (1981). Simultaneous factor analysis of dichotomous variables in several groups. *Psychometrika, 46*, 407–419.

Muthén, B., & Sheddon, K. (1999). Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics, 55*, 463–469.

Muthén, B., & Muthén, L. (2012). *The Mplus 7.0 computer program for structural equation modeling*. Santa Monica, CA: Muthén & Muthén Publishing.

Nagin, D. S. (1999). Analyzing developmental trajectories: Semi-parametric, group-based approach. *Psychological Methods, 4*, 139–177.

Nagin, D. S. (2005). *Group-based modeling of development*. Cambridge, MA: Harvard University Press.

Neale, M.C. (1993). *Mx Statistical Modeling.* Unpublished Manuscript, Virginia Commonwealth University, Richmond, VA

Neale, M.C., Boker, S. M., Xie, G., & Maes, H. H. (1999). Mx statistical modeling (5th ed.). Unpublished program manual, Virginia Institute of Psychiatric and Behavioral Genetics, Medical College of Virginia, Virginia Commonwealth University, Richmond, VA.

Rao, C. R. (1958). Some statistical methods for the comparison of growth curves. *Biometrics, 15*, 1–17.

Raftery, A. E. (1996). Approximate Bayes factors and accounting for model uncertainty in generalised linear models. *Biometrika, 83*(2), 251–266.

Reise, S. P., Widaman, K. F., & Pugh, P. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin, 114*(2), 552–566.

Rodgers, W., Ofstedal, M. B., & Herzog, A. R. (2003). Trends in scores on tests of cognitive ability in the elderly U.S. population: 1993-2000. *Journals of Gerontology: Social Sciences, 52B*(6), S338–S346.

Rogosa, D. R. (1988). Myths about longitudinal research. In K. W. Schaie, R. T. Campbell, W. Meredith, & S. C. Rawlings (Eds.), *Methodological issues in aging research* (pp. 171–209). New York: Springer.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton-Mifflin.

Tu, Y. K., Blance, A., Clerehugh, V., & Gilthorpe, M. S. (2005). Statistical power for analyses of changes in randomized controlled trials. *Journal of Dental Research, 84*(3), 283–287.

Tucker, L. R. (1958). Determination of parameters of a functional relation by factor analysis. *Psychometrika, 23*, 19–23.

Tucker, L. R. (1960). *Determination of generalized learning curves by factor analysis*. New Jersey: Princeton University.

Tucker, L. R. (1966). Learning theory and multivariate experiment: Illustration by determination of generalized learning curves. In R. B. Cattell (Ed.), *Handbook of multivariate experimental psychology* (pp. 476–501). Chicago: Rand McNally.

Tucker, L. R. (1992). Remarks on the studies of the variety of individuals. *Multivariate Behavioral Research, 27*(4), 635–647.

Wang, L., Zhang, Z., McArdle, J. J., & Salthouse, T. A. (2008). Investigating ceiling effects in longitudinal data analysis. *Multivariate Behavioral Research, 43*(3), 476–496.