Published in final edited form as:

Dev Psychol. 2018 November ; 54(11): 2193–2206. doi:10.1037/dev0000583.

Identifying atypical change at the individual level from childhood to adolescence

Eduardo Estrada¹, Emilio Ferrer¹, Bennett A. Shaywitz^{2,3,4}, John M. Holahan^{2,3}, and Sally E. Shaywitz^{2,3}

¹Department of Psychology, University of California, Davis

²Department of Pediatrics, Yale University School of Medicine

³Yale University School of Medicine

⁴Department of Neurology, Yale University School of Medicine

Abstract

Identifying change at the individual level is an important goal for researchers, educators, and clinicians. We present a set of statistical procedures for identifying individuals who depart from a normative change. Using Latent Change Scores models (*LCS*), we illustrate how the Individual Likelihood computed from a statistical model for change (*IL*) and from an alternative unrestricted model (*ILsat*) can be used to identify atypical trajectories in situations with several measurement occasions. Using *LCS* and linear regression, we also show how the observed and latent change residuals can be used to identify atypical individual change between two measurement occasions. We apply these methods to a measure of general verbal ability (from *WISC-R*), from a large sample of individuals assessed every two years from grade 1 to 9. We demonstrate the efficiency of these techniques, illustrate their use to identify individual change in longitudinal data, and discuss potential applications in developmental research.

Keywords

Reading abilities; Dyslexia; Individual change; Longitudinal data analysis; Structural equation modeling

Identifying individuals who do not change at the expected pace is an important goal for researchers, educators, and clinicians. A physician, for example, would want to identify patients who are not responding to a given treatment. Similarly, a teacher would be interested in determining whether the increase in reading proficiency displayed by a particular student from one year to the next is adequate for his age, or is falling behind a reference group. Both these questions imply identifying each individual as following a *typical* or an *atypical* change. To determine the magnitude of change that can be considered atypical, individual change can be compared to the change observed in a normative group. By doing this, the researcher can establish criteria for atypical change that either does not reach or exceeds the expected norm.

In developmental and educational settings, the processes of interest are often examined as they unfold over time (Lonigan, Burgess, & Anthony, 2000). For example, cognitive abilities such as reading, verbal skills, or mathematical reasoning are processes that change over time. In some scenarios, the goal is to detect individuals who change above a normative group (e.g., detecting individuals with high capacities, Crain-Thoreson & Dale, 1992). Yet, other situations focus on detecting individuals who change below typical thresholds –e.g., not improving as much as they should (Pennington & Lefly, 2001; Torppa, Poikkeus, Laakso, Eklund, & Lyytinen, 2006; Tressoldi, Stella, & Faggella, 2001).

In education, federal and state laws mandate accountability and demand that schools demonstrate that they are teaching children effectively. But just how is one to determine if a child is improving enough from year to year? More specifically, how can educators know which children are not changing according to the group norm and should receive an intervention?

Different authors have pointed out the importance of studying change at the *individual* level (Hamaker, 2012; Molenaar, 2004; Nesselroade, Gerstorf, Hardy, & Ram, 2007; Voelkle, Brose, Schmiedek, & Lindenberger, 2014). However, most research in developmental psychology is still conducted to address questions at the *group* level. In this paper we propose several statistical procedures that allow using information from the entire group to identify individuals whose developmental trajectories depart from that of the group. We apply these procedures to a large and representative sample of individuals who were assessed on verbal ability at five occasions (every 2 years, from grade 1 to grade 9).

Our goal in this paper is twofold. First, we capitalize on this large sample to investigate the efficacy of a set of statistical procedures for correctly identifying individuals as following atypical change. Second, we illustrate how these procedures can be implemented in empirical research.

Identifying individual change: Trajectories vs. pre-post changes

When investigating individual change in developmental phenomena, at least two different approaches can be implemented. The first approach involves studying individuals' trajectories over time and identifying the characteristics of normative development. The second approach involves measuring the change between two time-points and identifying individuals who deviate from the typical magnitude of change, either above or below the expected normative range.

The following example illustrates the difference between these two perspectives. Suppose reading ability was measured in a sample of individuals every two years from grade 1 to grade 11. The scores were scaled so the mean and standard deviation at grade 1 were 100 and 15, respectively. The scores were not age-normed, so are expected to show yearly improvements. The means at each grade (1, 3, 5, 7, 9 and 11) are 100, 135, 161, 181, 169 and 207 points, respectively. This mean trajectory is represented in line *a* of Figure 1. Suppose one child had a medical leave during grades 1 and 2, and could not attend school or receive compensatory instruction. As a consequence, her reading ability is atypically low at

grade 3. However, from that moment on, she is able to receive instruction and catch up with the group in subsequent years. In other words, although she experienced an atypical change from grade 1 to 3, her overall trajectory can be considered typical. This illustrative example is represented by line *b* in Figure 1. Now suppose another child (line *c* in Figure 1) showed a low rate of improvement, consistently from grade 1 to 11. In this case, the change from grade 1 to 3 might not be low enough to be considered atypical, but the entire trajectory certainly is.

A standard statistical approach intended to capture the characteristics of the full trajectory is not adequate for detecting the atypical change from grade 1 to 3 shown in line *b*. That particular trajectory might or might not be identified as atypical depending on the number of time points and the extent to which the change from grade 1 to 3 is atypical and affects the estimates for the whole trajectory. On the other hand, focusing only on change from grade 1 to 3 may lead to considering line *c* as typical, if the amount of change between grades 1 and 3 does not depart enough from the reference group. In other words, the two approaches – trajectories vs. two-occasion changes– provide different information and do not necessarily lead to similar conclusions regarding who is identified as typical, especially if the trajectory is comprised of many time-points and atypical change occurs between two of such points only. Next, we describe statistical tools that can be applied to study individual change from both perspectives.

Identifying atypical trajectories in repeated measures data

When data involve multiple measurement occasions, a reasonable data analytic approach is to examine the trajectory using a longitudinal model of change –e.g., a latent growth curve, or a latent change score model. This approach is useful because it allows modeling the information from the repeated measures using a single mathematical representation. It also permits testing complex hypothesis about change, modeling systems with more than one process –e.g., development of reading ability and abstract reasoning, and comparing alternative models of change, each one representing a different theory (E. Ferrer & McArdle, 2003; McArdle, 2009).

One flexible model for characterizing developmental change is the latent change score model (LCS; McArdle, 2001; McArdle & Hamagami, 2001). A path diagram of this model for five measurement occasions is depicted in Figure 2¹. A LCS model represents the process of interest as a dynamical system in which the changes, instead of the levels, are the focus. At each repeated occasion, a latent variable representing changes (Δy_t) in a latent process (y_t) is specified. Thus, at each occasion, the latent process is a function of the initial unobserved level, y_0 , plus the accumulation of changes up to that occasion. In Figure 2, such changes are a function of: a) an additive linear effect captured by the latent variable y_s , and b) a proportional effect from the latent level of the process at the previous occasion – captured by the parameter β . The means of the latent intercept and slope capture, respectively, the mean level in the process at the first occasion, and the average additive

¹Note that the model can be modified to include any number of occasions. The values 1-9 represent school grades from the empirical data to be described in later sections.

component at every repeated occasion. The variances of these latent variables denote individual differences in such initial level and additive change.

When modeling development in cognitive abilities and achievement during school years, the linear effect is typically positive. This entails an overall positive trend –i.e., growth– in the scores over time. In turn, the parameter β is negative, representing a dampening effect. Because this negative effect is proportional to the level at the previous occasion, it becomes increasingly stronger over time. The combined effect of the linear and proportional components of change leads to a non-linear trajectory with less overall gains over time. This joint effect is illustrated in the model implied trajectories of Figure 4. For more information on LCS models and their application to developmental change, see McArdle (2001), McArdle & Hamagami (2001), E. Ferrer & McArdle (2004), E. Ferrer et al. (2007), E. Ferrer et al. (2010), E. Ferrer & McArdle (2010), and Kievit et al. (2017).

Although LCS models have been widely used to study various aspects of psychological development (Ahmed, Wagner, & Lopez, 2014; E. Ferrer & McArdle, 2003; Quinn, Wagner, Petscher, & Lopez, 2015), they are seldom used as a tool to extract information at the individual level. Similarly, although many indices exist to evaluate model fit in structural equation modeling (SEM), those designed to examine such fit at the individual level are few and far between in the literature. One of such individual indices is the maximum likelihood estimation for the scores of a case given a structural model fitted to a sample (Arbuckle, 1996; Coffman & Millsap, 2006; Lange, Westlake, & Spence, 1976; Mehta & Neale, 2005; Sterba & Pek, 2012). This statistic requires fitting a statistical model for change –e.g., LCS–, and estimating the *model-implied* means, variances and covariances. It provides a direct estimation of how likely a person's scores are in the context of a particular statistical model and a particular set of such model-implied statistics. It has been termed Case-Wise Likelihood (Arbuckle, 1996), Person Specific Likelihood (Mehta & Neale, 2005), or Individual Likelihood (McArdle, 1998; Mehta & Neale, 2005), and it was suggested early on as a measure of person misfit to identify atypicality (McArdle, 1998). However, it is rarely used in empirical research and, to the best of our knowledge, this is its first application to study the development of cognitive abilities. We refer to this index as the Individual Likelihood (*IL*).

The concept of Individual Likelihood can be applied to our research problem in a different way: The *IL* of a case can be computed based on the *observed statistics* in the sample, instead of the model-implied expectations. This version of the *IL* is conceptually and mathematically equivalent to the *IL* of a saturated –i.e., unrestricted– SEM model fitted to the sample data. We will also compute this index, and will refer to it as *ILsat*. By computing *ILsat*, one can quantify the extent to which the scores of each case are likely to happen in the context of the sample from which the observed statistics were computed. Importantly, unlike the *IL*, the computation of *ILsat* does not require to fit any statistical model to the data. We describe both procedures in detail in the Methods section.

Identifying atypical change between two occasions

Although desirable, collecting data at multiple occasions is not always possible. Often, individuals are measured at two time-points only and information about changes in the scores is used to make decisions –e.g. teachers may want to detect whether children are “falling behind” in, say reading proficiency from first to second grade, and need a reading intervention. In principle, the LCS models described in the previous section can be also used to detect atypical change between two measurement occasions. (McArdle, 2001, 2009; McArdle & Hamagami, 2001; McArdle & Prindle, 2008). Specifically, if the variance of the latent change γ is estimated, individual differences are allowed in the change between occasions. A diagram of this model is represented in Figure 3.

In the model depicted in Figure 3, three latent variables are relevant: the latent trait at the first and second occasions (y_1 and y_2), and the latent change between them (γ). The mean, intercept, and variances of y_1 and change (μ_{y1} , τ_γ , σ^2_{y1} and σ^2_γ) capture the expected value and the individual differences in the level at the first occasion and in the pre-post change. The regression parameter β provides an estimation of the relation between the initial level and the change (a covariance between them would be an alternative specification of the model, McArdle, 2009).

The fact that the model incorporates a measurement error at both occasions, makes it possible to separate error variance from variance in the latent process. However, in order to do this, one additional piece of information is required. In the two-occasion LCS from Figure 3, there are six parameters to be estimated. However, only five degrees of freedom are available in the data: two means, two variances and one covariance. Hence, the model is not identified. Different solutions exist to this problem. One option is to use the observed variance at the first occasion and the reliability estimate from the test manual (r_{ic} , based on internal consistency) to fix the error variance in the model to $\sigma_e^2 = \sigma^2_{y1} (1 - r_{ic}^2)$. This is a reasonable approach when no other information is available, and it was recommended in the original definition of the LCS model (McArdle, 2001; McArdle & Woodcock, 1997). In the present study, where we had information from the full-trajectory LCS model in Figure 2, we used the estimate of the intercept variance from the full trajectory model. Specifically, because the first time-point is the same for both models (two- and multiple-occasion models), we could use the estimate of the intercept variance from the full trajectory model (σ^2_{y0}) to estimate the error variance in the two-occasion model as $\hat{\sigma}_e^2 = \sigma_{y1}^2 - \hat{\sigma}_{y0}^2$, where σ^2_{y1} is the observed variance in the first occasion. By doing this, the two-occasion LCS is exactly identified, and the latent variable γ becomes an estimation of the actual pre-post change, after the error variance has been accounted for.

Using this two-occasion LCS, latent changes γ can be obtained. Such latent changes capture the estimated change in the process between the two occasions. If the intercept is subtracted from such scores, the *latent* residual change for each individual ($r_i = \gamma_i - \tau_\gamma$) is obtained. This is a measure of the departure of each case from the expected group change, $r_i = 0$ represents the expected average change. Cases changing above and below this average will have positive and negative r_i values, respectively.

Another approach to assess change between two occasions is a simple linear regression model,

$$Y_{2i} = b_0 + b_1 \cdot Y_{1i} + \varepsilon_i \quad [1]$$

where Y_{2i} is an individual's observed score at time 2, b_0 and b_1 are the coefficients representing the intercept and slope in the linear function, and ε is the random error or *discrepancy*, assumed to be NID $(0, \sigma^2)$, and uncorrelated with Y_1 . After the coefficients for the b_0 and b_1 are estimated for the sample, the point estimate at time 2 for each person ($Y_{2i} = b_0 + b_1 \cdot Y_{1i} + \varepsilon_i$) is computed as a function of the value at time 1 (Y_{1i}). The difference between the predicted and observed value is the *observed* residual $R_i = \hat{Y}_{2i} - Y_{2i}$. Again, this can be interpreted as a measure of individual departure from the expected group change.

The observed residuals R_i have been proposed for the identification of atypical change, particularly in the context of neuropsychological evaluation. For example, Crawford and collaborators have shown how to compute standard errors for R_i and how to use them for statistical inference at the individual level (Crawford & Garthwaite, 2006, 2007; Crawford, Garthwaite, & Ryan, 2011; Crawford & Howell, 1998). Recent research (R. Ferrer & Pardo, 2014) showed that these inferential techniques lead to adequate rates of false positives even with highly non-normal pre and post scores: values of $\alpha = .05$ led to false positive rates around or below 5% under the null hypothesis in the simulated conditions.

Despite its benefits, the R_i method has one important shortcoming: it assumes no error of measurement. Unlike in the LCS approach, the discrepancy between predicted and observed scores is entirely attributed to prediction error –i.e., all the observed variance in the residuals is considered “true” variance in change. This is an unrealistic assumption (Cronbach & Furby, 1970; Rogosa, Brandt, & Zimowski, 1982; Rogosa & Willett, 1983).

In the rest of this paper, we implement the two proposed approaches to identify deviations from typical trajectories in multi-occasion and two-occasion data. For this, we use empirical data on verbal abilities. For each research question –trajectories over multiple occasions and pre-post changes–, we apply the proposed methods, describe the results in our sample, and discuss both the methods and the findings in the context of developmental research.

Methods

The data for our study are part of a large longitudinal sample, the Connecticut Longitudinal Study (CLS; S.E. Shaywitz et al., 1992; 1990). This study involves a cohort of Connecticut children and focuses on learning, attention, and reading (S.E. Shaywitz et al., 1992; 1990; E. Ferrer et al., 2015, 2010). In this manuscript we present analyses using data from the original sample of individuals who were first assessed in 1st grade and followed up annually (mean age at 1st grade = 6.98 years; SD = .38). The sample consists of 55.2% of females and 44.8% males, and the ethnic composition is: Caucasians (84.3%), African Americans (11.2%), Asians (0.9%), Hispanics (2.0%), and other children with unknown ethnicity

(1.6%). This sample from Connecticut was similar to the racial and ethnic composition of the nation at the time of the study.

Parents or caretakers provided written consent for their children to participate in the study, and children also provided assent. The study protocol was approved by each site's ethics review board (Institution: Yale Center for Dyslexia & Creativity, protocol number: 0104012404, Project name: "Connecticut Longitudinal Study: Outcome of Learning Disabilities in Adulthood"). This study was conducted in accordance with the ethical principles that have their origin in the Declaration of Helsinki and are consistent with good clinical practices and applicable laws and regulations.

Criteria for identification of atypical cases

Dyslexia was evaluated using a composite of the word reading subtests from the Woodcock-Johnson battery (Woodcock, Johnson, & Mather, 1989; reading cluster, a composite of passage comprehension, word identification, and word attack) and the full-scale IQ score from the Wechsler Intelligence Scale for Children-Revised (*WISC-R*; Wechsler, 1999). All the information about the psychometric properties of the tests can be found in their manuals. Dyslexia was determined if a participant's score was below 90 on the reading cluster score or if there were a difference of 1.5 SDs between the IQ and the reading cluster score, a difference that refers to the standardized residuals calculated from the regression of IQ on the reading scores. If participants met criteria for dyslexia in either grade 2 or grade 4, they were considered *dyslexic* or *atypical* (for more details, see Ferrer et al. 2015). Sixty-two cases were considered atypical (*mean age* at grade 1 = 7.07 years, *sd* = .45 years, 48.4% females), and 142 were considered typical (*mean age* = 6.92 years, *sd* = .33 years, 58.5% females).

Measures

To study individual trajectories in the present study, we chose the scores from the verbal scale of the WISC-R (Wechsler, 1999) as a measure of general verbal ability. We used the scores at grades 1, 3, 5, 7, and 9. Note that these measurement occasions were different than the occasions used for dyslexia identification. Because we were interested in characterizing the growth in verbal ability, we rescaled the raw scores so they were *not* age-normed. Specifically, at first grade we fixed the *mean* to 100 and the *sd* to 15, and the rest of occasions were scaled to this metric. This is equivalent to an IQ metric, and we refer to the scores as *verbal IQ* or *VIQ*, but such scores are expected to have *mean* = 100 and *sd* = 15 at grade 1 only. Table 1 shows the descriptive statistics for both groups at all occasions. Figure 4 depicts the individual growth trajectories.

Statistical Analysis

Calibration and test samples.—We randomly divided our sample in three different groups. The first group was the *typical calibration sample (CSt)*, which included 75% of all the typical cases, randomly selected ($n = 106$). The second group was the *atypical calibration sample (CSa)*, comprised of 75% of the atypical cases, also randomly selected ($n = 46$). The third group was the *test sample (TS)*, including all remaining cases ($n = 52$; 36 typical and 16 atypical).

This random sampling was repeated $k = 1000$ times in a bootstrap analysis. For each bootstrap sample, we fitted the LCS model depicted in Figure 2 separately to the *CSt* and *CSa* samples, using all five measurement occasions. From each pair of calibration samples, we then obtained two separate sets of *model implied* and *observed* statistics (described in more detail below). We used these statistics to compute the individual level statistics described below for all cases of the corresponding *TS*.

Computation of Individual Likelihood (IL): When a longitudinal model is fitted to data from a particular sample, a model-implied vector of means $\hat{\mu}_y$ and variance-covariances matrix $\hat{\Sigma}_{yy}$ is obtained. The *IL*, or fit function for each individual within a sample, is computed as

$$IL_i = -\frac{1}{2} \left[t \ln(2\pi) + \ln |\hat{\Sigma}_{yy}| + (\mathbf{y}_i - \hat{\mu}_y)' \hat{\Sigma}_{yy}^{-1} (\mathbf{y}_i - \hat{\mu}_y) \right], \quad [2]$$

where t is the number of repeated measures, $\ln$ is the natural logarithm $\mathbf{y}_i$ is the vector of observed scores for case i .

Given the observed scores of one case, *IL* is a direct estimation of how likely those scores are in the context of a particular statistical model, and a particular set of model-implied means and covariances. Values of *IL* farther away from zero imply lower likelihood for those scores—note that, as -2Log of the likelihood is not taken, equation [2] will always yield negative values. When a SEM model is fitted to a sample, the Maximum Likelihood value or fit function for the tested model equals the sum of the *IL* for all cases within the sample (Coffman & Millsap, 2006).

This procedure provides direct likelihood estimation, for each new trajectory, of being observed in each of the two model-implied set of parameters. Although the ML fit function does not have an invariant metric, the two values of *IL* can be directly compared because the same LCS model is estimated for the same set of variables in both calibration samples. After we have a *IL* under both models, we can simply classify the case in the category for which a higher—closer to zero—*IL* value is found.

Individual likelihood from the saturated model (ILsat): The last part of Equation 2, $(\mathbf{y}_i - \hat{\mu}_y)' \hat{\Sigma}_{yy}^{-1} (\mathbf{y}_i - \hat{\mu}_y)$, is the Mahalanobis distance (Mahalanobis, 1936) for the case i in the model-implied multivariate space (Coffman & Millsap, 2006; Pek & MacCallum, 2011). In Equation 2, this term, as well as the determinant of the covariance matrix, are computed from the *model-implied* statistics. However, it can also be computed based on the sample *observed* means, variances and covariances:

$$ILsat_i = -\frac{1}{2} \left[t \ln(2\pi) + \ln |\Sigma_{yy}| + (\mathbf{y}_i - \mu_y)' \Sigma_{yy}^{-1} (\mathbf{y}_i - \mu_y) \right]. \quad [3]$$

By computing $ILsat$ for each case under the two different sets of sample statistics, one can quantify to what extent a given row of observed individual scores are likely in the context of the typical and atypical calibration samples. By comparing the two $ILsat$ (from the typical and atypical groups) the case can be assigned to a group for which a higher – closer to zero – $ILsat$ value is obtained. Unlike IL , $ILsat$ is not based on any statistical model. Thus, it is not informed by any theory about change. On the other hand, it has the advantage of not requiring fitting a model to the data for its computation.

Trajectory analyses of the test sample.—The procedures described below were repeated for each of the $k=1000$ bootstrap samples. For each case in the TS, we computed the IL under each of the two models, CSt and CSa. We then compared the ILs and assigned each case to the group for which the IL was closer to zero –i.e., the model under which that particular observed vector was more likely. For each case in the TS, we also computed two values of $ILsat$ using the observed vector of means μ_y and variance-covariances matrix Σ_{yy} , in the CSt and CSa, respectively. We assigned each case to the group for which the $ILsat$ was closer to zero. Finally, we compared the categories assigned by the IL and $ILsat$ methods with the actual group. To evaluate this procedure, we computed the *balanced accuracy* in the TS. This index is equivalent to the mean of the specificity and sensitivity, and denotes the percentage of correctly identified cases, corrected for uneven group sizes. It is computed as (Brodersen, Ong, Stephan, & Buhmann, 2010):

$$Acc = (1/2)[nTrue\ typ / (nTrue\ typ + nFalse\ atp) + nTrue\ atp / (nTrue\ atp + nFalse\ typ)]$$

Two-occasion change.—To study individual changes in two-occasion data, we implemented the two described methods to the typical CS only. The goal here was not to identify typical and atypical cases –because this distinction was made based on the full trajectories–, but to make decisions at the individual level about change between the first two time points. We fitted the LCS described in Figure 3 and also implemented the pre-post linear regression from Equation 1 using scores at grades 1 and 3. We then used the obtained information to evaluate atypical change for each case in the TS between grades 1 and 3.

We fitted this model to the CSt data and extracted the estimates for the intercept of latent change (τ_δ), the error variance (σ_e^2) and the self-feedback β . Then, we fitted the model to the TS data, constraining these three parameters but freeing σ_{y1}^2 . The purpose of these constraints was to compute the factor scores² for the TS cases in the latent change variable. Last, we computed the latent residuals as $r_i = y_i - \tau_{y_i}$.

On a final step, we fitted the linear regression model from Equation 1 to the CSt data. We then applied the regression formula, with the CSt estimate for b_1 , to predict grade 3 scores for every case in the TS. We computed the observed residual as l . Last, we studied the

²Factor scores were computed with the Empirical Bayes Modal method. Note that, for this model, the regression method would lead to the exact same scores as EBM, whereas the Bartlett method would lead to the same latent difference for all cases. For more information about these methods, see Bollen & Arminger (1991) and Coffman & Millsap (2006).

distribution of r_i and R_i , and computed effect size measures to quantify the extent to which the typical and atypical cases differed in their residuals.

All the analyses were implemented using the *R* programming language (R Core Team, 2017). Latent variable analyses were performed with the *lavaan* package v0.5-23 (Rosseel, 2012) using Full Information Maximum Likelihood estimation. Linear regression analyses were performed with the *lm* function from the *stats* package v3.4.1. The *R* code for all the analysis can be accessed at <https://github.com/EduardoEstradaRs/individualChange>

Results

Description of the sample

Table 1 shows descriptive statistics for the full typical and atypical samples at all occasions. Figure 4 depicts the individual growth curves for all cases in this study. The mean for the dyslexic group was 16.3 points lower at grade 1. This difference increased over time up to 33.2 points at grade 9. Interestingly, the atypical group had higher values for all standard deviations and correlations –i.e., this group was more heterogeneous, and had stronger relations among measurement occasions.

Figure 4 illustrates differences between groups in the overall shape of change: the typical group has less variability at all occasions and shows higher levels as well as rate of change. However, because dyslexia was diagnosed based on other scales beyond VIQ, many of the trajectories displayed in Figure 4 could be assigned to either group, if there were no additional information.

Table 2 reports the parameters and fit indices from the full trajectory LCS models (five time-points, from grade 1 to 9) in the complete typical and atypical groups. The model fit was considerably better for the atypical group. This is expected, as this group had larger variances and correlations –i.e., there was more variability, and such variability was associated across occasions.

After accounting for measurement error, the mean of the latent initial level μ_{y0} was 15.32 points higher in the typical group than in the atypical group. The former group also had a higher component of additive change, captured by μ_{ys} (59.89 points every 2 years vs. 45.7 points in the atypical group). The self-feedback (β) was negative in both groups –i.e., individuals with higher level showed less increase at the following assessment–, and this effect was stronger in the typical sample. This effect was stronger in the typical group ($\beta = -.245$ vs. $\beta = -.216$ in the atypical group). In both groups, the initial level y_0 and the additive change component' y_s were positively correlated, with similar values. This indicates that individuals with higher initial level also experienced faster growth in verbal ability. The model-implied mean trajectories for both groups are depicted in Figure 4.

Identifying individuals with atypical trajectories in one sample

For each of the bootstrap replications, we computed *IL* values for every case *in the test sample* based on the model-implied means, variances and covariances from both calibration samples. We also computed two values of *ILsat* based on the two sets of CSs observed

statistics. To illustrate how these two statistics can be used, we report the results for one of the random bootstrap replications in Table 3, Figure 5 and Figure 6. Table 3 includes the means and standard deviations of *ILsat* and *IL* for the new cases in the TS. Each case in the TS was assigned to the group for which *ILsat* and *IL* were closer to zero. Then, this group was compared with the group predetermined using previous research criteria. The confusion matrix for the *ILsat* and *IL* categorization in this bootstrap sample is included in Table 3.

Values in Table 3 indicate that means were always closer to zero –i.e., more likely— and the standard deviations were lower in the cells with correct combinations (typical-typical and atyp-atyp). Figure 5 displays scatterplots showing the distribution of the *ILsat* (panel a) and *IL* (panel b) under both models for all cases in the test sample. The case number is displayed for the incorrectly classified individuals. All of them are close to the diagonal line. Note that this line indicates equal degree of likelihood when compared with both samples. The trajectories of these individuals, along with the rest of the TS, are depicted in Figure 6.

Focusing on the atypical group (panels *b* and *d*), case 173 was incorrectly classified as typical with both procedures. *IL* also identified the case 418 as typical (panel *d*). These two individuals have trajectories similar to those in the typical group. Focusing on the typical group (panels *a* and *c*), *ILsat* identified five cases as atypical (panel *a*), whereas *IL* misidentified four cases (66, 94, 263 and 276, panel *c*). These four cases have low levels of VIQ throughout their trajectories, similar to the atypical group. In other words, *IL* misidentified only the cases with ambiguous trajectories –i.e., low scores in typical cases and high scores in atypical cases. For *ILsat*, the same pattern is observed with the atypical cases, but no clear pattern is observed in the typical group.

Performance of *IL* and *ILsat* for identifying typical and atypical cases

From the $k=1000$ bootstrap samples, the model converged 877 times for the CSt ($n=106$) and 791 times for the CSa ($n=46$). These rates of convergence are adequate considering the sample sizes –especially in CSa– and the FIML estimation procedure. Here we report the results for the 736 samples for which the model converged in both groups. The empirical distributions of the balanced accuracy for the classification based on *IL* and *ILsat* are depicted in Figure 7 (panel a).

Random identification of the cases would lead to balance accuracy = .5, whereas perfect identification would lead to balanced accuracy = 1. Panel a of Figure 7 shows that both *IL* and *ILsat* performed very well. The mean balanced accuracy across all samples were .82 for *ILsat* and .83 for *IL*. ($SD = .05$ for both methods). The central 95% interval of the empirical values were .73-.90 for *ILsat* and .74-.90 for *IL*. These values imply that, for these data, the *Individual Likelihood* is an excellent method for identifying cases from the typical and atypical populations. Interestingly, using a restricted statistical model for change -in which the sample means, variances and covariances are explained only by 7 parameters- leads to a marginal increase in accuracy.

Identifying individuals with atypical change in two-occasion data.

Table 4 includes parameters from the two-occasion LCS model fitted to VIQ data from grades 1 and 3. This model was fitted to the full typical group. The model was exactly

identified ($df=0$) and therefore it had perfect fit (McArdle, 2001; McArdle & Woodcock, 1997; see our introduction).

The LCS model was fitted to the CSt in each of the bootstrap replications. From each model, we retained the intercept of latent change τ_{y^*} , the regression coefficient β , and the error variance σ^2_e . Then, we estimated the model again using the corresponding test samples, constraining these three parameters to the values obtained with the CSt. We computed the scores in the latent change for all individuals, and computed their latent residuals as $r_i = y_i - \tau_{y^*}$.

The linear regression (Eq. 1, scores in grade 3 regressed on grade 1) for the full typical group yielded the regression parameters $b_0=42.84$ ($se=8.15$) and $b_1=.88$ ($se=.08$), both of them significant with $p<.001$. The adjusted R^2 value was .484. This model was fitted to the CSt in each of the bootstrap replications. From each model, we retained the b_1 parameter and used it to predict the grade 3 VIQ score for all cases in the corresponding TS. We computed the observed residuals as $R_i = \hat{Y}_{3i} - Y_{3i}$. To illustrate these two approaches, we report the results for one of the random bootstrap replications in Table 5 and Figure 8. The means and standard deviations for R and r for the two groups in the TS are reported in Table 5.

In this bootstrap sample, the mean difference in residuals between the groups is larger for observed R than for latent r residuals. However, because the standard deviations were also higher in R , the actual difference between the groups is larger in r than in R . This is reflected by the higher values in the two measures of effect size computed. The first is Cohen's d for independent samples (Cohen, 1988); the second is the Probability of Superiority (PS , Delaney & Vargha, 2002; Vargha & Delaney, 2000). The former is a standardization of the mean difference (typical minus atypical), using the pooled standard deviation as denominator. The latter quantifies the probability of taking one individual at random from each group and finding a higher residual in the typical individual.

The rank-order of the residuals was very similar using a LCS and a linear regression ($r=.952$). This similarity could lead to the conclusion that both methods yield similar results in the detection of atypical change. However, differences in Cohen's d and the probability of superiority indicate that the typical and atypical cases were better distinguished when using the latent residuals (r) from the LCS model. $d_R = .54$ vs. $d_r = .84$, $PS_R = .64$ vs. $PS_r = .74$, both differences in favor of the typical group. The PS values indicate that, if one case from each group is selected at random, the probability of finding a higher residual in the typical case is .64 for observed residuals, and .74 for latent residuals.

Figure 8 displays the residuals distributions for both groups. Even if the Pearson's correlation coefficient is very high, the density plots show a higher degree of overlap of the distributions of the observed residuals R . This is consistent with the higher Probability of Superiority for the latent residuals r .

Differences between the latent and observed residuals in two occasions across all the bootstrap samples

Figure 7 depicts the empirical distributions of the Probability of Superiority (panel b), and Cohen's d (panel c) computed for the observed and latent residuals across all the bootstrap samples. Based on the observed residuals, the mean of PS distribution is .696. Based on the latent residuals the mean PS is .762 ($SD=.07$ in both distributions). Regarding Cohen's d , the distribution means are .722 and .984 for the observed and latent residuals, respectively ($SD=.27$ and $SD=.30$ respectively). Overall, these results indicate that the typical and atypical cases are more different using the latent than the observed residuals, and provide evidence for the idea that the LCS model is useful for discriminating variance due to measurement error from variance in pre-post change.

Discussion

Summary of results

In this study we used two methods for: *a*) analyzing longitudinal trajectories in multi-occasion data and identifying individuals who depart from a typical trajectory; and *b*) examining change between two time-points and identifying individuals in a sample whose change departs from the expected, given the sample. We applied these methods to a measure of general verbal ability from a large sample assessed every two years from grade 1 to 9. We demonstrated the efficiency of these techniques and their use to identify individuals in longitudinal data, either with two time points or with multiple measurement occasions.

Theoretical and methodological considerations

Our results show that the Individual Likelihood is a very effective statistic for identifying typical and atypical cases when data from multiple occasions are available. In this study, we computed it from the expectations from a LCS model for growth –i.e., the model-implied means, variances and covariances (IL), and the observed statistics in the sample, equivalent to the saturated model implied statistics ($ILsat$). Our results across the random bootstrap samples show that both methods lead to excellent rates of correct identification. Although IL appears to perform the best, $ILsat$ is almost as good. This implies that: *a*) at least with these data and model, the individuals are almost equally well identified with or without a statistical model of growth, *b*) although the difference between IL and $ILsat$ is very small, having a model –and thus constraining the means, variances and covariances in the data to be explained by just few parameters– does lead to better identification of typical and atypical cases.

An interesting and convenient feature of these two procedures is the fact that they do not require fitting a statistical model for the evaluation of new cases, as long as they come from the same population as the calibration samples. The only required information are the mean vector and the variance-covariance matrix computed in a typical and atypical calibration sample. In the case of IL , this information is obtained from a statistical model of change fitted to a calibration sample. In the case of $ILsat$, it is directly observed in the calibration sample –therefore, $ILsat$ has the advantage of not requiring any statistical model.

The individual fit function or *IL* has been proposed as a method for evaluating model fit at the group level (Coffman & Millsap, 2006), and for identifying outliers within a sample (Sterba & Pek, 2012). However, to the best of our knowledge, it has not been used to identify individuals with atypical trajectories in a sample.

In principle, *IL* and *ILsat* can be also used to identify atypical cases in processes that are not expected to grow over time. However, both measures are based on a comparison of the individual vectors with the mean vector between the typical and atypical groups. Hence, they require that the two groups have a different mean in at least one of the occasions –the more occasions with different group means, and the more different the means, the better for these purposes. On the other hand, other change models such as latent growth curves or auto-regressive models may be more appropriate in situations with different trajectories or no systematic change.

The Individual Likelihood from the saturated model (*ILsat*), used as a tool for classifying cases into groups, is mathematically related to Fisher's *Linear Discriminant Analysis* (LDA; Fisher, 1936; Flury & Riedwyl, 1985). In both methods, the observed means, variances and covariances are used to compute the Mahalanobis Distance (*MD*) for each case. Then, both approaches use *MD* as a measure of the likelihood of belonging to each group, and for computing a linear discrimination function. However, there are several differences between the two approaches: a) *ILsat* also considers the determinant of the variance-covariance matrix, b) *ILsat* does not assume homoscedasticity –i.e., the variance-covariance matrix is allowed to differ across groups. In this regard, *ILsat* is more similar to Fisher's *Quadratic Discriminant function*; and, c) LDA applications typically consider the sample sizes of the different groups as prior probabilities. This information was not used for *ILsat* in our study, although the method could be extended to include it. Given the similarities between the two methods, it would be possible to conduct Linear or Quadratic Discriminant Analysis using the *IL* index (computed from the restricted multi-occasion LCS model), instead of *ILsat*. This is an interesting application to Discriminant Analysis that deserves investigation.

Growth Mixture Models (GMM, Muthén et al., 2002; Muthén & Muthén, 2000; Nagin, 1999) have been proposed as a technique for identifying unobserved heterogeneity in populations. The goal of GMM is to find subpopulations “in a post-hoc manner”, trying to describe group differences in longitudinal change between and within those unobserved groups (e.g., Ram & Grimm, 2009). GMM could be applied to our research question as well. For example, it would be possible to fit the LCS multi-occasion model separately to the typical and atypical calibration samples, and then fit a LCS Mixture Model with two classes, constraining the parameters from the calibration samples to those of the classes. Then, the probability of each case of belonging to each group could be computed and this information could be used to classify each case. This potential application of GMM is an interesting line of research to be explored in the future.

Regarding the two-occasion data, the results from the full-trajectory LCS model showed that typical cases have a higher rate of change. Hence, we expected to find higher residuals in the typical group. This was indeed the case with both the observed and latent approaches. Furthermore, our results show that the two-occasion LCS and the linear regression are not

equivalent methods. We found that, although the residuals from each method are highly correlated: a) individuals were not ranked in the exact same order under both approaches, and b) the typical and atypical groups were more different regarding their latent than their observed residuals, as the bootstrap results in Figure 7 show. Both the *probability of superiority* as well as Cohen's *d* differed between the LCS and the linear regression: the values were always larger when using the latent model. These higher effect sizes for the *latent* residuals provide evidence for the idea that the LCS was more accurate than the pre-post linear regression in detecting atypical change. In other words, partitioning the observed variance into measurement error and variance of change led to cleaner estimates of individual change.

Limitations and future directions

Because both *IL* and *ILsat* were designed for continuous variables –and multivariate normality is assumed in maximum likelihood estimation–, we do not recommend these methods when the variables are categorical or have non-ignorable skewness or kurtosis. Future research should evaluate the performance of these procedures in those scenarios, and clarify what types of differences in the trajectories can be detected with *IL* and *ILsat*.

As explained previously, in the framework of simple linear regression and observed residuals *R*, inferential methods are available for making decisions at the individual level (Crawford & Garthwaite, 2006, 2007; Crawford et al., 2011; Crawford & Howell, 1998; R. Ferrer & Pardo, 2014). These methods are based on using the data from a normative sample to computing the standard error of the predicted value for a new case. This information allows constructing a confidence interval for the individual time 2 score, and consider as “significant” any case with a time 2 score beyond those confidence limits. In principle, these procedures can be applied to the latent residuals *r* obtained from our two-occasion LCS. However, developing these methods is beyond the scope of this manuscript. Also, it must be noted that the latent residuals *r* are a transformation of the factor scores in the latent change variable. Because factor scores are indeterminate by definition (Grice, 2001), evaluating the degree of determinacy of latent residuals is a topic that deserves further exploration.

It is important to note that some statistical properties of the two-occasion LCS approach (such as bias or model fit) are not fully known in the context of detecting atypical development in cognitive abilities. A full investigation of this important issue would require a comprehensive simulation under varying conditions and is, thus, beyond the scope of the present paper. However, we believe this is a promising line for future research.

The methods proposed in this manuscript can be applied to evaluate the effectiveness of an intervention. Suppose a new sample of children with dyslexia receives an intervention. It would be possible to fit the LCS and the linear regression models to a *dyslexic* calibration sample –instead of the typical CS used in this study. Then, the parameters from those models could be used to evaluate the extent to which the change in the individuals from the intervention is atypical (too large) in reference to the dyslexic CS, not undergoing any intervention. An effective intervention would hopefully lead to few or no cases with negative residuals. And positive high residuals would identify those individuals for whom the intervention is being more effective.

Conclusion

Identifying change at the individual level is an important goal for psychological research. However, most statistical tools available focus on assessing change at the group level. In this manuscript, we proposed a set of procedures that allow the identification of atypical individual change in situations with two or multiple measurement occasions. The application of these procedures can help researchers identify individuals who are not changing as much as they should, in the light of their group norm. We hope that our work will help researchers expand their statistical toolbox and conduct further individual-based research.

References

- Ahmed Y, Wagner RK, & Lopez D (2014). Developmental relations between reading and writing at the word, sentence, and text levels: A latent change score analysis. *Journal of Educational Psychology*, 106(2), 419–434. 10.1037/a0035692 [PubMed: 24954951]
- Arbuckle JL (1996). Full information estimation in the presence of incomplete data In Marcoulides GA & Schumacker RE (Eds.), *Advanced structural equation modeling: Issues and techniques* (Vol. 243, pp. 243–277). Mahwah, NJ: Erlbaum.
- Bollen KA, & Arminger G (1991). Observational Residuals in Factor Analysis and Structural Equation Models. *Sociological Methodology*, 21, 235 10.2307/270937
- Brodersen KH, Ong CS, Stephan KE, & Buhmann JM (2010). The Balanced Accuracy and Its Posterior Distribution. In 2010 20th International Conference on Pattern Recognition (pp. 3121–3124). IEEE 10.1109/ICPR.2010.764
- Coffman DL, & Millsap RE (2006). Evaluating Latent Growth Curve Models Using Individual Fit Statistics. *Structural Equation Modeling: A Multidisciplinary Journal*, 73(1), 1–27. 10.1207/s15328007sem1301_1
- Cohen J (1988). *Statistical power analysis for the behavioral sciences* (2nd ed). Hillsdale, NJ: L. Erlbaum Associates.
- Crain-Thoreson C, & Dale PS (1992). Do early talkers become early readers? Linguistic precocity, preschool language, and emergent literacy. *Developmental Psychology*, 28(3), 421–429. 10.1037/0012-1649.28.3.421
- Crawford JR, & Garthwaite PH (2006). Comparing patients' predicted test scores from a regression equation with their obtained scores: A significance test and point estimate of abnormality with accompanying confidence limits. *Neuropsychology*, 20(3), 259–271. 10.1037/0894-4105.20.3.259 [PubMed: 16719619]
- Crawford JR, & Garthwaite PH (2007). Using regression equations built from summary data in the neuropsychological assessment of the individual case. *Neuropsychology*, 21(5), 611–620. 10.1037/0894-4105.21.5.611 [PubMed: 17784809]
- Crawford JR, Garthwaite PH, & Ryan K (2011). Comparing a single case to a control sample: Testing for neuropsychological deficits and dissociations in the presence of covariates. *Cortex*, 47(10), 1166–1178. 10.1016/j.cortex.2011.02.017 [PubMed: 21458788]
- Crawford JR, & Howell DC (1998). Regression equations in clinical neuropsychology: An evaluation of statistical methods for comparing predicted and obtained scores. *Journal of Clinical and Experimental Neuropsychology (Neuropsychology, Development and Cognition: Section A)*, 20(5), 755–762. 10.1076/jcen.20.5.755.1132
- Cronbach LJ, & Furby L (1970). How we should measure "change": Or should we? *Psychological Bulletin*, 74(1), 68–80. 10.1037/h0029382
- Delaney HD, & Vargha A (2002). Comparing several robust tests of stochastic equality with ordinally scaled variables and small to moderate sized samples. *Psychological Methods*, 7(4), 485–503. 10.1037/1082-989X.7.4.485 [PubMed: 12530705]
- Ferrer E, & McArdle JJ (2003). Alternative structural models for multivariate longitudinal data analysis. *Structural Equation Modeling-a Multidisciplinary Journal*, 10(4), 493–524. 10.1207/S15328007SEM1004_1

- Ferrer E, & McArdle JJ (2004). An experimental analysis of dynamic hypotheses about cognitive abilities and achievement from childhood to early adulthood. *Developmental Psychology*, 40(6), 935–952. 10.1037/0012-1649.40.6.935 [PubMed: 15535749]
- Ferrer E, & McArdle JJ (2010). Longitudinal Modeling of Developmental Changes in Psychological Research. *Current Directions in Psychological Science*, 19(3), 149–154. 10.1177/0963721410370300
- Ferrer E, McArdle JJ, Shaywitz BA, Holahan JM, Marchione K, & Shaywitz SE (2007). Longitudinal models of developmental dynamics between reading and cognition from childhood to adolescence. *Developmental Psychology*, 43(6), 1460–1473. 10.1037/0012-1649.43.6.1460 [PubMed: 18020824]
- Ferrer E, Shaywitz BA, Holahan JM, Marchione KE, Michaels R, & Shaywitz SE (2015). Achievement Gap in Reading Is Present as Early as First Grade and Persists through Adolescence. *The Journal of Pediatrics*, 167(5), 1121–1125. 10.1016/j.jpeds.2015.07.045 [PubMed: 26323201]
- Ferrer E, Shaywitz BA, Holahan JM, Marchione K, & Shaywitz SE (2010). Uncoupling of Reading and IQ Over Time: Empirical Evidence for a Definition of Dyslexia. *Psychological Science*, 21(1), 93–101. 10.1177/0956797609354084 [PubMed: 20424029]
- Ferrer R, & Pardo A (2014). Clinically meaningful change: False positives in the estimation of individual change. *Psychological Assessment*, 26(2), 370–383. 10.1037/a0035419 [PubMed: 24490678]
- Fisher RA (1936). The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, 7(2), 179–188. 10.1111/j.1469-1809.1936.tb02137.x
- Flury B, & Riedwyl H (1985). T 2 Tests, the Linear Two-Group Discriminant Function, and Their Computation by Linear Regression. *The American Statistician*, 39(1), 20–25. 10.1080/00031305.1985.10479381
- Grice JW (2001). Computing and evaluating factor scores. *Psychological Methods*, 6(4), 430–450. 10.1037/1082-989X.6.4.430 [PubMed: 11778682]
- Hamaker EL (2012). Why researchers should think “within-person.” In Mehl MR & Conner TS (Eds.), *Handbook of research methods for studying daily life*. New York, NY: Guilford.
- Kievit R, Brandmaier A, Ziegler G, van Harmelen A-L, de Mooij S, Moutoussis M, ... Dolan R (2017). Developmental cognitive neuroscience using Latent Change Score models: A tutorial and applications. 10.1101/110429
- Lange K, Westlake J, & Spence MA (1976). Extensions to pedigree analysis III. Variance components by the scoring method. *Annals of Human Genetics*, 39(4), 485–491. 10.1111/j.1469-1809.1976.tb00156.x [PubMed: 952492]
- Lonigan CJ, Burgess SR, & Anthony JL (2000). Development of emergent literacy and early reading skills in preschool children: Evidence from a latent-variable longitudinal study. *Developmental Psychology*, 36(5), 596–613. 10.1037/0012-1649.36.5.596 [PubMed: 10976600]
- Mahalanobis PC (1936). On the generalised distance in statistics. *Proceedings of the National Institute of Sciences of India*, 1936, 49–55.
- McArdle JJ (1998). Modeling longitudinal data by latent growth curve methods. *Modern Methods for Business Research*, 359–406.
- McArdle JJ (2001). A latent difference score approach to longitudinal dynamic structural analysis In Cudeck R, du Toit S, & Sörbom D (Eds.), *Structural equation modeling, present and future: a festschrift in honor of Karl Jöreskog*. Lincolnwood, IL: Scientific Software International.
- McArdle JJ (2009). Latent Variable Modeling of Differences and Changes with Longitudinal Data. *Annual Review of Psychology*, 60(1), 577–605. 10.1146/annurev.psych.60.110707.163612
- McArdle JJ, & Hamagami F (2001). Latent difference score structural models for linear dynamic analyses with incomplete longitudinal data In Collins LM & Sayer AG (Eds.), *New methods for the analysis of change*. (pp. 139–175). Washington: American Psychological Association, 10.1037/10409-005
- McArdle JJ, & Prindle JJ (2008). A latent change score analysis of a randomized clinical trial in reasoning training. *Psychology and Aging*, 23(4), 702–719. 10.1037/a0014349 [PubMed: 19140642]

- McArdle JJ, & Woodcock RW (1997). Expanding test-retest designs to include developmental time-lag components. *Psychological Methods*, 2(4), 403–435. 10.1037/1082-989X.2.4.403
- Mehta PD, & Neale MC (2005). People are variables too: Multilevel structural equations modeling. *Psychological Methods*, 10(3), 259–284. 10.1037/1082-989X.10.3.259 [PubMed: 16221028]
- Molenaar PCM (2004). A Manifesto on Psychology as Idiographic Science: Bringing the Person Back Into Scientific Psychology, This Time Forever. *Measurement: Interdisciplinary Research & Perspective*, 2(4), 201–218. 10.1207/s15366359mea0204_1
- Muthén B, Brown CH, Masyn K, Jo B, Khoo S-T, Yang C-C, ... Liao J (2002). General growth mixture modeling for randomized preventive interventions. *Biostatistics*, 3(4), 459–475. 10.1093/biostatistics/3.4.459 [PubMed: 12933592]
- Muthén B, & Muthén LK (2000). Integrating Person-Centered and Variable-Centered Analyses: Growth Mixture Modeling With Latent Trajectory Classes. *Alcoholism: Clinical and Experimental Research*, 24(6), 882–891. 10.1111/j.1530-0277.2000.tb02070.x
- Nagin DS (1999). Analyzing developmental trajectories: A semiparametric, group-based approach. *Psychological Methods*, 4(2), 139–157. 10.1037/1082-989X.4.2.139
- Nesselroade JR, Gerstorf D, Hardy SA, & Ram N (2007). Focus Article: Idiographic Filters for Psychological Constructs. *Measurement: Interdisciplinary Research & Perspective*, 5(4), 217–235. 10.1080/15366360701741807
- Pek J, & MacCallum RC (2011). Sensitivity Analysis in Structural Equation Models: Cases and Their Influence. *Multivariate Behavioral Research*, 46(2), 202–228. 10.1080/00273171.2011.561068 [PubMed: 26741328]
- Pennington BF, & Lefly DL (2001). Early Reading Development in Children at Family Risk for Dyslexia. *Child Development*, 72(3), 816–833. 10.1111/1467-8624.00317 [PubMed: 11405584]
- Quinn JM, Wagner RK, Petscher Y, & Lopez D (2015). Developmental Relations Between Vocabulary Knowledge and Reading Comprehension: A Latent Change Score Modeling Study. *Child Development*, 86(1), 159–175. 10.1111/cdev.12292 [PubMed: 25201552]
- R Core Team. (2017). R: A language and environment for statistical computing. Vienna, Austria: R foundation for statistical computing Retrieved from <https://www.R-project.org/>
- Rogosa DR, Brandt D, & Zimowski M (1982). A growth curve approach to the measurement of change. *Psychological Bulletin*, 92(3), 726–748. 10.1037/0033-2909.92.3.726
- Rogosa DR, & Willett JB (1983). Demonstrating the Reliability the Difference Score in the Measurement of Change. *Journal of Educational Measurement*, 20(4), 335–343. 10.1111/j.1745-3984.1983.tb00211.x
- Rosseel Y (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, 48(2). <https://doi.org/10.18637/jss.v048.i02>
- Shaywitz SE, Escobar MD, Shaywitz BA, Fletcher JM, & Makuch R (1992). Evidence that dyslexia may represent the lower tail of a normal distribution of reading ability. *The New England Journal of Medicine*, 326(3), 145–150. 10.1056/NEJM199201163260301 [PubMed: 1727544]
- Shaywitz Sally E., Shaywitz BA, Fletcher JM, & Escobar MD (1990). Prevalence of Reading Disability in Boys and Girls: Results of the Connecticut Longitudinal Study. *JAMA*, 264(8), 998–1002. 10.1001/jama.1990.03450080084036 [PubMed: 2376893]
- Sterba SK, & Pek J (2012). Individual influence on model selection. *Psychological Methods*, 17(4), 582–599. 10.1037/a0029253 [PubMed: 22845875]
- Torppa M, Poikkeus A-M, Laakso M-L, Eklund K, & Lyytinen H (2006). Predicting delayed letter knowledge development and its relation to Grade 1 reading achievement among children with and without familial risk for dyslexia. *Developmental Psychology*, 42(6), 1128–1142. 10.1037/0012-1649.42.6.1128 [PubMed: 17087547]
- Tressoldi PE, Stella G, & Faggella M (2001). The Development of Reading Speed in Italians with Dyslexia: A Longitudinal Study. *Journal of Learning Disabilities*, 34(5), 414–417. 10.1177/002221940103400503 [PubMed: 15503590]
- Vargha A, & Delaney HD (2000). A Critique and Improvement of the “CL” Common Language Effect Size Statistics of McGraw and Wong. *Journal of Educational and Behavioral Statistics*, 25(2), 101–132. 10.2307/1165329

- Voelkle MC, Brose A, Schmiedek F, & Lindenberger U (2014). Toward a Unified Framework for the Study of Between-Person and Within-Person Structures: Building a Bridge Between Two Research Paradigms. *Multivariate Behavioral Research*, 49(3), 193–213. 10.1080/00273171.2014.889593 [PubMed: 26735189]
- Wechsler D (1999). Wechsler Abbreviated Scale of Intelligence. San Antonio, TX: Harcourt Brace and Company.
- Woodcock RW, Johnson MB, & Mather N (1989). Woodcock-Johnson Psycho-Educational Battery, Revised. Allen, TX: Developmental Learning Materials.

Figure 1.

Three examples of reading ability trajectories (grades 1 to 11).

Note: Line *a* represents a typical trajectory. Line *b* represents a typical trajectory with an atypical change from grade 1 to 3. Line *c* represents an atypical trajectory (change below the reference group).

Figure 2.

Path diagram of a Latent Change Score (LCS) model for a process (Y) measured every other grade, from grades 1 to 9.

Figure 3.
Path diagram of a Latent Change Score (LCS) model for the change between two different occasions.

Figure 4.

Developmental trajectories of the verbal WISC-R scale for the full typical and atypical samples from grades 1 to 9

Note: The thick lines depict the model implied mean trajectories from the full LCS model (Figure 2) for each group. Thin lines represent observed individual trajectories.

Figure 5.

Bivariate distributions of the Individual Likelihood for the saturated model (IL_{sat} , panel a) and test model (IL , panel b) in the test sample for one of the bootstrap replications.

Note: Values computed from the data in typical and atypical calibration samples (CS), for each new case in the test sample. The diagonal line represents cases with the same value under both models. Each case was assigned to the group for which IL and IL_{sat} were closer to zero. The case number is displayed for cases incorrectly categorized.

Figure 6.
Individual trajectories in the test sample for one of the bootstrap replications.
Note: Black lines represent individuals classified incorrectly.

Figure 7.

Empirical distribution of *balanced accuracy* (a), *probability of superiority* (b), and *standardized mean difference* (c) for all the bootstrap samples.

Note: $k = 736$ samples. Panel a: *Balanced accuracy* computed using the *IL* and *ILsat* methods for case identification from the whole trajectory LCS. Panels b and c: *Prob. of Sup.* and Cohen's d computed from the observed residuals from linear regression and latent residuals from 2-occasion LCS.

Figure 8.

Bivariate distribution of the observed and latent residuals for the groups in the test sample for one of the bootstrap replications.

Note: The diagonal line indicates the same observed and latent residual.

Table 1.
Descriptive statistics for WISC-R verbal scale scores (VIQ) for typical and atypical full groups

	Typical group			Dyslexic / Atypical group			Full Information Pearson Correlations						
	<i>n</i>	<i>mean</i>	<i>sd</i>	<i>n</i>	<i>mean</i>	<i>sd</i>	Grade 1	Grade 3	Grade 5	Grade 7	Grade 9		
Grade 1	140	104.84	10.73	59	88.51	17.36		.886	.813	.801	.730		
Grade 3	141	135.34	13.37	62	114.63	17.77	.699		.861	.876	.784		
Grade 5	141	165.97	15.05	60	135.87	21.16	.551	.695		.892	.811		
Grade 7	140	184.83	14.75	60	153.82	22.23	.576	.628	.767		.882		
Grade 9	142	197.67	14.59	61	164.49	24.56	.518	.625	.752	.796			

Note: Scores rescaled to have *mean* = 100 and *sd* = 15 at grade 1 for the full sample. Typical and atypical group correlations are shown in the lower and upper half of the matrix, respectively. The differences in *n* are due to missing data.

Table 2.

Parameter estimates and fit indices for the full trajectory LCS model for both groups in the full sample

Parameter	Typical		Atypical	
	<i>Estimate</i>	<i>S.E.</i>	<i>Estimate</i>	<i>S.E.</i>
μ_{Y_0}	103.83	.99	88.51	2.20
μ_{Y_s}	59.89	1.67	45.70	2.76
B	-.245	.01	-.216	.02
$\sigma^2_{y_0}$	87.05	15.36	243.78	50.85
$\sigma^2_{y_s}$	17.79	2.88	37.85	8.74
σ_{Y_0, Y_s}	25.83	4.89	68.31	16.86
ρ_{Y_0, Y_s}	.66		.71	
σ^2_{eY}	55.63	3.83	54.43	5.76
<i>n</i>	142		62	
<i>LogL</i>	-2625.6		-1164.3	
$\chi^2(13 \text{ df})$	71.5		25.6	
<i>p</i>	<.001		.019	
<i>CFI</i>	.878		.965	
<i>RMSEA</i>	.178		.125	

Note: Model estimated with 5 time points, from grade 1 to 9, for the typical and atypical calibration samples. All parameters statistically significant with $p < .003$.

Table 3. Descriptive statistics and confusion tables for the cases in the test sample (TS) for one of the bootstrap replications.

<i>ILsat</i> from model estimated in			<i>IL</i> from model estimated in			Confusion matrix			
Typ CS	Mean	Sd	Atyp CS	Mean	Sd	Based on <i>ILsat</i>	Based on <i>IL</i>	Typ	Atyp
						Typ	Atyp	Typ	Atyp
Test typical cases ($n=36$)	-18.4	1.5	-19.5	1.7	1.8	-19.8	2.6	31	5
Test atypical cases ($n=16$)	-21.9	3.6	-18.6	1.6	4.0	-18.6	1.8	1	15

Note: Left and center sections: Means, standard deviations for the Individual Likelihood from the saturated model (*ILsat*) and the Individual Likelihood from the test model (*IL*) for the cases in the Test Sample (TS). Right section: number of cases from the TS considered typical and atypical with each method. The cells with correct combinations are shaded.

Table 4.

Parameter estimates from the LCS model for change in VIQ between grades 1 and 3.

Parameter	Estimate	SE	<i>p</i>
μ_{y1}	104.83	.90	<.001
τ_y	13.31	11.72	.256
β	.17	.11	.139
σ^2_{y1}	85.47	13.50	<.001
χ^2_y	34.93	12.11	.004
σ^2_e	28.04	Fixed	

Note: Data from the complete typical group ($n=142$).

Table 5.

Descriptive statistics for the observed and latent residuals in the test sample (TS) in one of the bootstrap replications.

	<i>n</i>	Observed Residuals (<i>R</i>)		Latent Residuals (<i>r</i>)	
		<i>Mean</i>	<i>Sd</i>	<i>Mean</i>	<i>Sd</i>
Typical	34	1.04	10.99	.57	5.04
Atypical	16	-4.53	9.18	-3.87	5.06
Cohen's <i>d</i>		.542		.837	
<i>Prob. of Superiority</i>		.636		.735	

Note: Lower section: Cohen's *d* and Probability of Superiority of the typical group.