Number Sense Growth in Kindergarten: A Longitudinal Investigation of Children at Risk for Mathematics Difficulties

Nancy C. Jordan, David Kaplan, Leslie Nabors Oláh, and Maria N. Locuniak

University of Delaware

Number sense development of 411 middle- and low-income kindergartners (mean age 5.8 years) was examined over 4 time points while controlling for gender, age, and reading skill. Although low-income children performed significantly worse than middle-income children at the end of kindergarten on all tasks, both groups progressed at about the same rate. An exception was story problems, on which the low-income group achieved at a slower rate; both income groups made comparable progress when the same problems were presented nonverbally with visual references. Holding other predictors constant, there were small but reliable gender effects favoring boys on overall number sense performance as well as on nonverbal calculation. Using growth mixture modeling, 3 classes of growth trajectories in number sense emerged.

Mathematics difficulties are widespread in the United States as well as in other industrialized nations. The consequences of such difficulties are serious and can be felt into adulthood (Dougherty, 2003; Murnane, Willett, & Levy, 1995). Low math achievement is especially pronounced in students from low-income households (National Assessment of Educational Progress, 2004). Children with weaknesses in basic arithmetic may not develop the conceptual structures required to support the learning of advanced mathematics. Although competence in high-level math serves as a gateway to a myriad careers in science and technology (Geary, 1994), many students never reach this stage. Some children gradually learn to avoid all things involving math and even develop math anxieties or phobias (Ashcraft, 2002; Ginsburg, 1997).

In recent years, there has been growing interest in children with mathematics difficulties, in part because of the relatively large body of research on children’s mathematical cognition more generally (Geary, 2000; Jordan, Blanteno, & Uberti, 2003). In general, math difficulties have been defined as below-average performance on a standardized achievement test (Hanich, Jordan, Kaplan, & Dick, 2001). Research on children with mathematics difficulties has been influenced by the wealth of studies in phonetic-based reading difficulties, studies that have led directly to the development of evidence-based methods for determining who is going to need support and how to provide help (e.g., Foorman, 2003; Torgeson, 2002). As in reading, longitudinal research is critical for understanding how math difficulties develop and change over time (Jordan, Kaplan & Hanich, 2002). For example, Geary and colleagues have shown that some mathematics difficulties are relatively fluid in elementary school whereas others are highly persistent and are characterized by basic cognitive deficits (Geary, 1990; Geary, Brown, & Samaranayake, 1991; Geary, Hamson, & Hoard, 2000). Jordan and colleagues found that although children with mathematics difficulties who are good readers can use their verbal strengths to compensate for some deficiencies in math, they have persistent problems mastering number combinations, such as 3 + 8 = 11 or 12 – 9 = 3 (Hanich et al., 2001; Jordan et al., 2002; Jordan, Hanich, & Kaplan, 2003a, 2003b). Difficulties with number combinations reflect basic weaknesses in number sense, such as the ability to grasp counting principles or to manipulate quantities mentally (Geary, Bow-Thomas, & Yao, 1992; Jordan et al., 2003a). There may also be underlying problems in working memory, spatial representation, and attention (Geary, 1994; Mix, Huttenlocher, & Levine, 2002). Longitudinal research on subgroups of mathematics difficulties is beginning to bear fruit for the development of instructional methods. For example, Fuchs, Fuchs, and Prentice (2004) demonstrate that children with mathematics difficulties who are good readers respond differently to
instruction than children with mathematics difficulties who are also poor readers.

This investigation was concerned with the emergence of mathematics difficulties. If children’s learning needs can be identified early on, we may be able to design interventions that prevent failure in math. The importance of different components of “number sense” to mathematics achievement is not well understood, although the aforementioned research on mathematics difficulties in elementary school is suggestive. Although no two researchers define number sense in exactly the same way (Gersten, Jordan, & Flojo, 2005), most agree that the ability to subitize small quantities, to discern number patterns, to compare numerical magnitudes and estimate quantities, to count, and to perform simple number transformations are key elements of number sense in young children (Berch, 2005; Case, 1998). Most children develop fundamental number sense before they receive formal instruction in elementary school, although there is significant variation by social class and cognitive ability (Ginsburg & Golbeck, 2004; Ginsburg & Russell, 1981; Huttenlocher, Jordan, & Levine, 1994; Jordan, Huttenlocher, & Levine, 1994). Even infants appear to be sensitive to small numbers and number transformations (e.g., Wynn, 1992). Preschool children learn basic counting principles (e.g., Fuson, 1988; Wynn, 1990) and can perform addition and subtraction calculations (e.g., Levine, Jordan, & Huttenlocher, 1992; Mix, Levine, & Huttenlocher, 1999). These foundational aspects of number sense are important to the “higher order” mathematical thinking (e.g., fluency and flexibility with operations and procedures) that results from formal education (Berch, 2005).

In this study, we examined the development of number sense in kindergartners. Our battery, although not exhaustive, assessed skills that have been validated by research and are relevant to the math curriculum in primary school (Griffin & Case, 1997) rather than basic cognitive abilities (e.g., general working memory). The areas included counting, number knowledge, number transformation, estimation, and number patterns and are summarized in Table 1.

### Components of Number Sense

#### Counting

According to Baroody (1987), “Counting puts abstract number and simple arithmetic within the reach of the child” (p. 33). Development of counting is a critical pathway to learning about numbers and counting weaknesses have been linked to mathematics difficulties (Geary, 2003). Most children develop knowledge of three important “how to count” principles before they enter kindergarten (Gelman & Gallistel, 1978), including the one–one principle, the stable-order principle, and the cardinality principle. Typically, children learn the count sequence by rote and then discover counting principles through informal experiences with numbers and counting (Briars & Siegler, 1984). As children progress through kindergarten and early elementary school, they gradually acquire more advanced counting abilities. They learn to count backward, to count by twos, and to enumerate object sets greater than 10. They also learn the words for decades and the rules for combining number words (e.g., combining 30 with 3 to make the larger number 33) (Ginsburg, 1989).

Counting skills are fundamental to learning base-10 concepts. Early difficulties in counting portend later difficulties with arithmetic operations (Geary, Hoard, & Hamson, 1999).

#### Number Knowledge

Children as young as 4 years of age recognize and describe global differences in quantities (Case & Griffin, 1990; Griffin, 2002, 2004). For example, they can tell which of two stacks of chips has more or fewer. Younger children rely on visual perception

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Key Elements of Number Sense in Young Children</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Area</strong></td>
<td><strong>Components</strong></td>
</tr>
<tr>
<td>Counting</td>
<td>Grasping one to one correspondence</td>
</tr>
<tr>
<td></td>
<td>Knowing stable order and cardinality principles</td>
</tr>
<tr>
<td></td>
<td>Knowing the count sequence</td>
</tr>
<tr>
<td>Number knowledge</td>
<td>Discriminating and coordinating quantities</td>
</tr>
<tr>
<td></td>
<td>Making numerical magnitude comparisons</td>
</tr>
<tr>
<td>Number transformation</td>
<td>Transforming sets through addition and subtraction</td>
</tr>
<tr>
<td></td>
<td>Calculating in verbal and nonverbal contexts</td>
</tr>
<tr>
<td></td>
<td>Calculating with and without referents (physical or verbal)</td>
</tr>
<tr>
<td>Estimation</td>
<td>Approximating or estimating set sizes</td>
</tr>
<tr>
<td></td>
<td>Using reference points</td>
</tr>
<tr>
<td>Number patterns</td>
<td>Copying number patterns</td>
</tr>
<tr>
<td></td>
<td>Extending number patterns</td>
</tr>
<tr>
<td></td>
<td>Discerning numerical relationships</td>
</tr>
</tbody>
</table>
rather than on counting to make this judgment (Xu & Spelke, 2000). By 6 years of age, however, children integrate their global quantity and counting schemas into a mental number line (Siegler & Booth, 2004). This superordinate structure, referred to as a “central structure for whole numbers,” allows children to make better sense of their quantitative worlds (Griffin, 2002). Children gradually learn that numbers later in the counting sequence have larger quantities than earlier numbers. They come to see that numbers themselves have magnitudes, such that 8 is bigger than 5 or that 6 is smaller than 9. Children use these skills in a wide range of contexts and eventually coordinate quantities to construct a linear representation of numerical magnitude, to understand place value, and to perform mental calculations. Number knowledge helps children think about mathematical problems, and its development is sensitive to children’s early experiences with number (Griffin, Case, & Siegler, 1994; Siegler & Booth, 2004). Griffin et al. (1994) found that middle-income children enter kindergarten with better developed number knowledge (e.g., the ability to discriminate which of two numbers is bigger) than low-income children. Number knowledge development is linked to the amount of informal instruction children receive at home on number concepts (Saxe et al., 1987). Number knowledge is a strong predictor of arithmetic achievement in first grade, with a zero-order predictive validity correlation of .73 (Baker et al., 2002).

Number Transformations

During preschool, children acquire quantitative abilities that are relevant to learning conventional arithmetic operations. However, children have limited success in performing verbally presented calculation problems, such as story problems (“Mike had 2 pennies. Jen gave him 3 more pennies. How many pennies does he have now?”) and number combinations (“How much is 2 and 3?”) (e.g., Ginsburg & Russell, 1981; Levine, Jordan, & Huttenlocher, 1992). Although this might indicate a lack of calculation skill, a number of other factors could compromise the child’s ability to solve story problems and number combinations. For example, children might not understand the words and syntactic structure of a problem and/or may have trouble accessing mental representations of quantities when physical referents are not provided. Levine et al. (1992) developed a “nonverbal” calculation task that eliminated these sources of difficulty. The task requires a child to reach an exact solution to a calculation problem rather than to make a judgment about the effects of the addition or subtraction transformation. Young children’s success in solving nonverbal calculations depends on the ability to hold and manipulate quantitative representations in working memory as well as an understanding of number transformations (Klein & Bisanz, 2000). The ability to solve nonverbal calculation problems develops earlier than the ability to solve comparable story problems and number combinations in most children (Levine et al., 1992). Moreover, nonverbal calculation ability varies less across social classes than does the ability to solve verbal calculations (which clearly favors middle- over low-income children) (Jordan, Huttenlocher, & Levine, 1992; Jordan et al., 1994).

Estimation

There is a significant, positive relationship between estimation abilities and skill with arithmetic operations, which requires an exact knowledge of numbers (Dowker, 1997; Rubenstein, 1985). Before learning conventional arithmetic, children as young as 4 1/2 years old can estimate concrete set sizes and use reference points (Baroody & Gatzke, 1991). For example, they can estimate the number of dots in a set (e.g., 8) within 25% of the actual value, gauge whether a set of dots was larger or smaller than a stated reference point (e.g., “Is this group of dots more than 5 or less than 5?”), and judge whether a set fits within two reference points (e.g., “Is this group of dots less than 5, between 5 and 10, or more than 10?”). Performance varies according to set size and number of reference points (i.e., one reference point is easier than two).

Number Patterns

Recognition and use of number patterns increases skill with number combinations (Threfall & Frobisher, 1999). There are patterns within number combinations and relationships between them. For example, patterns in combinations of 6 include 3+3=6; therefore 4 (which is 1 more than 3)+2 (which is 1 less than three) = 6, 5+1 = 6, so 1+5 = 6, and so forth. Children with an intuitive grasp of number patterns can readily derive answers from known combinations to solve unknown ones. This ability, in turn, helps them master or become fluent with number combinations (Gray & Tall, 1994; Jordan et al., 1994). Sense of number patterns is a key component of early mathematical knowledge (Ginsburg, 1997). Children as young as 4–5 years of
age show skill in copying and extending simple number patterns (e.g., ABAB patterns), and this skill is sensitive to background variables, such as social class (Starkey, Klein, & Wakely, 2004). Good ability to discern number patterns should help children master number combinations.

**Research Plan**

In this study we gave our number sense battery to a large group of kindergartners four times during the school year. The study is the first part of a multiyear longitudinal project by our research team. Although we expected the various tasks to be related, we were interested to see if the tasks clustered together to form a “unitary” construct of number sense or whether more than one factor would emerge. For example, some tasks may rely more heavily on conventional knowledge whereas others may involve basic number abilities (Huttenlocher et al., 1994; Levine et al., 1992).

Longitudinal assessment with four time points allowed us to examine growth curves on the various tasks from the beginning of kindergarten, when children entered school with a wide range of experiences but with limited formal instruction, to the end of kindergarten, when all children had one year of exposure to the same math curriculum. Because early number tasks are sensitive to income level (Starkey et al., 2004), and many low-income children are at risk for developing mathematics difficulties (Jordan et al., 1994), a primary focus was on comparing performance and growth in children from low- and middle-income families. Our low-income participants were mainly African American and Latino children from urban areas. Although we anticipated that low-income kindergartners, on average, would enter school at a lower level than middle-income kindergartners (Starkey et al., 2004), we were interested in the extent to which low-income kindergartners catch up with their middle-income peers during the course of the year.

We also considered gender in our analyses. Previous work with second- and third-grade children (Jordan et al., 2003a) showed that boys had a small but consistent advantage over girls on estimation and place value tasks and that boys used concrete strategies less often than girls on number combinations. In the present investigation, we sought to determine if gender differences are evident in younger children on related tasks. Another variable we considered was reading skill. As noted earlier, children’s rate of achievement in primary school math is associated with their skill in reading (e.g., Jordan et al., 2002). Reading skill may reflect some general level of competence or may be relevant primarily to conventional number tasks that have a basis in language (e.g., counting, story problems). Finally, we covaried age of kindergarten entry because children start kindergarten at very different ages. Holding other factors constant, we expected older children to have an advantage over younger children on our tasks.

Our statistical approach centered around two sets of analyses: conventional growth curve modeling and growth mixture modeling. Conventional growth curve modeling gives an estimate of the average level of number sense competency at a chosen time point (in this case, the end of kindergarten) as well as the average growth rate over time for all children (Hedeker, 2004; Raudenbush & Bryk, 2002). We performed conventional growth curve analyses for the different areas of number sense described previously (e.g., counting, number knowledge, number transformation, etc.). A limitation of conventional growth curve modeling is that it assumes that there is one average growth trajectory that describes the population under investigation. It may be more realistic, however, to examine whether there are distinct growth trajectory classes. For example, some children may have rapid rates of growth in number sense that level off quickly, others may have relatively steady rates of growth, while still others may have very slow rates of growth. In longitudinal studies of early reading, researchers have identified at least three unique trajectory classes (Kaplan, 2003; Leppanen, Niemi, Aunola, & Nurmi, 2004; Muthén, Khoo, Francis, & Boscardin, 2002). Distinct trajectory classes appear to emerge in number sense development as well (Aunola, Leskinen, Lerkanen, & Nurmi, 2004). Identification of distinct trajectory classes is important because it could lead to the finding that predictors such as age-of-entry, gender, reading proficiency, and income level do not demonstrate “one-size-fits-all” effects. Rather, the influence of these predictors may very well be different depending on the trajectory class.

To address the issue of whether there are distinct trajectory classes, we used growth mixture modeling (GMM) (Muthén, 2004) as implemented in the Mplus software program (Muthén & Muthén, 2004). The use of GMM, however, does not obviate the need to describe the sample using conventional growth curve modeling. Indeed, conventional growth curve modeling serves as an important starting point for growth mixture modeling. A technical discussion of growth mixture modeling is given in the appendix.
Method

Participants

Recruitment packages were distributed to all incoming kindergartners in six schools in the same school district. These sites were selected for several reasons. First, each school used the Trailblazers kindergarten math curriculum (Teaching Integrated Mathematics and Science Curriculum, 2004), allowing us to control for the type of math instruction that the children received. Second, we chose schools whose children varied both in income status (as measured by the percentage of children receiving free/reduced lunch) and in racial/ethnic composition (as reported by parents). Two schools were located in an urban, low-income area, and one of these schools ran full-day kindergarten programs. The other four schools were located in suburban locations and ran part-day kindergarten programs. One of the urban schools was also host to the school district’s Spanish–English bilingual kindergarten program, in addition to their regular kindergarten program. The full-day program differed from the part-day program in terms of special subjects rather than academic instruction (i.e., the full-day program included art classes, etc. while the part-time program did not).

Consent forms were returned for 66% of the children, resulting in an initial sample of 418 children. Seven of these children withdrew from school before the first round of data collection, leaving a final sample of 411 children, with an average age of 5 years 7 months. The background characteristics of these children are given in Table 2.

We sent surveys to families at the beginning of kindergarten to learn about the time that caregivers spent with their children on math and reading activities. The surveys were attached to the initial parental consent forms. This survey listed five math activities (e.g., counting objects, talking about numbers) and five literacy activities (e.g., reading books, saying/singing the alphabet), and asked parents to circle the frequency (rarely, sometimes, frequently) that best described the amount of time they spend on each activity with their child. Spanish-language surveys were sent to the parents of children in the bilingual program. Surveys were returned for 88% of the children (for 90% of the middle-income families and for 85% of the low-income families). Twelve returned surveys were excluded from the analysis because they contained incomplete data. We found a significant difference in the reported frequency of math and literacy activities between low-income ($M = 13.40$, $SD = 4.23$) and middle-income families ($M = 14.63$, $SD = 3.57$), with middle-income families reporting more frequent activities than low-income families, $t(349) = 2.662$, $p = .008$, $d = .32$. No significant effect was found for the gender of the child on reported frequency of literacy and numeracy activities (boys: $M = 14.21$, $SD = 3.941$; girls: $M = 14.28$, $SD = 3.711$; $t(349) = 0.167$, $p = .867$, $d = .02$).

To document the amount of children’s exposure to math in kindergarten, we observed all of the kindergarten classrooms twice during the year: once in January and once in May. Five fully trained testers who had previously taught in classrooms or had experience with school observations conducted the observations. We recorded the total number of minutes spent on math activities during the kindergarten day as well as the number of math activities in each class. Observers noted the beginning and end of math activities and used a checklist based on the Trailblazers curriculum to indicate what content was

<table>
<thead>
<tr>
<th>Table 2</th>
<th>Demographic Information of Participants by Income Status ($n = 411$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Low income</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>68</td>
</tr>
<tr>
<td>Female</td>
<td>69</td>
</tr>
<tr>
<td>Race</td>
<td></td>
</tr>
<tr>
<td>Minority$^b$</td>
<td>122</td>
</tr>
<tr>
<td>Nonminority</td>
<td>13</td>
</tr>
<tr>
<td>Unknown</td>
<td>2</td>
</tr>
<tr>
<td>Special education services</td>
<td></td>
</tr>
<tr>
<td>Receiving services</td>
<td>8</td>
</tr>
<tr>
<td>Not receiving services</td>
<td>129</td>
</tr>
<tr>
<td>Bilingual program</td>
<td></td>
</tr>
<tr>
<td>Participating</td>
<td>27</td>
</tr>
<tr>
<td>Not participating</td>
<td>108</td>
</tr>
<tr>
<td>Kindergarten program</td>
<td></td>
</tr>
<tr>
<td>Part day</td>
<td>70</td>
</tr>
<tr>
<td>Full-day</td>
<td>66</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Mean time 4 raw reading score (DIBELS) (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Letter naming fluency</td>
<td>32.46 (15.21)</td>
</tr>
<tr>
<td>Phoneme segmentation fluency</td>
<td>26.73 (18.47)</td>
</tr>
<tr>
<td>Nonsense word fluency</td>
<td>18.98 (13.14)</td>
</tr>
</tbody>
</table>

$^a$Income status was determined by the participants’ status in free/reduced lunch programs. “Low income” refers to students receiving free or reduced lunch, whereas “Middle income” refers to students not receiving this benefit.

$^b$Minority refers to participants who self-identified as black (36.1%), Asian (5.1%), or Hispanic (15.7%) on District registration forms.
covered during each activity. To establish reliability, pairs of observers conducted simultaneous observations in the same classrooms during their first experience in the field. Inter-rater reliability for the training round, averaged across observers, was .96, with a range from .86 to 1.00. Following these initial classroom visits, discrepancies were discussed and resolved by mutual agreement. In addition, the observation team held two calibration meetings: once after the training round and again before the scoring of the observation data.

The average math activity lasted just over 16 min ($SD = 12.63$), and children were exposed to an average of six math activities (or activities that included math) per day, with a range of 2–12 activities within any one kindergarten day. Math activities were defined as explicit math instruction, as well as math instruction that was embedded in other content areas or daily classroom activities (e.g., integrating math when reading a storybook, counting off while lining up to leave the classroom). Analyzing the data at the class level, children in the full day program participated in an average of five activities per day ($SD = 1.60$) and those in the part day program participated in an average of 7 ($SD = 2.40$) activities per day, $t(40) = 2.09, p = .043, d = .66$. Examining the data at the activity level, the mean length of an activity was 20 min ($SD = 12.99$) for the full-day program and 15 min ($SD = 12.25$) for the part-day program, $t(265) = 3.15, p = .002, d = .46$. Therefore, although children in the part-day program actually participated in more activities on average, these activities were shorter than those conducted in the full-day classrooms. As would be expected in kindergarten, the majority of activities were led by the teacher and were structured around whole-class instruction or center-based groups of children working under the supervision of the teacher or another adult. There was little variation in these patterns according to school.

**Materials and Procedure**

Children were assessed on the number sense battery four times during the kindergarten school year: September (Time 1), November (Time 2), February (Time 3), and April (Time 4). Children’s reading skills were assessed at Time 4. Schools were tested in the same order for each 1-month testing period. Children were given the battery in school by 1 of 11 researchers who were fully trained in the testing procedures. Native or near-native speakers of Spanish administered the battery to the children in the bilingual program. Although the number sense instructions were read in English to all participants, children in the bilingual program were allowed to ask that they be clarified in Spanish and/or to answer in Spanish. If a child did not understand the task items in English, the items were also administered in Spanish. The reading measure, however, was administered to all children in English as prescribed by the school district. Children were tested individually at each of the four time points. The testing time was about 30 min per session. The number tasks were presented as games to hold children’s attention. Short breaks were provided between tasks as necessary.

**Number Sense Measures**

Children were given the same number sense tasks at each time point. Because of the young age of the children and the 2-month period of time between assessments, we did not use alternate forms. The tasks were presented to each child in the following order: counting skills, number recognition, number knowledge, nonverbal calculation, story problems, number combinations, estimation, and number patterns. The number sense battery was shown to kindergarten teachers in all of the participating schools, who verified its relevance to the kindergarten math curriculum. Items were added or removed from the battery on the basis of teachers’ suggestions.

### Counting skills

The counting skills portion of the number sense battery comprised three sections: enumeration, count sequence, and counting principles. These sections were collapsed to produce one counting skills score. The total possible score on counting skills was 13 points.

#### Enumeration

There were four items in this section, adapted from Jordan, Levine, and Huttenlocher (1994). Children were shown five stars on a paper and asked to touch each star as they counted them. The paper was then hidden from view, and the children were then asked how many stars were on the paper (cardinality principle). The same procedure was followed using seven stars. Children were scored correct/incorrect for each question, resulting in a possible score of four points (two for counting and two for cardinality).

#### Count sequence

There was one item in this section. Children were asked to count to 10 and were given one point if they succeeded in doing so. Children were allowed to restart counting only once, but they were allowed to self-correct any number that they were producing.

#### Counting principles

This section consisted of eight items. The task was adapted from Geary, Hoard, and
presented (2, 6 – 4). The examiner placed a quantity of chips on her mat (in a horizontal line) and told the child how many chips were on the mat. The examiner then covered the chips with the box lid. We either added or removed chips, one at a time, and told the child how many chips were being added or removed. In order to keep the task as straightforward as possible, all addition items were given before proceeding with the subtraction items. For each item, the children were asked to say or show how many chips were left under the box. Children’s incorrect performances were corrected only on the initial addition and subtraction problems.

The item was scored as correct if the child displayed the correct number of disks and/or gave the right number word. In cases where children gave both types of responses, inconsistent answers were very infrequent (<1%). Children received scores of correct/incorrect for each of the addition and subtraction items, allowing a possible eight points.

**Story problems.** Children were given four addition and four subtraction story problems. The calculations were the same as the ones used in nonverbal calculation. The addition story problems were presented before the subtraction story problems. The addition problems were phrased as follows: “Jill has two pennies. Jim gives her one more penny. How many pennies does Jill have now?” while the subtraction problems were phrased: “Mark has three cookies. Colleen takes away one of his cookies. How many cookies does Mark have now?” Children received a score of correct/incorrect for each story problem, for a possible total of eight points.

**Number sense.** This task was adapted from Baroody and Gatzke’s (1991) estimation task. We presented children with five 15 × 23 cm cards with 3, 8, 15, 25, or 35 dots placed haphazardly on each card. The cards were presented to the child in the above order so that the number of dots increased each time they were shown a new card. The examiner displayed each card for as long as it took to say, “About how many dots do you see?” The child received a correct score if he or she estimated within 25% of the actual number of faces on the card. For example, in
response to the card with three dots, “3” was the only acceptable answer, whereas, in response to the card with 25 dots, any number in the range of 19–31 was scored as correct. The total possible score on Estimation was eight points.

Number patterns. The number patterns task was adapted from Starkey et al. (2004). Children were shown pictures of red (R), blue (B), and yellow (Y) beads on a string. Primary colors were chosen for ease of color distinction. In addition, a brief color-matching test was administered to each child before the first item was given in order to screen for color blindness. The tester pointed to each color bead, one at a time, and asked the child to point to the matching color at the bottom of the page. This was done for the first two items so that every color was accounted for. No children were excluded for color blindness.

For the number patterns task items, the beads were arranged in repeating patterns involving two colors (RBRBBRB) and three colors (RBBYRBRBYRYY); increasing patterns involving two colors (RBBRBRRBB) and three colors (RBYRRBBYRRRB); and decreasing patterns involving two colors (BBBRBB) and three colors (RBYYYRBYRBY). In each of the seven trials, one bead was left uncolored (e.g., RRBYYRBYRBY). Children had to look at the red, blue, and yellow color choices at the bottom of each page and point to which color the uncolored bead should be in order to complete each number pattern correctly. Children’s responses were scored as correct or incorrect for each pattern, for a total possible score of seven points.

Number recognition. This subtest was added to our battery at the suggestion of kindergarten teachers in the school district, who indicated that number recognition is emphasized in the kindergarten math curriculum, and consisted of four items. Children were shown the numbers 2, 8, 9, and 13 and asked to name the numbers. Children received a score of correct/incorrect for each number, for a total of four possible points.

Reading Measure

The Dynamic Indicators of Basic Early Literacy Skills (DIBELS) 6th ed. (Good & Kaminski, 2002) was given to children at the end of kindergarten. The DIBELS assesses letter naming fluency, phoneme segmentation fluency, and nonsense word fluency. The fluency scores are the total number of letters, phonemes, or nonsense words, respectively, identified in 1 min. Test–retest reliability at the end of kindergarten was .93 for letter naming fluency, .88 for phoneme segmentation fluency, and .92 for nonsense word fluency. On the basis of normative data, end of kindergarten children are considered at “some risk” if they identify fewer than 40 letters per minute (<25 high risk), fewer than 35 sounds per minute (<10 high risk), and fewer than 25 nonsense words (<15 high risk) (Good, Simmons, Kame’enui, Kaminski, & Wallis, 2002). We combined the score for each fluency area to get a total reading score. The combined measure gave us a reliable indication of general English reading proficiency.

Data Analytic Methods

As noted earlier, we used conventional growth curve modeling in the initial statistical analyses. The power of conventional growth curve modeling notwithstanding, a fundamental limitation of the method is that it assumes that the observed growth trajectories are a sample from a single finite population of individuals characterized by a single average status parameter and a single average growth rate. However, it may be the case that the sample is derived from a mixture of populations, each having its own unique growth trajectory. If this is so, then conventional growth curve modeling applied to a mixture of populations will result in biased estimates of growth. Moreover, from an intervention perspective, the use of conventional growth curve modeling in the presence of mixtures of populations could result in a lack of power to detect the influence of interventions on growth factors. For example, an intervention may be beneficial for children with problematic growth trajectories but irrelevant for children with positive and steep growth trajectories. Therefore, it is necessary to relax the assumption of a single population of growth and allow for the possibility of mixtures of different populations.

Results

The internal reliability estimates for the full number sense battery as well as the subtests, by time point, are given in Table 3. Internal consistency for the full battery is sufficiently high across all four time points (at least ≥.8). Reliabilities for individual subtests were somewhat lower, and thus the data should be viewed in a cautionary light. Alpha coefficients for counting and number recognition at Time 1 were particularly low and internal reliability for number patterns was low across the time points.

Although not the central purpose of this paper, it was of interest to examine the dimensionality of the number sense battery. An exploratory factor analysis
with maximum likelihood estimation and oblique rotation via PROMAX was conducted for all four time points. The results are shown in Table 4. We initially extracted two and three factors. In all cases, the three 3-factor solution resulted in a so-called Heywood case, namely a negative estimated residual variance. This problem is typically due to an over-extraction of the number of factors. The two-factor solution, however, did not result in a Heywood case.

The results in Table 4 show that across the four time points, the two-factor solution remains stable, with the first factor indicating “basic number skills” (counting, number recognition, number knowledge, nonverbal calculation, estimation, and number patterns) and the second factor indicating “conventional verbal arithmetic” (story problems and number combinations). The two-factor solution also shows adequate fit to the data as indicated by the root mean square error of approximation (RMSEA) (see lower section of Table 4). Not surprisingly, the factors are highly correlated, but we do find that the factor correlations are relatively stable over the time periods. The factor analyses suggest a certain degree of construct validity for our number sense battery. Nevertheless, we view it as important to the main purpose of this study to examine differences in growth trajectories on the component subskills of number sense.

The findings for the baseline conventional growth curve models are presented next, followed by conventional growth curve models predicting number sense development with the addition of the covariates of gender, age of entry into kindergarten, and reading status. We then examine the added contribution of income status to number sense development over and above the first three covariates. Following the presentation of the conventional growth curve models, we provide the results of the growth mixture models for the full number sense battery as well as for a selected set of the subtests.

**Table 3**
Cronbach’s Alpha for the Full Number Sense Battery and Individual Subtests by Time of Testing

<table>
<thead>
<tr>
<th></th>
<th>Time 1</th>
<th>Time 2</th>
<th>Time 3</th>
<th>Time 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full battery</td>
<td>.80</td>
<td>.87</td>
<td>.88</td>
<td>.82</td>
</tr>
<tr>
<td>Counting skills</td>
<td>.25</td>
<td>.39</td>
<td>.41</td>
<td>.49</td>
</tr>
<tr>
<td>Number recognition</td>
<td>.29</td>
<td>.72</td>
<td>.70</td>
<td>.63</td>
</tr>
<tr>
<td>Number knowledge</td>
<td>.57</td>
<td>.59</td>
<td>.62</td>
<td>.58</td>
</tr>
<tr>
<td>Nonverbal calculation</td>
<td>.51</td>
<td>.58</td>
<td>.53</td>
<td>.64</td>
</tr>
<tr>
<td>Story problems</td>
<td>.58</td>
<td>.71</td>
<td>.73</td>
<td>.77</td>
</tr>
<tr>
<td>Number combinations</td>
<td>.78</td>
<td>.78</td>
<td>.82</td>
<td>.85</td>
</tr>
<tr>
<td>Estimation</td>
<td>.42</td>
<td>.37</td>
<td>.41</td>
<td>.47</td>
</tr>
<tr>
<td>Number patterns</td>
<td>.20</td>
<td>.33</td>
<td>.31</td>
<td>.31</td>
</tr>
</tbody>
</table>

**Table 4**
Oblique Rotated Factor Loadings for the Number Sense Battery across Times 1 – 4

<table>
<thead>
<tr>
<th></th>
<th>Factor 1</th>
<th>Factor 2</th>
<th>Factor 1</th>
<th>Factor 2</th>
<th>Factor 1</th>
<th>Factor 2</th>
<th>Factor 1</th>
<th>Factor 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Counting skills</td>
<td>.977</td>
<td>−.067</td>
<td>.946</td>
<td>−.083</td>
<td>.899</td>
<td>−.124</td>
<td>−.946</td>
<td>−.105</td>
</tr>
<tr>
<td>Number recognition</td>
<td>.652</td>
<td>.220</td>
<td>.856</td>
<td>.011</td>
<td>.866</td>
<td>.014</td>
<td>.900</td>
<td>.012</td>
</tr>
<tr>
<td>Number knowledge</td>
<td>.665</td>
<td>.278</td>
<td>.827</td>
<td>.092</td>
<td>.761</td>
<td>.156</td>
<td>.761</td>
<td>.177</td>
</tr>
<tr>
<td>Nonverbal calculation</td>
<td>.529</td>
<td>.339</td>
<td>.495</td>
<td>.396</td>
<td>.567</td>
<td>.349</td>
<td>.459</td>
<td>.448</td>
</tr>
<tr>
<td>Story problems</td>
<td>.103</td>
<td>.673</td>
<td>−.105</td>
<td>.927</td>
<td>.058</td>
<td>.788</td>
<td>−.066</td>
<td>.940</td>
</tr>
<tr>
<td>Number combinations</td>
<td>−.077</td>
<td>.881</td>
<td>.075</td>
<td>.724</td>
<td>−.025</td>
<td>.893</td>
<td>.048</td>
<td>.793</td>
</tr>
<tr>
<td>Estimation</td>
<td>.622</td>
<td>.216</td>
<td>.483</td>
<td>.299</td>
<td>.498</td>
<td>.199</td>
<td>.533</td>
<td>.187</td>
</tr>
<tr>
<td>Patterns</td>
<td>.769</td>
<td>−.031</td>
<td>.553</td>
<td>.265</td>
<td>.670</td>
<td>.079</td>
<td>.657</td>
<td>.076</td>
</tr>
</tbody>
</table>

|                      |          |          |          |          |          |          |          |          |
| r                    | .679     | .692     | .650     | .657     |
| RMSEA (p-value)      | .062 (.202) | .048 (.509) | .034 (.775) | .059 (.266) |
| CI (RMSEA)           | .036, 0.088 | .018, 0.076 | .000, 0.065 | .032, 0.085 |

Note. CI = confidence interval; RMSEA = root mean square error of approximation.
overall number sense was statistically significant as is the case for all subtests except nonverbal calculation, estimation, and number patterns. In general, the rate of acceleration slowed over time as evidenced by the negative and statistically significant acceleration parameters for overall number sense, number recognition, number knowledge, nonverbal calculation, and number patterns.

Table 6 presents the results of the conventional growth curve models with the addition of gender, grand-mean centered kindergarten start age, and reading proficiency (z score). The coefficients given in the top portion of the table indicate the predicted values for girls of an average kindergarten start age (5 years 7 months) with an average reading score. Focusing on the effects of the covariates, we found gender differences at the end of kindergarten favoring males for overall number sense, counting skills, number knowledge, nonverbal calculation, estimation, and patterns. For example, while girls of average start age and reading scores were predicted to score 10.92 on the counting skills subtest at kindergarten exit, boys of average start age and reading scores were predicted to score 0.35 points higher on this measure. With the exception of overall number sense, boys did not differ from girls in the linear rate of growth over time, and there were no gender differences in acceleration over time. With regard to kindergarten start age, age of entry into kindergarten measured in months was positively associated with end of kindergarten performance on overall number sense, number recognition, number knowledge, nonverbal calculation, story problems, and number combinations. For every month older that a child

Table 6

<table>
<thead>
<tr>
<th>Number sense total</th>
<th>Counting skills</th>
<th>Number recognition</th>
<th>Number knowledge</th>
<th>Nonverbal calculation</th>
<th>Story problems</th>
<th>Number combinations</th>
<th>Estimation</th>
<th>Patterns</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>37.21*</td>
<td>10.92*</td>
<td>3.54*</td>
<td>5.92*</td>
<td>4.67*</td>
<td>2.83*</td>
<td>3.23*</td>
<td>2.17*</td>
</tr>
<tr>
<td>Slope</td>
<td>0.36</td>
<td>0.15*</td>
<td>0.04</td>
<td>0.01</td>
<td>−0.01</td>
<td>0.10</td>
<td>0.04</td>
<td>0.02</td>
</tr>
<tr>
<td>Var (intercept)</td>
<td>49.75*</td>
<td>1.92*</td>
<td>0.38*</td>
<td>1.26*</td>
<td>**</td>
<td>**</td>
<td>3.78*</td>
<td>0.62*</td>
</tr>
<tr>
<td>Var (slope)</td>
<td>1.49</td>
<td>0.06</td>
<td>0.04</td>
<td>0.05</td>
<td>**</td>
<td>**</td>
<td>0.07</td>
<td>0.11*</td>
</tr>
<tr>
<td>Acceleration variable</td>
<td>−0.11*</td>
<td>−0.00</td>
<td>−0.01*</td>
<td>−0.02*</td>
<td>−0.03*</td>
<td>−0.00</td>
<td>−0.02*</td>
<td>−0.00</td>
</tr>
<tr>
<td>Intercept on male</td>
<td>2.77*</td>
<td>0.35*</td>
<td>0.09</td>
<td>0.30*</td>
<td>0.59*</td>
<td>0.41</td>
<td>0.44</td>
<td>0.31*</td>
</tr>
<tr>
<td>Slope on male</td>
<td>0.08*</td>
<td>0.10</td>
<td>0.05</td>
<td>0.00</td>
<td>0.12</td>
<td>0.07</td>
<td>0.19</td>
<td>0.05</td>
</tr>
<tr>
<td>Acceleration on male</td>
<td>0.07</td>
<td>0.01</td>
<td>0.01</td>
<td>0.00</td>
<td>0.01</td>
<td>0.00</td>
<td>0.03</td>
<td>0.00</td>
</tr>
<tr>
<td>Intercept on K start age</td>
<td>0.38*</td>
<td>0.03</td>
<td>0.03*</td>
<td>0.06*</td>
<td>0.06*</td>
<td>0.06*</td>
<td>0.08*</td>
<td>0.03</td>
</tr>
<tr>
<td>Slope on K start age</td>
<td>0.01</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>Acceleration on K start age</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Intercept on reading score</td>
<td>5.17*</td>
<td>0.56*</td>
<td>0.41*</td>
<td>0.70*</td>
<td>0.76*</td>
<td>1.03*</td>
<td>1.22*</td>
<td>0.29*</td>
</tr>
<tr>
<td>Slope on reading score</td>
<td>0.06</td>
<td>0.03</td>
<td>−0.02</td>
<td>−0.05</td>
<td>0.03</td>
<td>0.05</td>
<td>0.04</td>
<td>0.03</td>
</tr>
<tr>
<td>Acceleration on reading score</td>
<td>−0.01</td>
<td>0.00</td>
<td>0.00</td>
<td>−0.01</td>
<td>−0.00</td>
<td>−0.00</td>
<td>−0.00</td>
<td>0.01</td>
</tr>
</tbody>
</table>

Note. Var () stands for the variance of the parameters in parentheses.
*p < .05. **Initial fitting of these models resulted in negative variances; therefore, models were re-estimated after fixing variance components to zero.
entered kindergarten, her total battery score increased by 0.38 points. Age of entry showed no significant relationship to linear growth or acceleration on any of the subtests. In other words, older children began kindergarten slightly more advanced in number sense skills than their younger classmates, and this difference persisted throughout the year. With regard to reading proficiency, we found that better performance on the reading test was associated with positive and statistically significant performance at the end of kindergarten on overall number sense and all subtests. Reading proficiency was not significantly associated with a linear trend or with acceleration over time.

Table 7 presents the growth curve results when income status is added to the model. The trajectories by task, income level, and gender are plotted in Figure 1. Children in the middle-income group performed significantly better than children from the low-income group at the end of kindergarten on the overall number sense battery and on all of the subtests. Furthermore, a significant effect of income status on the slope suggests that low-income children experienced less growth on story problems throughout the kindergarten year than middle-income children. No acceleration effects of income were revealed. The addition of the income status did not appreciably change the findings of previous covariates entered into the model.

Growth Mixture Modeling Results

As noted earlier, a goal of this paper was to explore the extent to which there exist underlying latent classes of growth trajectories in number sense and subtest performance, and how these latent growth trajectory classes might be differentially predicted by the background variables of gender, kindergarten start age, reading proficiency, and income status. In addition to total battery performance, we selected the nonverbal calculation, story problems, and number combinations subtests for further analysis for two reasons. First, our growth curve analyses showed divergent trajectories of growth on these tasks based on gender and income (see Figure 1), which suggested that a growth mixture approach to the tasks would provide useful information on how gender and income status may play a role in defining a population of children who are at risk for mathematics difficulties. Moreover, these three tasks assessed manipulation of identical sets of calculations in three different contexts.

The results of the growth mixture modeling are displayed in Table 8 for the total number sense and for the three selected subtests. Frequencies of children placed in each class by demographic group and task are presented in Table 9. The number of children assigned to classes (369) is slightly smaller than the
Figure 1. Fitted trajectories controlling for grand-mean centered kindergarten start age and time 4 reading score for number sense battery and subtasks by gender and income status.
Table 8
Results of Three-Class Growth Mixture Model with Effects of Gender, Grand Mean-Centered Kindergarten Start Age, Reading Score at Exit (z-score), and Income Status Number Sense Battery and Subtest Scores

<table>
<thead>
<tr>
<th></th>
<th>Number sense battery</th>
<th>Nonverbal calculation</th>
<th>Story problems</th>
<th>Number combinations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Class#1 Low/flat</td>
<td>Class#2 Avg/mod</td>
<td>Class#3 High/mod</td>
<td>Class#1 Low/flat</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Class#1 Low/flat</td>
</tr>
<tr>
<td></td>
<td>Class#1 Low/flat</td>
<td>Class#2 High/steep</td>
<td>Class#3 High/flat</td>
<td>Class#1 low/flat</td>
</tr>
<tr>
<td>Coefficient</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$I$</td>
<td>28.92*</td>
<td>38.20*</td>
<td>49.88*</td>
<td>3.49*</td>
</tr>
<tr>
<td>$S$</td>
<td>0.19</td>
<td>0.72</td>
<td>0.92</td>
<td>-0.51</td>
</tr>
<tr>
<td>$Q$</td>
<td>0.05</td>
<td>0.02</td>
<td>-0.12</td>
<td>-0.10</td>
</tr>
<tr>
<td>$I$ on MALE</td>
<td>1.54</td>
<td>0.50</td>
<td>-0.20</td>
<td>0.11</td>
</tr>
<tr>
<td>$S$ on MALE</td>
<td>0.98</td>
<td>0.72</td>
<td>0.16</td>
<td>-0.03</td>
</tr>
<tr>
<td>$Q$ on MALE</td>
<td>0.08</td>
<td>0.13</td>
<td>0.04</td>
<td>-0.00</td>
</tr>
<tr>
<td>$I$ on K start age</td>
<td>0.81*</td>
<td>0.89*</td>
<td>0.45*</td>
<td>-0.04</td>
</tr>
<tr>
<td>$S$ on K start age</td>
<td>0.01</td>
<td>0.12</td>
<td>-0.05</td>
<td>-0.03</td>
</tr>
<tr>
<td>$Q$ on K start age</td>
<td>-0.00</td>
<td>0.01</td>
<td>0.00</td>
<td>-0.00</td>
</tr>
<tr>
<td>$I$ on reading</td>
<td>2.73*</td>
<td>-0.51</td>
<td>-0.18</td>
<td>0.55*</td>
</tr>
<tr>
<td>$S$ on reading</td>
<td>0.31</td>
<td>-0.46</td>
<td>-0.12</td>
<td>0.04</td>
</tr>
<tr>
<td>$Q$ on reading</td>
<td>0.04</td>
<td>-0.02</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>$I$ on LOW INCOME</td>
<td>1.02</td>
<td>-2.85</td>
<td>-2.00</td>
<td>-0.23</td>
</tr>
<tr>
<td>$S$ on LOW INCOME</td>
<td>0.20</td>
<td>-0.59</td>
<td>-0.38</td>
<td>0.21</td>
</tr>
<tr>
<td>$Q$ on LOW INCOME</td>
<td>0.01</td>
<td>-0.08</td>
<td>-0.05</td>
<td>0.05*</td>
</tr>
<tr>
<td>Final class proportions</td>
<td>0.29</td>
<td>0.40</td>
<td>0.31</td>
<td>0.44</td>
</tr>
</tbody>
</table>

Note. $I =$ intercept; $S =$ linear slope; $Q =$ quadratic.

*p < 0.05.
sample size of 411 children because missing data prevented some children from being assigned to a particular trajectory class.

Considering the total number sense battery, we found three distinct classes that are labeled according to their kindergarten exit level relative to the average (low, average, or high) as well as their growth rate (flat, moderate, or steep growth). We characterized these three classes as “low/flat”, “avg/mod,” and “high/mod” based on the shapes of the trajectories.

Table 9
Frequencies of Children Placed in Each Class, by Demographic Group, for Number Sense Battery, Nonverbal Calculation, Story Problems, and Number Combinations (n = 369)

<table>
<thead>
<tr>
<th>Class</th>
<th>Number Sense Battery</th>
<th>Low/flat 102 children</th>
<th>Avg/mod 155 children</th>
<th>High/mod 112 children</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td></td>
<td>48 (47%)</td>
<td>77 (50%)</td>
<td>68 (61%)</td>
<td>193 (52%)</td>
</tr>
<tr>
<td>Female</td>
<td></td>
<td>54 (53%)</td>
<td>78 (50%)</td>
<td>44 (39%)</td>
<td>176 (48%)</td>
</tr>
<tr>
<td>Income status</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mid-income</td>
<td></td>
<td>30 (29%)</td>
<td>121 (78%)</td>
<td>97 (87%)</td>
<td>248 (67%)</td>
</tr>
<tr>
<td>Low-income</td>
<td></td>
<td>72 (71%)</td>
<td>34 (22%)</td>
<td>15 (13%)</td>
<td>121 (33%)</td>
</tr>
<tr>
<td>Mean age at kindergarten entry (years)</td>
<td>5.75 (0.33)</td>
<td>5.50 (0.25)</td>
<td>5.58 (0.25)</td>
<td>5.58 (0.33)</td>
<td></td>
</tr>
<tr>
<td>Mean letter naming fluency score</td>
<td>28.82 (14.80)</td>
<td>37.56 (14.97)</td>
<td>52.12 (14.72)</td>
<td>39.57 (17.35)</td>
<td></td>
</tr>
<tr>
<td>Mean phoneme segmentation score</td>
<td>20.17 (16.83)</td>
<td>30.79 (18.76)</td>
<td>39.66 (14.14)</td>
<td>30.54 (18.45)</td>
<td></td>
</tr>
<tr>
<td>Mean nonsense word fluency score</td>
<td>14.78 (11.65)</td>
<td>21.40 (13.15)</td>
<td>40.20 (23.99)</td>
<td>25.28 (19.69)</td>
<td></td>
</tr>
</tbody>
</table>

| Class | Nonverbal calculation | Low/flat 91 children | High/steep 148 children | High/mod 168 children | Total |
|-------|-----------------------|                      |                       |                       |       |
| Gender          |                       |                       |                      |                       |       |
| Male          |                       | 38 (42%)              | 74 (50%)             | 102 (61%)             | 214 (53%) |
| Female        |                       | 53 (58%)              | 74 (50%)             | 66 (39%)              | 193 (47%) |
| Income status |                       |                       |                      |                       |       |
| Mid-income    |                       | 44 (48%)              | 99 (67%)             | 128 (76%)             | 271 (67%) |
| Low-income    |                       | 47 (52%)              | 49 (33%)             | 40 (24%)              | 136 (33%) |
| Mean age at kindergarten entry (years) | 5.50 (0.25) | 5.58 (0.33) | 5.67 (0.33) | 5.58 (0.33) |
| Mean letter naming fluency score | 29.57 (16.28) | 39.31 (16.67) | 45.40 (15.83) | 39.60 (17.31) |
| Mean phoneme segmentation score | 22.72 (18.76) | 31.40 (19.52) | 33.95 (16.55) | 30.47 (18.49) |
| Mean nonsense word fluency score | 14.54 (12.34) | 25.57 (18.40) | 30.89 (21.63) | 25.57 (19.70) |

| Class | Story problems | Low/flat 260 children | High/mod 54 children | High/flat 55 children | Total |
|-------|----------------|-----------------------|                      |                       |       |
| Gender          |                       |                       |                      |                       |       |
| Male          |                       | 121 (46%)             | 53 (98%)            | 19 (35%)              | 193 (52%) |
| Female        |                       | 139 (54%)             | 1 (2%)              | 36 (65%)              | 176 (48%) |
| Income status |                       |                       |                      |                       |       |
| Mid-income    |                       | 148 (57%)             | 50 (93%)            | 50 (91%)              | 248 (67%) |
| Low-income    |                       | 112 (43%)             | 4 (7%)              | 5 (9%)                | 121 (33%) |
| Mean age at kindergarten entry (years) | 5.58 (0.25) | 5.67 (0.33) | 5.67 (0.33) | 5.58 (0.33) |
| Mean letter naming fluency score | 35.33 (15.83) | 48.54 (16.81) | 50.78 (16.64) | 39.57 (17.35) |
| Mean phoneme segmentation score | 26.65 (19.42) | 31.82 (15.94) | 38.93 (14.12) | 30.54 (18.45) |
| Mean nonsense word fluency score | 19.10 (13.32) | 25.98 (16.49) | 39.26 (25.69) | 25.28 (19.69) |

| Class | Number combinations | Low/flat 223 children | High/steep 50 children | High/mod 96 children | Total |
|-------|---------------------|-----------------------|                       |                       |       |
| Gender          |                       |                       |                      |                       |       |
| Male          |                       | 109 (49%)             | 27 (54%)            | 57 (59%)              | 193 (52%) |
| Female        |                       | 114 (51%)             | 23 (46%)            | 39 (41%)              | 176 (48%) |
| Income status |                       |                       |                      |                       |       |
| Mid-income    |                       | 119 (53%)             | 38 (76%)            | 91 (95%)              | 248 (67%) |
| Low-income    |                       | 104 (47%)             | 12 (24%)            | 5 (5%)                | 121 (33%) |
| Mean age at kindergarten entry (years) | 5.58 (0.33) | 5.50 (0.25) | 5.75 (0.25) | 5.58 (0.33) |
| Mean letter naming fluency score | 34.29 (15.65) | 39.42 (16.70) | 51.90 (15.17) | 39.57 (17.35) |
| Mean phoneme segmentation score | 26.65 (19.42) | 31.82 (15.94) | 38.93 (14.12) | 30.54 (18.42) |
| Mean nonsense word fluency score | 19.10 (13.32) | 25.98 (16.49) | 39.26 (25.69) | 25.28 (19.69) |

Note. Percentages inside parentheses refer to within task model classification. Other values in parentheses indicate standard deviations.
of the trajectories (see Figure 2). The bottom of Table 8 shows that approximately 29%, 40%, and 31% of the sample fell into these three classes, respectively. Class labels and latent class proportions for the remaining subtests are also displayed in Table 8. The trajectories for total number sense as well as for the three subtests are plotted by class in Figure 2.

Turning to the influence of the predictor variables within each class for total number sense, there were no gender or income differences across the three trajectory classes. There was a positive and statistically significant association between the age of entry into kindergarten and exit number sense score across the three classes of total number sense, meaning that children who are older when they enter kindergarten demonstrate significantly better exit number sense scores than their younger counterparts. We found that reading proficiency was significantly associated with total number sense score at the end of kindergarten, but only for children in the low/flat class. In other words, children in the low/flat class who were better readers ended kindergarten with better number sense performance. This suggests that skill development in reading may be beneficial for low-performing children, but has no statistically significant effect for children already performing better in math.

We also found three distinct classes with respect to the nonverbal calculation task: “low/flat,” “high/steep,” and “high/mod.” Approximately 44%, 30%, and 26% of our sample falls into these classes, respectively. Although there were two classes of children who did not demonstrate statistically significant growth in nonverbal calculation in kindergarten, there was also a group of children (high/steep) whose skills in this area grew throughout kindergarten, allowing them to exit with above-average scores on this task.

The only gender difference occurred on the nonverbal calculation exit score for the high/steep class, but this difference of 0.10 points on the subtask score was minimal. Likewise, the negative effect of kindergarten start age on task performance for children in the class was not large: for every month older that a child began kindergarten, that child scored an average of 0.10 points lower on nonverbal calculation at the end of the year. The more interesting findings occurred in reading. Reading score at kindergarten exit is significantly and positively associated with the nonverbal calculation performance of children in the low/flat and high/steep classes. Furthermore, the children in the high/steep class received an additional benefit of reading on nonverbal calculation growth and acceleration. In sum, reading ability played a significant role in the growth and exit performance of nonverbal calculation skills, but only for children in the high/steep class.

Analysis of children’s performance on story problems also revealed three distinct classes: low/flat, high/mod, and high/flat. Of the children in the sample, 70%, 14%, and 16% of the children fell into these three classes, respectively. As can be seen from Table 5, gender appeared to have a very large and significant effect on story problems skills at the end.
of kindergarten for children in the high/mod class. Upon further investigation, however, we found that this class consisted of 53 boys and only 1 girl; therefore, the effect of gender was overestimated. Kindergarten start age had a small but statistically significant effect on growth and acceleration in story problems performance, but only for the children in the high/mod class. Reading showed significant effects only on story problems exit score for the low/flat class alone. Income status, however, had a more notable effect on this task: children in the low/flat class from low-income households scored an average 0.68 points lower at the end of kindergarten and grew an average of 0.29 points slower per month in story problems skill throughout kindergarten than did children from middle-income households in the same class.

On number combinations, children fell into three classes: low/flat, high/steep, and high/mod. The distribution of children into these classes was 61%, 13%, and 26%, respectively. Significant linear growth was not observed for any class; there was a four-point discrepancy on exit scores between the low/flat class on the one hand and the high/steep and high/mod classes on the other, as can be seen by comparing the data presented in the first line of Table 8. Gender had an effect on number combinations score for the low/flat class alone. Among these children, boys scored an average 0.38 points lower than girls at the end of the kindergarten year. Among the high/steep class, boys experienced slightly faster growth during the year. Age of kindergarten entry seemed particularly relevant for children in the high/steep class: children scored an average of 0.32 points higher on number combinations for every month later they enrolled in kindergarten. The children in this class also gained an additional 0.19 points per kindergarten month and an acceleration of 0.02 points per month for every month later they started kindergarten. Kindergarten start age also had a small but positive effect on number combinations score for children in the low/flat class. Reading, on the other hand, affected only the exit score in number combinations for children in the low/flat class. Income had different effects on kindergarten exit score depending on class. Children from low-income households in the low/flat class, for example, scored 0.40 points lower on number combinations at the end of the year than did children from mid-income households in the same class. Low-income children in the high/mod class, however, scored 1.14 points higher at exit than did middle-income children in the same class.

### Table 10

<table>
<thead>
<tr>
<th>Odds of Class Assignment (with Class #3 as Comparison) for Number Sense Battery, Nonverbal Calculation, Story Problems, and Number Combinations</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Number Sense Battery</strong></td>
</tr>
<tr>
<td>Class #1 (low/flat)</td>
</tr>
<tr>
<td><strong>MALE</strong> 0.33</td>
</tr>
<tr>
<td>K start age 1.32*</td>
</tr>
<tr>
<td>Time 4 reading score 0.07*</td>
</tr>
<tr>
<td>LOW INCOME 17.29*</td>
</tr>
<tr>
<td><strong>Nonverbal Calculation</strong></td>
</tr>
<tr>
<td>Class #1 (low/flat)</td>
</tr>
<tr>
<td><strong>MALE</strong> 0.49</td>
</tr>
<tr>
<td>K start age 0.87</td>
</tr>
<tr>
<td>Time 4 reading score 0.47</td>
</tr>
<tr>
<td>LOW INCOME 1.30</td>
</tr>
<tr>
<td><strong>Story Problems</strong></td>
</tr>
<tr>
<td>Class #1 (low/flat)</td>
</tr>
<tr>
<td><strong>MALE</strong> 1.09</td>
</tr>
<tr>
<td>K start age 0.92</td>
</tr>
<tr>
<td>Time 4 reading score 0.43*</td>
</tr>
<tr>
<td>LOW INCOME 4.39*</td>
</tr>
<tr>
<td><strong>Number Combinations</strong></td>
</tr>
<tr>
<td>Class #1 (low/flat)</td>
</tr>
<tr>
<td><strong>MALE</strong> 0.46*</td>
</tr>
<tr>
<td>K start age 0.90</td>
</tr>
<tr>
<td>Time 4 reading score 0.30*</td>
</tr>
<tr>
<td>LOW INCOME 10.49*</td>
</tr>
</tbody>
</table>

*Comparisons are made with Number Sense Battery Class #3 (high/moderate).
*Comparisons are made with Nonverbal Calculation Class #3 (high/moderate).
*Comparisons are made with Story Problems Class #3 (high/flat).
*Comparisons are made with Number Combinations Class #3 (high/moderate).
*Story problems, Class #2 consisted of 53 boys and 1 girl; therefore, the effect of gender is overestimate.
*p < .05.

### Multinomial Logistic Regression Analyses

An important additional analysis derived from the application of growth mixture modeling is the prediction of class membership as a function of the background variables via multinomial logistic regression. Multinomial logistic regression provides the odds of a child being assigned to one growth trajectory class versus another based on whether the child is male (or female), low income (or middle income), a good reader (or a poor reader), or started kindergarten earlier (or later). Multinomial logistic regression coefficients were estimated within the MPlus software program, and the odds ratios are displayed in Table 10. For these analyses, Class #3 for the total number sense battery and subtests was used as the comparison group (see Table 8).
Examining the results for total number sense, we found that children who were older when they entered kindergarten, had lower reading proficiency and were of low-income status had greater odds of being in the low/flat class versus the high/mod class. For example, the kindergarten start age coefficient can be interpreted as indicating that for every month older a child began kindergarten that child had 1.32 times greater odds, or was approximately one-third times more likely, of falling into the low flat versus the high/mod class. The low-income coefficient indicates that children of low-income status were 17.29 times more likely to be assigned to the low/flat class rather than the high/mod class when compared with their middle-income peers. In contrast, children with better reading scores had lower odds of being in both the low/mod or avg/mod class relative to being in the high/mod class. Reading skill was the only variable that differentiated the avg/mod class from the high/mod class in number sense, but kindergarten start age, reading skill, and income status differentiated the low/flat class from the high/mod class. In other words, better reading performance will increase a child’s odds of a high number sense exit score whether or not the child is in the avg/mod or the low/flat class. Low-income status, on the other hand, gives a child greater odds of having low end-of-kindergarten number sense scores.

Looking at the coefficients for the nonverbal calculation task, we found that children with lower kindergarten exit scores in reading had lower odds of being in the high/steep versus the high/mod class. Results from the story problems task showed that poorer readers had lower odds of being in the high/flat class versus the low/flat class. Low-income children had greater odds of being in the low/flat class versus the high/flat class. (The significant gender difference on story problems [between the high/mod and the high/flat class] must be interpreted cautiously, as the high/mod class contained only one girl.)

The results from the number combinations task showed that low-income girls and poorer readers had higher odds of being in the low/flat versus the high/mod class. Children who entered kindergarten at an older age, children with lower reading scores, and middle-income children had lower log-odds of being in the high/steep versus the high/mod class. In sum, low income status and poor reading skill consistently increase children’s odds of starting and ending kindergarten at a low level on the number tasks, with the exception of nonverbal calculation. It should be noted, however, that the effects for income are much greater than the effects for other predictors.

Discussion

We tracked the development of number sense in 411 children over four time points during kindergarten. Children’s performance and growth were examined while controlling for background variables of income level, gender, age, and reading skill. We looked at children’s level of performance at the end of kindergarten as well as their rates of growth from the beginning to the end of kindergarten. All children received essentially the same math curriculum throughout the school year. There did not appear to be large differences between classes or schools in terms of math exposure. Although our number sense battery showed good reliability across the various time points (> .8), individual subtests were less reliable and should be interpreted with caution.

Two-Dimensional Model of Number Sense

A factor analysis addressed the extent to which our number sense tasks cluster together. Our battery was two dimensional across all time points: the first dimension involving basic number skills (counting, number recognition, number knowledge, nonverbal calculation, estimation, and number patterns) and the second dimension involving conventional arithmetic (story problems and number combinations). This two-dimensional model corresponds with quantity discrimination and verbal sequential factors of mathematics proficiency reported by Okamoto and Case (1996) using similar methods of analysis. The model is also in keeping with Berch’s (2005) discussion of the lower and higher order features of number sense. Lower order number sense is characterized by elementary knowledge about quantity, such as counting and comparing numerical magnitudes (Dehaene, 2001), whereas secondary skills that result from conventional educational activities characterize higher order number sense. Nonverbal calculation involves a more basic arithmetic ability, one that emerges after the approximate skills of infancy but before the conventional skills of arithmetic (Huttenlocher et al., 1994). Likewise, counting skills, estimation, and knowledge of number magnitudes underpin school arithmetic (Aunola et al., 2004; Geary, 1995).

Development of Number Sense

Growth curve analyses revealed significant linear growth throughout kindergarten in most areas of number sense. We found three distinct growth classes for our overall battery as well as for the se-
lected calculation subtests (nonverbal calculation, story problems, and number combinations). Although the characteristics of growth classes varied somewhat between tasks, we generally found a group of children who ended kindergarten at a low level with flat growth, another group of children who ended kindergarten at a high level with moderate to steep growth, and a final group of children who ended kindergarten at an average level with moderate growth. Previous longitudinal work with a sample of Finnish children (Aunola et al., 2004) identified two classes of growth in math achievement, namely low performers and high performers. Our data, using a more heterogeneous sample, further captured a class that starts kindergarten at a relatively low level but makes moderate to fast growth during the year. It is likely that our multi-ethnic U.S. sample had more diversity in terms of preschool experiences than the Finnish sample, which could account for the identification of a third class in our growth mixture results. Identification of these unique growth classes is important for differentiating children in obvious need of intervention (e.g., low/flat) from those who are likely to make progress with minimal or no special help (e.g., avg/mod).

Social Class

A primary goal of our study was to examine number sense development in children from low-income households. Although social class differences in number-related skills have been documented (e.g., Ginsburg & Russell, 1981; Jordan et al., 1994; Starkey et al., 2004), few investigations have studied young children longitudinally with multiple time points. Holding other background variables constant, low-income children in the present study fared poorly on most of our number sense tasks relative to their middle-income peers. Yet children in both income groups progressed at similar rates on many tasks. A notable exception was on story problems, where low-income kindergartners showed almost no growth from the beginning to the end of the year, even when we accounted for reading proficiency. Story problems are a conventional task that requires children to perform arithmetic calculations embedded in sentences. Story problems are demanding not only in terms of basic arithmetic but also in terms of language and auditory attention (Fuchs et al., in press; Levine et al., 1992). Although many kindergartners fell into the low-flat growth class for story problems, a low-income kindergartner was four times more likely to be in this class than his or her middle-income peers. A similar pattern was observed on number combinations, where a low-income child was much more likely to fall in the low/flat class than a middle-income child. Interestingly, both income groups made comparable progress when the same calculations were presented in a nonverbal context with visual referents. The findings corroborate those of Huttenlocher and colleagues (Huttenlocher et al., 1994; Levine et al., 1992), which revealed that nonverbal calculation abilities are less sensitive to social class than are conventional story problems. Our divergent story problem trajectories show that the income gap widens in kindergarten, despite comparable math instruction, and low-income children are likely to enter first grade with a genuine disadvantage in problem solving. It should be noted that story problems and number combinations were not introduced into the curriculum until relatively late in the year. Therefore, a possible scenario is that middle-income children are acquiring these skills outside the classroom while low-income children are not. In other words, it is not necessarily the case that low-income children are unresponsive to instruction in these areas. It should also be noted that there may have been classroom differences between low- and middle-income children that were not revealed in the present study. For example, many of our low-income children attended a full-time kindergarten program (with the addition of special subjects, such as art) whereas all of our middle-income children attended a part-time program. In any case, because math becomes more dependent on verbal skills in primary school and difficulties in story problems and arithmetic combinations are defining characteristics of math difficulties (Jordan et al., 2002), kindergarten or early first grade may be an important time to intervene with high-risk children.

Gender

Overall, our findings suggest that gender differences in math emerge as early as kindergarten. There were small, but statistically reliable, gender effects on kindergarten level performance on overall number sense, nonverbal calculation, and estimation. In each case, boys showed an advantage over girls, and the findings held above and beyond income level, age, and reading ability. There is also a greater percentage of boys in high performing classes on overall number sense (61% vs. 39%) and on nonverbal calculation (61% vs. 39%) when background variables are controlled for. A gender effect was also reported in previous work with school-age children (Jordan et al., 2003a) where third-grade boys performed better.
than girls on tasks assessing place value, estimation (i.e., approximate arithmetic), and mental computation. Although Aunola et al., (2004) did not find gender differences in level of performance, they found that boys achieved at a faster rate in primary-school math, especially among high performing classes of children. We did not see gender differences in growth rates in kindergarten. Although our gender effects may be small in practical terms, they are consistent and could help explain some of the gender differences in math problem solving that are present in later years (Hyde, Fennema, & Lamon, 1990).

Note that the skills on which boys seem to have an advantage (e.g., nonverbal calculation) involve spatial reasoning as well as some manipulation of quantities or numerical magnitudes (Baroody & Gatzke, 1991; Dehaene, Spelke, Pinel, Stanescu, & Tsivkin, 1999). The finding raises the question of whether there are cognitive explanations for gender differences in early math (Geary, 1998). Interestingly, a consistent finding in the literature on spatial cognition is that males have an advantage over females in spatial cognition when it is measured by mental rotation tasks or by spatial judgments about moving objects (Casey, Nuttal, & Pezaris, 1997; Royer, Tronisky, Chan, Jackson, & Marchant, 1999). Of course, there may also be explanations related to socialization and motivation such that boys are encouraged more than girls on number tasks, even at an early age (Aunola et al., 2004; Eccles, Janus, & Harold, 1990).

Reading

Reading proficiency at the end of kindergarten was a significant predictor of performance on all number sense measures (but not of rate of growth) even when we considered other background variables. Leppanen et al. (2004) also reported a relationship between early reading and number skills and Jordan et al. (2002) found that children with mathematics difficulties who are good readers make better progress in math achievement than children with mathematics difficulties who are poor readers. Recall that we assessed reading with a standard measure of word-level fluency skills, that is, letter naming, nonsense word reading, and phoneme segmentation. It is possible that the kindergarten reading score reflected some kind of general verbal cognitive ability. Supporting this notion, Jordan and Hanich (2003) report that elementary school children with both reading and math difficulties have lower IQ scores than do children with mathematics difficulties only or children with reading difficulties only. However, the data may also reflect a relationship between school-related skills and early home experiences (Saxe et al., 1987). In any case, reading difficulties are a correlate of math difficulties (Leppanen et al., 2004) and should be viewed as a risk factor.

Age

As expected, age was a significant predictor of overall number sense performance at the end of kindergarten. Older children began kindergarten slightly more advanced than their younger peers and this difference held throughout the school year. Of course, the age effects are small in practical terms, and it remains to be seen whether age will predict math achievement in subsequent grades. Reading researchers have found that although older kindergartners have stronger emergent literacy skills than younger kindergartners, age is not a good predictor of reading achievement at the end of the first and second grades (Crone & Whitehurst, 1999).

The present kindergarten study is the initial phase of a longitudinal investigation of children’s mathematics from kindergarten through third grade. Remaining questions include the extent to which kindergarten number sense predicts achievement in primary school, the extent to which growth classes in kindergarten predict growth classes in elementary school, and the extent to which income, gender, and age effects hold, become greater, or attenuate in elementary school. We will also continue to look at children’s reading development as well as the role of general cognitive abilities in children’s rate of achievement in math.

References


Muthén, B. O., Khoo, S. T., Francis, D., & Boscardin, C. K. (2002). Analysis of reading skills development from...


Appendix: Growth Mixture Modeling

This appendix provides a brief technical overview of growth mixture modeling. The growth mixture model is part of a general analytic framework for the analysis of continuous and categorical latent variables. For an overview of the general analytic framework including growth mixture modeling, see Muthén (2004).

In line with the discussion given in Kaplan (2003) growth mixture modeling begins by combining conventional growth curve modeling (e.g. Raudenbush & Bryk, 2002) with latent class analysis (e.g., Clogg, 1995) under the assumption that there exists a mixture of populations each defined by unique growth trajectories. In this respect, growth mixture modeling is more flexible than conventional growth curve modeling which assumes a single population generating the empirical growth trajectories.

An extension of latent class analysis sets the groundwork for growth mixture modeling. Specifically, latent class analysis can be applied to repeated measures at different time points. This is referred to as latent class growth analysis (LCGA) (see e.g., Nagin, 1999). As with latent class analysis, LCGA assumes homogenous growth within classes. Growth mixture modeling, by contrast, relaxes the assumption of homogenous growth within classes and is capable of capturing two important forms of heterogeneity. The first form of heterogeneity is captured by individual differences in growth through the specification of the conventional growth curve model. The second form of heterogeneity is more fundamental representing heterogeneity in classes of growth trajectories.

The specification of the growth mixture model is similar to that given for the conventional growth curve model. The difference, as noted, lies in allowing there to be different growth trajectories for different classes. Specifically, we hypothesize there exists a latent categorical variable $c_i$ composed of $K$ classes ($k = 1, 2, \ldots K$) The $i$th student belongs to one of the $K$ classes and the classes are mutually exclusive and exhaustive. The growth parameters are obtained via the conventional growth curve model defined as

$$y_{it} = \eta_{0i} + \eta_{1i}a_{it} + e_{it},$$

where $y_{it}$ is a continuous outcome measure at time $t$ for student $i$, $a_{it}$ is a time measure for person $i$ such as age, grade, time point of study, etc., where for simplicity, we will require $a_{it} = a_t$ meaning that the time metric is common to all individuals. The random growth factors $\eta_{0i}$ and $\eta_{1i}$ represent the status and rate of growth, respectively, and $e_{it}$ is a random error component. An acceleration parameter can also be accommodated in this specification.

The specification of the growth mixture model allows the random growth factors $\eta_{0i}$ and $\eta_{1i}$ to vary over latent classes through their means, variances, and covariances. Specifically, let

...
\[ \eta_{0i} = \alpha_{0k} + \gamma_{0k}x_i + \zeta_{0i}, \]
\[ \eta_{1i} = \alpha_{1k} + \gamma_{1k}x_i + \zeta_{1i}, \]

where \( \alpha_{0k} \) and \( \alpha_{1k} \) are the average intercept and average rate of growth that vary over the classes, respectively. The parameters \( \gamma_{0k} \) and \( \gamma_{1k} \) relate the time-invariant covariates \( x \) to the random growth factors and also allowed to vary over classes, and the \( \zeta \)'s are random disturbance terms which are assumed to be constant over the latent classes.

The specification in Equation (2) demonstrates the flexibility of the growth mixture modeling approach—namely that the role of covariates in predicting growth may be different across substantively different trajectory classes.

The growth mixture model also allows covariate information contained in \( x \) to influence class membership directly via a multinomial logistic regression. Specifically, the multinomial logistic regression giving the probability of membership in class \( k \) given background covariate information \( x \) can be written as

\[ p(c_i = k|x_i) = \frac{e^{\beta_{0k} + \beta_{1k}x_i}}{\sum_{k=1}^{K} e^{\beta_{0k} + \beta_{1k}x_i}}, \]  

where a given class \( K \) is chosen as the reference class, with coefficients \( \beta_{0k} \) and \( \beta_{1k} \) fixed to zero.

Estimation of the growth mixture model is carried out using maximum likelihood of the EM algorithm (Dempster, Laird, & Rubin, 1977) as implemented in the Mplus software program (Muthén & Muthén, 2004). Technical details can be found in Muthén and Shedden (1999).

A sequence of modeling steps have been suggested for the application of growth mixture modeling and is utilized in this paper. To begin, the researcher specifies an a priori number of trajectory classes based on theory or inspection of empirical growth trajectories. Selecting among models with differing numbers of classes can be determined in three ways. First, one can select the best model by an inspection of the Bayesian Information Criterion (BIC). The BIC is a measure that balances the fit of the model with a penalty function for adding parameters to the model. If one attempts to improve the fit of the model by adding a new trajectory class, then the BIC will increase unless the additional trajectory class improves the fit of the model. Thus, we are interested in the smallest BIC among a variety of competing specifications. A plot of the BIC values against models that specify varying numbers of classes can be used to ascertain the number of classes. The model with the lowest BIC value is recommended.

Second, one can look at the posterior probability of being assigned to a particular class. The precision of group assignment is suggested by classes with substantively large numbers of assigned units.

Third, one can examine the utility of the number of classes in terms of substantive considerations. For example, prior knowledge of the incidence of problematic math development in the population could be used to determine if certain classes contain reasonable numbers of children.

Following the determination of the number of latent classes, one can add predictors of the latent classes model, thus fully specifying a growth mixture model. However, the addition of covariate information can influence latent trajectory class determination, and so it is essential to re-examine the statistical and substantive utility of the latent trajectory classes using the methods just outlined.
Methods for analyzing longitudinal data provide researchers with powerful tools for describing developmental patterns and identifying predictors of development. A wide variety of analytic methods are available for describing developmental patterns and identifying predictors of development. The purpose of this chapter is to provide an overview of five of the currently available analytic approaches to estimate growth curves for continuous outcomes. We focus on describing the assumptions and outline the limitations a researcher faces when using univariate and multivariate repeated measures analysis of variance, hierarchical linear models (HLM), latent growth curve models (LGC), and prototypic or growth mixture model methods. Using a simulated data set and data from a child care intervention study, we demonstrate the strengths and weaknesses of each approach.

Data analysis for developmental projects provides the mechanism for testing the research questions with project data. Ideally, statistical analysis integrates the three elements of a longitudinal study: theoretical model of change, longitudinal design, and statistical models (Burchinal & Appelbaum, 1991; Collins, 2006; Singer & Willett, 2003). The project should be based on a well-articulated theory of change that dictated the study design to ensure that data collection adequately capture the hypothesized change and processes implicated in that theory (Collins, 2006). Selection of the analytic methods follows logically in a well-designed study based on a clearly articulated theory of change.

A wide variety of analytic methods are readily available, but not all researchers know how to select a method that matches both their theory of change and the data they collected. The purpose of this chapter is to provide an overview of some of the currently available analytic approaches to estimating growth curves. We limit this presentation to methods for continuous outcomes, although there are a whole class of methods that address questions about when events occur (see Singer & Willett, 2003, for
an introduction to survival analysis), predictors of discrete events over time
(see Diggle, Liang, & Zeger, 1994, for an introduction to generalized es-
timating equations and longitudinal logistic regression), or transition over
time in latent class membership (see Clogg, 1995; Collins, 2006, for a dis-
cussion of latent class analysis methods).

GOALS OF LONGITUDINAL ANALYSES

The primary goal of longitudinal analysis of repeated measures and
growth curve analyses in particular is to describe patterns of change over
time. We would like to chart individual developmental patterns. Ideally, we
would like to describe the underlying developmental trajectory (Burchinal
& Appelbaum, 1991; Blanton & Jaccard, 2006). This trajectory provides
the mathematical description of the development of a given attribute over time.
The trajectory specifies the type of equation or form of the developmental
trajectory, and is uniquely defined for a given person by that individual’s
parameters. Nonlinear functions such as exponential or logistic functions
have been very useful because they describe change that is very rapid at first
and then slows down until the final level is reached and growth stops
(Willett, Singer, & Martin, 1997; Vonesh, 1996). The parameters of that
trajectory describe important characteristics of growth such as the age at
which the growth spurt starts, the rate of change during the growth spurt,
and the final level of the attribute at the end of the growth spurt. Other
functions, such as polynomial growth curves, provide reasonable descrip-
tions of growth and are especially useful when our measurement of the
underlying attribute is not precise or when relatively few repeated assess-
ments were collected (Singer & Willett, 2003). These functions have pa-
rameters that describe the initial or mean level of the attribute, rate of linear
change over time, and rates of higher order patterns of change (e.g., quad-
ratic, cubic, etc.). Higher order polynomial growth curves can describe
growth patterns in which the rate of change is not constant over time.

For example, height is an attribute with a developmental trajectory that
is well described. We can measure height accurately on a ratio scale, and our
measurement reflects exactly how tall the individual is at that time. We
know that changes in height over time can be well described with a series of
linked nonlinear functions (Bock, Wainer, Peterson, Thissen, Jurray, &
Roch, 1973). The first three-parameter logistic nonlinear trajectory de-
scribes the infant growth spurt between birth and 3 years of age. Growth is
very rapid during early infancy and slows as the child approaches his or her
third birthday. Growth tends to be linear between the end of the infant
growth spurt and the beginning of the adolescent growth spurt. The ad-
olescent growth spurt is well described, again, by a three-parameter logistic
nonlinear function. Children tend to follow the same growth pattern, so population curves have been estimated and are used by pediatricians as an index of well-child development, to determine whether a given child seems to be growing as expected, and also to compare groups of individuals or identify predictors of developmental patterns.

In addition to describing development, we want to identify predictors of patterns of change. We would like to know what personal, family, school, or community characteristics are related to individual developmental patterns. Are there differences between groups in terms of rate of change or level of development? Are there characteristics of the individual, schools or employment, family, or community that predict developmental trajectories? For example, if you were studying height, you might want to know if smaller children at birth show a bigger growth spurt during infancy.

LIMITING FACTORS IN DESCRIBING PATTERNS OF CHANGE

Our ability to measure the outcome often limits the extent to which we can describe patterns of change. It is far more difficult to describe development of many psychological attributes other than height because they cannot be measured as accurately as height or weight (Blanton & Jaccard, 2006). We have good instruments for measuring characteristics such as language, intelligence, academic achievement, behavior problems, and prosocial skills. These measures have good reliability and reasonable validity, but they provide approximate, not exact or isomorphic, assessments of the trait. When measurement is approximate, we can meaningfully describe individual differences in developmental patterns based on norm-referenced or criterion-referenced assessments even when we cannot describe the underlying developmental trajectory accurately. Growth curve analyses on instruments that provide approximate, not exact, measurement can tell us a great deal about both intra-individual developmental trajectories and inter-individual differences in patterns of change.

Sample size and number of repeated assessments often limit our ability to describe individual patterns of change. Most longitudinal research includes relatively few repeated measures on most individuals and modest to moderate sample sizes. The accuracy of our statistical model to describe an individual’s data is extremely limited when there are only a few repeated assessments. Similarly, our ability to identify individual differences is limited when our samples are small. Longitudinal research is expensive and humans tend to grow slowly, so it is very costly to collect large numbers of repeated assessments on large samples. Statistical methods vary in their ability to provide precise and valid test statistics when fit to short time series or with small samples. Accelerated longitudinal designs (i.e., designs in
which multiple cohorts are observed for shorter periods of time allow for describing growth across all ages accessed) can provide efficient means to describe developmental patterns even with relatively few assessments per individual (see Graham, Taylor, & Cumsville, 2001 for details).

Finally, any longitudinal method must account for correlations among repeated measures. Almost all repeated assessments of humans are correlated because skill levels at one time are correlated with skill levels at other times. It is this dependency that is being modeled when we estimate the developmental trajectory. Failure to adequately account for these correlations often renders test statistics invalid because variability is often underestimated, resulting in overestimating statistics (Tabachnick & Fidell, 2007). When that happens, $p$-values as an index of whether associations between predictors and developmental patterns are not statistically valid. Longitudinal methods vary markedly in their ability to appropriately account for correlations in repeated measures.

**STATISTICAL ASSUMPTIONS FOR LONGITUDINAL MODELS**

Most statistical methods for testing hypotheses share a common set of assumptions that need to be met when analyzing both cross-sectional and longitudinal data (Tabachnick & Fidell, 2007). These assumptions are presented and then many of them are discussed in the context of specific methods for analyzing longitudinal data. The mostly commonly used cross-sectional and longitudinal methods assume that the developmental outcome variables are normally distributed and measured at the interval or ratio level; they may also be a transformation of those variables that is normally distributed. This assumption can be relaxed somewhat, but variables need to have roughly symmetric, unimodal distributions in which more data are in the middle of the distribution than at either end of the distribution. The scale of measurement can be ordinal as long as there are at least four levels represented within the data and the distribution is roughly normal (Gebotys, 1993). Parameter coefficients cannot be trusted to describe population tendencies when the outcome has a distribution that is markedly skewed or when very few distinct levels are represented in the data. A wide variety of methods also exist for analyzing data with other distributions but require more specialized techniques (e.g., autoregressive models to examine stability and change, poison regression for count data, exponential or logistic methods for nonlinear data, censored models for data with a large number of values at the minimum value; see McArdle & Nesselroade, 2003, for details). Similarly, most longitudinal methods require an appropriate between-subjects model that includes relevant main effects and interactions (Tabachnick & Fidell, 2007). The model coefficients
are biased when important covariates or interactions are omitted. Finally, it
is also assumed that at some level in the data there are independent units of
observations. The test-statistics’ \( p \)-values are usually too small when inde-
pendence of observations is falsely assumed. The consequence of violating
any of these assumptions is that test statistics are not valid and cannot be
trusted to indicate whether predictors relate to patterns of change.

Several additional assumptions must also be met when repeated meas-
ures are analyzed with methods that test hypotheses about change over
time. First, it is assumed that you have specified a longitudinal model that
can adequately describe change over time for individuals (see Burchinal &
Appelbaum, 1991 or Singer & Willett, 2003, for more detail), but the type
and level of that model is limited by the number of repeated assessments in
several ways. With two repeated assessments, we can only estimate a sep-
arate intercept or slope for each individual, but not both (note: while some
methodologists such as Rogosa, Brandt, & Willett, 1982, argue that you
must have more than two repeated assessments to use longitudinal meth-
ods, all argue that you must take the correlation among repeated measures
even with two assessments). With three repeated measures we typically es-
timate an individual linear growth curve, with a separate intercept and slope
that describe rate of change with respect to time or age for each individual.
With four repeated assessments, we can choose between nonlinear func-
tions with three parameters such as the exponential growth curve or the
quadratic polynomial function. For example, the three-parameter expo-
nential function describes monotonically increasing change as a function
time of onset of growth spurt, maximum rate of change, and a final, or
asymptotic, level. In contrast, the quadratic polynomial function describes
change as a function of level, linear change, and quadratic change, and it
estimates a parabolic function. With five or more repeated assessments, we
can fit nonlinear functions such as logistic growth curves or higher order
polynomial functions such as cubic growth curves. These functions can de-
scribe change that is S shaped. Thus, it is important to have enough re-
peated assessments to permit the estimation of hypothesized individual
growth curves.

It is imperative that the selected growth curve model accounts for the
pattern of correlations within repeated assessments to rely on the test sta-
tistics to identify predictors of developmental patterns (Singer & Willett,
2003). A model will completely account for these correlations if the number
of parameters estimated is one less than the number of repeated assess-
ments. On the other hand, a model with substantially fewer parameters than
repeated assessments will provide more power for hypothesis testing, as-
suming the model accurately describes developmental change. We offer a
couple of recommendations to balance power and model adequacy based
on our years of statistical experience. First, when the number of repeated
measures is three or fewer, we recommend that the number of parameters in the initial individual growth curve model be close to one minus the number of repeated measures. Second, when the number of repeated measures is four or greater, we suggest fitting a preliminary model that is at least one degree higher than the hypothesized polynomial growth curve model to test whether the hypothesized model is adequate. For example, if one believes change is approximately quadratic and there are at least four repeated measures, then you can fit a cubic model in a preliminary analysis. If the random variance and associated covariances for the cubic term are inestimable or nonsignificant, then you should refit the model as a quadratic model.

A related issue that must be addressed is deciding whether the population and individual growth curves have the same or different growth curve models. Typically, the same type of growth curve model is estimated to describe both individual and group patterns of change, but this is not necessary or even desirable in all cases. The individual growth curve parameters describe the extent to which that individual differs from the population for that index of change (e.g., mean level, slope), whereas the group growth curve parameters describe the overall shape of the population curve. Therefore, individual differences can be relatively minor for higher order terms in a polynomial model even when the group-level parameter is clearly needed to describe patterns of change. In this case, omitting the higher order term from the group model would bias all of the other parameter estimates, but including this term in the individual model decreases power and often may result in a model that cannot be estimated. Although uncommon in psychology, use of different individual and group growth curve models is a common approach in biostatistics. A leading statistician demonstrated that blood pressure can be characterized by a quadratic group curve and a linear individual curve because the overall shape of the curve was quadratic and there were marked individual differences in both the intercept and slope (Laird & Ware, 1982). Conversely, it is possible that substantial individual differences exist for higher order polynomial terms, but those terms average to zero. In that case, the population curve would not show that level of curvature even though individual growth curves would. In developmental psychology, Burchinal, Campbell, Bryant, Wasik, and Ramey (1997) published an example of the use of a quadratic group growth curve model and a linear individual growth curve model. The individuals had at least three repeated assessments and the developmental pattern was quadratic, but systematic individual variability in the quadratic term was trivial. Careful attention to both population growth curve coefficients and individual growth curve variances and covariances is necessary to identify the appropriate longitudinal model.
Another measurement assumption that must be met with longitudinal data analyses is that all repeated assessments must measure the same attribute in the same metric over time. An example of an attribute that meets this assumption is vocabulary acquisition (Huttenlocher, Vasilyeva, Cymerman, & Levine, 2002), and an example of an attribute that might violate it is cognition during infancy. Cognitive tests of young infants involve a major psycho-motor component whereas tests of older individuals involve a major verbal component (Neisser, Boodoo, Bouchard, Boykin, Brody, Ceci, Halpern, Loehlink, Perloff, Steinberg, & Urbina, 1996). Furthermore, the scores from these assessments need to be in the same metric over time. Literally, it is assumed that a one-point change in the score means the same thing across the entire scale and over age. All growth curve methods rely on computing differences in scores over time, and those differences are meaningful to the extent that the metric of the assessment is consistent. Therefore, the scale needs to be consistent, which raises concerns about using raw scores. For this reason, we believe developmental age scores, Rasch scores (see Chapter II for details), or even standard scores can be very useful for describing inter-individual differences in developmental patterns.

Some, but not all, growth curve methods require that all individuals be measured at the same time points or age. Many studies attempt to measure all participants at specified intervals, creating panel data. All growth curve methods can be used with such time-structured data. Time-structured data are repeated measures data in which individuals are measured at specific time points or ages, and thus time or age can be treated as a categorical variable. Panel data are an example of time-structured data. In contrast, time can be viewed as treating time or age as a time-varying variable in data in which age at assessment is allowed to vary, either deliberately or because some data were collected late. A few methods can use the actual age or time at which data were collected as a time-varying predictor, and the accelerated longitudinal design is example of that kinds of model that deliberately plans for missing data (Graham et al., 2001).

Finally, some longitudinal methods allow the inclusion of individuals with missing data when the assumption that data are ignorably missing is met (see the chapter by Widaman for a comprehensive discussion of issues related to analyzing data with missing data). Missing data almost always occurs in longitudinal research. Data are ignorably missing when the reason the data are missing is related to the predictors and is not related to the outcome variable of interest (Schafer & Graham, 2002). For example, missing data due to attrition are probably ignorably missing, but data missing due to failure to achieve a basal score on the test are not ignorably missing. Some of the growth curve methods described below allow for the inclusion of individuals with some missing data. How the statistical method handles
missing data becomes irrelevant if missing data imputation methods such as multiple imputation are used (see chapter by Widaman for details).

GROWTH CURVE METHODS

A wide variety of growth curve methods can be easily implemented with commercial software, but selection of the most appropriate methods depends on many factors. The various approaches differ in whether they allow for individual differences in intercept or rates of change (e.g., slope) and, correspondingly, whether they require a priori specification of the shape of the longitudinal trajectories. They also differ in whether they allow for ignorably missing data and for repeated measures of predictors (also called time-varying predictors) along with repeated measures of the developmental outcome. In the following pages, we briefly discuss some of the most widely used approaches, including univariate and multivariate repeated measures analysis of variance, HLM, LGC, and prototypic or mixture growth models. Table 1 provides a summary of each method’s attributes and lists the currently available software that can be used to implement the method.

<table>
<thead>
<tr>
<th>Models Allow For:</th>
<th>Type of Growth Curve Method</th>
<th>Univariate</th>
<th>Profile</th>
<th>HLM</th>
<th>LGC</th>
<th>Prototypic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Individual differences</td>
<td></td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Intercept</td>
<td></td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Slope</td>
<td></td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Specifying within-subjects model</td>
<td></td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Time-varying covariates</td>
<td></td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Ignorably missing data</td>
<td></td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Some</td>
<td>Some</td>
</tr>
<tr>
<td>Assessment times that vary</td>
<td></td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Some</td>
<td>No</td>
</tr>
<tr>
<td>Measurement error in predictors</td>
<td></td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Test statistics are “valid” for moderate to large samples</td>
<td></td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>NA</td>
</tr>
</tbody>
</table>

Useful Software:

- SAS (procedure) GLM GLM Mixed HLM CALIS Cluster
- SPSS (procedure) GLM GLM HLM AMOS Cluster
- HLM Yes
- MPlus Yes Yes Yes
- AMOS Yes
- Trajectory SAS Macro

TABLE 1
CHARACTERISTICS OF SELECTED GROWTH CURVE METHODS
To illustrate the various growth curve methods, the simulated data and data from a published longitudinal study were analyzed. The simulated data represents developmental data from six children who were to be measured every six months between 24 and 60 months. Because most longitudinal studies are unable to assess all children at all occasions, we have allowed for missing data (i.e., ignorably missing). Figure 1a plots the data for these six children. Most children have some missing data and they vary in level and rate of change over time. The other example data set is from the Abecedarian and CARE Projects, conducted at the Frank Porter Graham Child Development Institute (Burchinal et al., 1997). At birth, 161 low-income children were randomly assigned to a child care intervention or control group. Table 2 contains descriptive statistics of the sample and Figure 2a shows plots of scores for randomly selected individuals.

Figure 1.—Estimated growth curves from example data: (a) longitudinal data; (b) “univariate” growth curves; (c) profile analysis growth curves; (d) hierarchical linear models (HLM) growth curves; (e) latent growth curves (LGC); and (f) prototypic growth curves.
Children’s intelligence was tested annually between two and five years of age and then every 18 months after entry to school. About one-third of the children missed at least one assessment, but none of the missing data was due to factors associated with the child’s intelligence (e.g., can be considered ignobly missing). In most cases, it was due to the family’s temporary relocation from the area. We also measured aspects of the child’s life that were changing (i.e., quality of home environment as measured by the Home Observation for Measuring the Environment [HOME; Caldwell & Bradley, 1984]) and aspects we presumed to be unchanging (e.g., maternal IQ). This example was used to demonstrate how the various methods deal with missing data, time-varying covariates, and appropriately adjust for correlations in the repeated assessments.

**Univariate Repeated Measures Approach**

The original repeated measures analysis method is called the univariate repeated measures analysis of variance or mixed models. It was widely used by developmentalists until about 1980 and continues to be used by other psychologists from other disciplines. This method assumes there are

---

### TABLE 2

**CHILD CARE INTERVENTION STUDY DATA**

<table>
<thead>
<tr>
<th>Age (in years)</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6.5</th>
<th>8</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>N</strong></td>
<td>163</td>
<td>160</td>
<td>159</td>
<td>155</td>
<td>150</td>
<td>174</td>
<td>174</td>
</tr>
<tr>
<td><strong>Home</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>M</strong></td>
<td>.61</td>
<td>.66</td>
<td>.71</td>
<td>.74</td>
<td>.74</td>
<td>.69</td>
<td>.69</td>
</tr>
<tr>
<td><strong>SD</strong></td>
<td>.12</td>
<td>.13</td>
<td>.12</td>
<td>.11</td>
<td>.11</td>
<td>.09</td>
<td>.09</td>
</tr>
</tbody>
</table>

**Child IQ by child care treatment**

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Stanford Binet</th>
<th>WPPSIWISC-R</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>M</strong></td>
<td>97.1</td>
<td>99.8</td>
</tr>
<tr>
<td><strong>SD</strong></td>
<td>11.3</td>
<td>12.9</td>
</tr>
<tr>
<td>Control</td>
<td>85.7</td>
<td>93.6</td>
</tr>
<tr>
<td><strong>M</strong></td>
<td>101.8</td>
<td>102.4</td>
</tr>
<tr>
<td><strong>SD</strong></td>
<td>12.9</td>
<td>10.5</td>
</tr>
<tr>
<td>4</td>
<td>101.1</td>
<td>12.9</td>
</tr>
<tr>
<td>5</td>
<td>102.4</td>
<td>10.5</td>
</tr>
<tr>
<td>6.5</td>
<td>99.8</td>
<td>12.9</td>
</tr>
<tr>
<td>8</td>
<td>98.4</td>
<td>12.0</td>
</tr>
<tr>
<td>12</td>
<td>95.8</td>
<td>10.2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Stanford Binet</th>
<th>WPPSIWISC-R</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>M</strong></td>
<td>94.6</td>
<td>93.6</td>
</tr>
<tr>
<td><strong>SD</strong></td>
<td>14.0</td>
<td>13.0</td>
</tr>
<tr>
<td>Control</td>
<td>90.8</td>
<td>12.4</td>
</tr>
<tr>
<td>5</td>
<td>94.6</td>
<td>14.0</td>
</tr>
<tr>
<td>6.5</td>
<td>93.6</td>
<td>13.0</td>
</tr>
<tr>
<td>8</td>
<td>92.2</td>
<td>12.4</td>
</tr>
<tr>
<td>12</td>
<td>90.6</td>
<td>11.5</td>
</tr>
</tbody>
</table>

*Note.* —40% of sample received the child care treatment. HOME was not administered at 6.5 or 12 years, so previous score was carried forward.
individual differences in the level or intercept of repeated measures, but not in rates of change over time. As a consequence, this method underestimates variability and overestimates test statistics when individuals have different slopes. This limitation has been widely recognized and various approaches have been developed to provide more appropriate test statistics (see Tabachnick & Fidell, 2007, for a discussion). These approaches require deleting the data from any individual who was not observed at all time points, estimating the degree of correlation among the repeated assessments, and adjusting the degrees of freedom for error accordingly.

Growth curves for the example data sets were estimated using this approach. First, the simulated data were fit to the model without the adjustment and results are shown in Figure 1b. The growth curves are constrained to have the same slope but allowed to have different intercepts. Thus, the individual growth curves with the simulated data are poorly estimated because the marked differences in slopes are ignored. Second, the child care intervention study was analyzed with univariate growth curve methods, using the corrections for type-I error inflation (Tabachnick & Fidell, 2007). This deletes all data for individuals who have any missing data on variables used as independent or dependent variables. Despite the fact that data were collected on 162 children, only the 139 subjects with

![Figure 2](image-url)
complete data could be included in the analysis. Time-varying predictors are not allowed, so we computed the mean HOME score from our repeated assessments, thereby creating a time-invariant predictor. The analysis included the child care treatment, maternal IQ, the mean of the repeated HOME scores, age at assessment as a categorical variable (i.e., the six dummy variables representing age contrasted each of the first six assessment ages with the final assessment age) and interactions between the other variables and age. The analysis ignored all individual differences in rates of change over time and poorly estimated the main effects for age and interactions involving age. The test statistics are shown in the first column of Table 3. As you can see, this analysis indicated that the child care treatment, $F(1, 135) = 29.6$, $p < .001$, maternal IQ, $F(1, 135) = 10.5$, $p < .001$, and HOME, $F(1, 135) = 24.2$, $p < .001$, were related to the level of the children’s IQ, but only the child care treatment was related to patterns of change over time, $F(6, 135) = 7.7$, $p < .001$. The analysis treated time as a categorical variable, so the tests for age were across all ages and did not allow us to distinguish among linear, quadratic, or higher order patterns of change.

**Multivariate Repeated Measures Analysis**

Once computers became faster and more accessible about 30 years ago, we were able to switch from these univariate growth curve methods to the
multivariate repeated measures approach, or profile analysis (Tabachnick & Fidell, 2007). The multivariate approach was a marked improvement over the univariate growth curve methods because individual differences in slopes were allowed. That is, separate growth curves are estimated for each individual, and those individual growth curve parameters are used to estimate group growth curves. The individual growth curve parameters are estimated using ordinary least squares, and the group growth curve parameters are estimated as unweighted or simple means of the corresponding individual growth curve parameters. For example, the intercept for a given group is estimated as the mean of the individual intercepts for all individuals in that group. This flexibility in modeling produces valid test statistics when sample sizes are moderate to large and appropriate polynomial models are selected to describe individual growth curves. However, the multivariate approach also has major limitations. Like the univariate approach, this method also excludes individuals with any missing data and cannot easily incorporate time-varying covariates (see Table 1).

Profile analyses were conducted using the simulated and intervention study data. Using the simulated data, growth curves are estimated only for the two individuals with complete data (see Figure 1c), but those two individuals were allowed to have different growth curves. Using the child care intervention study, the profile analysis included data from 139 children in the sample, excluding children with missing data. Orthogonal contrasts among the dependent variables computed linear and quadratic trends over time and allowed us to test whether the selected predictors were related to either linear or quadratic change. The \( F \)-tests from the analysis are shown in the second column of Table 3. Figure 2b shows the estimated growth curves for the treatment and control groups. As in the univariate method, the time-varying predictor, HOME scores, had to be averaged and included as a between-subjects predictor. The \( F \)-tests for child care treatment, maternal IQ, and HOME are constrained to be identical under the univariate and multivariate approaches (see columns 2 and 3 in Table 3). This is because the main effects with both models test for differences related to those factors in the individual growth curve intercept estimated as the mean of the repeated assessments for that individual. In contrast, the two approaches describe patterns of change quite differently because the profile analysis allows for individual differences in individual growth curve linear and quadratic slopes, using a multivariate test to ask if both linear and quadratic change is related to the predictors. The two child care groups differed in their linear, \( F(1, 135) = 14.3, p < .001 \), and quadratic, \( F(1, 135) = 18.9, p < .001 \), rates of change over time. Further, higher maternal IQ was associated with more linear gains over time, \( F(1, 135) = 4.8, p < .05 \), and higher averaged-over-time HOME scores were associated with more curvature, \( F(1, 135) = 6.1, p < .05 \).
Hierarchical Linear Models (HLM)

As computers became even faster and more accessible in the 1980s, methods that addressed many of these limitations of the univariate and multivariate approaches became feasible. General linear mixed models, called HLM by social scientists, were developed to describe intra-individual developmental patterns and identify inter-individual predictors of developmental patterns (Laird & Ware, 1982; Bryk & Raudenbush, 1992; Raudenbush & Bryk, 2002; Singer & Willett, 2003). Like the multivariate growth curve approach, HLM estimates an individual growth curve for each individual and a group growth curve from the individual growth curve parameters. Unlike the multivariate growth curve approach, the individual growth curves are estimated using empirical Bayesian or maximum likelihood methods and are weighted to include information from both the individual’s and the entire sample’s data. Individual growth curves are smoothed toward the group growth curve if they appear too different under the assumption that error accounts for the reason that the individual curve seems deviant. That is, group growth curve parameters are estimated as the weighted mean of the corresponding individual curve parameters from individuals in that group. Larger weights are given to individuals with more repeated assessments and smaller weights to the individuals whose data appear deviant compared with the rest of their group. This smoothing of individual curves and weighting of group curve parameters has been referred to as borrowing strength and can greatly increase the precision of parameter estimates and power to identify predictors of developmental patterns (Raudenbush & Bryk, 2002; Singer & Willett, 2003).

The advantages to using this approach can be seen in Table 1 and Figure 1d. Individual growth curves are allowed to vary in terms of both intercepts and slopes. This approach can be used when data are ignorably missing, eliminating the need to delete individuals with missing data. It can easily include time-varying covariates and assessment times that vary across individuals. The methods are asymptotic, but test statistics should be valid when sample sizes are moderate to large and distributional assumptions are met. The necessary sample size will depend on the number of individuals in the sample, the number of repeated assessments, the distribution of the variables, and the number of parameters estimated in the model. As a very rough guideline, group sizes of 30–50 should provide reasonable test statistics when variables are normally distributed and five or fewer variables are included to predict intercepts and slopes.

HLM analyses of the two example data sets demonstrate these characteristics. Figure 1d shows that we were able to estimate individual growth curves for all six individuals and they were allowed to vary in level and rates
of change over time. The HLM analysis of the child care intervention data predicted IQ trajectories with linear individual growth curves, quadratic group curves, time-varying assessments of the HOME, and a single time-invariant measure of maternal IQ. As shown in the third column of Table 3, the children who received the child care intervention tended to have higher IQ scores over time, $F(1, 464) = 30.8, p < .005$, to show more linear gain over time, $F(1, 464) = 35.1, p < .001$, and have more curvature in their trajectories, $F(1, 464) = 18.4, p < .001$. The treatment and control group growth curves were almost identical to those estimated using the multivariate approach shown in Figure 2b so an additional plot is not provided. In addition, results suggested that maternal IQ was directly related to children’s IQ trajectories, $F(1, 464) = 34.5, p < .001$, but was not related to patterns of change over time when the concurrent HOME was also considered. The HOME was related to children’s IQ overall, $F(1, 464) = 4.7, p < .05$, and to linear increases over time, $F(1, 464) = 9.2, p < .01$. Compared with the multivariate analysis approach, HLM indicated a different role for maternal IQ and the HOME in predicting cognitive trajectories because it related the HOME scores to the IQ scores from the same assessment time, whereas the multivariate approach included the HOME as the averaged-over-time score for each family. Finally, using a series of analyses to test mediation (Mackinnon, Lockwood, Hoffman, West, & Sheets, 2002), we were able to provide evidence supporting a mediated path from maternal IQ through HOME to cognitive development (Burchinal et al., 1997).

**Latent Growth Curves (LGC)**

LGC have been proposed and can be estimated with widely used software. This approach has been shown to be equivalent to HLM in approach and estimation of individual growth curves (Curran, 2003). Muthén and his colleagues (Muthén & Curran, 1997; Muthén & Muthén, 2000; Muthén, 2004) popularized a structural equation modeling (SEM) approach in which fixed paths in the measurement model estimate individual growth curves and estimated paths in the SEM describe direct and indirect associations among the latent variables. The model assumes that all individuals were measured at each time point, and estimates a latent intercept and slope for each individual by fixing the loadings of the observed variables on these latent variables. The most recent version of the SEM software MPlus allows for both ignorably missing data for time-structured data and for data in which assessment times are allowed to vary among individuals (Muthén, 2004). The loading of all repeated assessments is constrained to “1” to specify the latent intercept, and the latent linear slope loadings are constrained to the time of assessment. Nonlinear slopes can be specified, and alternative time metrics such as orthogonal polynomials can be used.
The LGC approach has numerous strengths (see Table 1). It allows for time-varying covariates by specifying latent intercepts and slopes in the measurement model for each time-varying covariate, and allowing the latent intercepts and slopes to be related across multiple time-varying variables. Essentially, both level and change in one variable can be used to predict level and change in other variables. Error in assessment is considered in the measurement model, and errors can be allowed to correlate.

This approach offers major advantages and a few disadvantages over other methods. The major advantages of this approach compared with all other approaches include its ability to account for some error in predictors and to test mediation hypotheses. The major disadvantage of this approach in easy-to-implement software packages includes its reliance on time-structured data and a slight reduction in power compared with HLM due to the fact that the HLM approach smooths the estimates parameters of the individual growth curves.

LGCs are presented using the simulated data and the child care intervention data. To compare the LGC and HLM approaches, we made two decisions in fitting the models. First, Figures 1d and e show that the estimated individual growth curves from the LGC and HLM approaches are identical. This occurred because we intentionally fixed the loadings in the LGC approach to equal those under the HLM approach. Second, we fitted a reduced model to the child care intervention data rather than the more complicated model that included several time-varying covariates that has been previously published using HLM (Burchinal et al., 1997). It was necessary to reduce the number of estimated parameters under the LGC approach because of the small sample size.

Results from the LGC are shown in the final column of Table 3. Individual quadratic growth curves were estimated by fixing paths to the repeated IQ assessments, using a value of “1” as the designated path for the intercept, age in years at assessment minus five (i.e., age at the end of treatment) as the designated path for the linear slope, and that age squared as the designated path for the quadratic slope. LGC were estimated for IQ and HOME scores. The treatment and control group growth curves were almost identical to those estimated using the multivariate and HLM approach shown in Figure 2b, so an additional plot is not provided. Overall, the results are quite similar to the multivariate and HLM analysis results, suggesting that child care treatment, maternal IQ, and the quality of the home environment were related to the overall level of intellectual development, and that the child care treatment and maternal IQ were related to patterns of change. Findings regarding the main effect of the HOME on IQ are more similar for the multivariate and LGC analyses than for the HLM analysis. This is because both the multivariate and LGC analyses included
the time-varying predictor, HOME, averaged over time for family. The LGC measurement model for the home estimates the HOME intercept as the mean of the repeated HOME scores for a given individual, producing the same under the model that we created when used the mean HOME scores in the multivariate analyses. In contrast, the HLM analysis linked each HOME score to the IQ score from that assessment point. Additionally, the LGC extends the findings from the multivariate and HLM analyses by testing the association between linear change over time on the HOME and IQ development. Furthermore, unlike the HLM analysis, we were able to test the mediation of maternal IQ through HOME (i.e., an indirect path) within the LGC model and did not have to use additional models to derive the indirect path as was necessary for HLM.

**Prototypic or Mixture Growth Curves**

The prototypic growth curve methods are person-oriented analyses of longitudinal data. They are based on the assumption that there are a small number of qualitatively different LGC that underlie development within the populations sampled (Nagin, 1999; Muthén, 2001). It is assumed that the LGC differ markedly from each other and account for much of the individual differences observed in patterns of change over time.

At least three approaches have been used to estimate prototypic growth curves. The p-type factor analysis was used in the 1960s and 1970s (e.g., the seminal monograph on cognitive development by McCall, Appelbaum, & Hogarty, 1973), but has not been used recently. This type of factor analysis involves factor analysis of the correlation matrix, estimating combinations of patterns of change on a single variable across people instead of factors across variables. Cluster analysis of longitudinal data seeks to identify homogeneous subgroups of people who show similar patterns of change over time on a single variable within groups and qualitatively different patterns across groups.

Semiparametric mixture models (also known as trajectory analysis) as popularized by Nagin and Tremblay (1999) has become a very popular method. Trajectory analysis assumes the presence of distinct groups with differing growth trajectories within the population. A polynomial or nonlinear model is used to relate age to the outcome. The method allows for missing observations and censored measurement distributions that result in clustering at the scale's minimum or maximum. This method requires the analyst to specify the number of groups present in the population. Essential to this method are posterior probability estimates of group membership that are produced for each individual (i.e., the probability of belonging to each group). The individual is consequently assigned
to the group for which they have the highest probability of membership. Once group membership has been assigned, multinomial regression can be performed to investigate the relationship between covariates and the group growth trajectories. Each individual’s probability of membership in their assigned group is used as the regression weight in the multinomial analysis.

A similar approach, developed by Muthén and Muthén (2000), is also quite popular. Their approach classifies each individual into one prototypic group and allows for individual variability within the prototypic group. Because the prototypic groups are defined within a measurement model, this approach also allows for modeling differences among prototypic groups within a SEM.

Analyses of the two example data sets using Nagin’s semiparametric mixture growth approach (Nagin, 1999) are presented. Analysis of the simulated data, shown in Figure 1f, reveals that two prototypic patterns describe the trajectories of all six children and values were assigned to each child that represented the extent their individual growth curves were similar to the each of the two prototypic curves. The trajectory analysis of the intervention study data included six latent growth trajectories based on the Bayesian Information Criterion. The analysis procedure estimated these six trajectories and assigned six probability values to each child representing the likelihood that the child’s individual growth curve resembled each LGC. Each child was assigned to the latent curve with the largest probability value. Figure 2c presents the estimated and actual mean growth trajectories for each group. Group 1 represents a small group of children \((n = 5)\) with high IQ scores at age 2 who continued to have high test scores throughout the observation period. Children in Group 2 are children who tested above average early but consequently regressed toward an average score by age 8. Groups 3 and 4 consist of children with similar below average test scores (approximately 90) at age 2 but later diverge such that group 3 shows increasing test scores which eventually achieve average scores. A similar pattern was observed for groups 5 and 6. The initial test scores are considerably lower (approximately 75), but group 5 test scores increase, while group 6 test scores remain nearly constant.

Of special interest for this analysis was identifying factors that discriminate between children showing different patterns of change over time (i.e., groups 3 and 5 vs. groups 4 and 6, respectively). To investigate these factors, we performed a multinomial analysis predicting group membership from child care intervention, mother’s IQ, and the averaged-over-time HOME scores. Child care treatment, mother’s IQ, and the HOME score were all significantly predictive of membership in group 3 (below average-to-increasing) compared with group 4 (below normal-to-maintaining). Children who received the child care treatment were four times more likely to be in group 3 than group 4.
**Comparison of Methods**

The degree to which the mixture models (i.e., person-centered approaches) are distinct from HLM or LGC (i.e., variable-centered approaches) has been debated. It is clear that all growth models are sensitive to distributional violations, but the mixture growth approaches are particularly vulnerable (Bauer & Curran, 2003). Some argue that all growth modeling methods should be labeled person-centered because the analysis is based on the repeated assessments of the individual (Molenaar, 2004), while others argue that the growth mixture models are convenient representations of complex individual trajectories (Eggleston, Laub, & Sampson, 2004). To some degree, such debate can be practically resolved if attention is paid to selecting the method most closely aligned with the theoretical model of change (Collins, 2006) and that provides the most parsimonious and clear interpretation of the findings (Cudeck & Henly, 2003).

The five growth curve methods differ in many ways but tend to yield similar conclusions when used to analyze these two data sets. The univariate and prototypic/person-oriented approaches provide the poorest estimation of individual growth curves but provide interesting groups showing different developmental patterns. Accordingly, the HLM and LGC approaches provide the best estimation of individual growth curves and will have the most power to identify predictors of individual developmental patterns unless there are unspecified subgroups in the sample that show very different patterns of change over time.

The univariate approach is clearly the least desirable method for estimating individual growth curves. Theoretically, it does not allow for individual differences in rates of change over time and, accordingly, will result in poorly specified growth curve models and associated test statistics. There seems to be no real advantage to using this approach in today’s world of fast computers. The multivariate approach improves upon the univariate approach but still is more restrictive than either the HLM or LGC approaches. The inability to easily accommodate time-varying predictors or individuals with some ignorably missing data makes this approach less desirable.

The mixture growth model identified latent profiles, not individual developmental trajectories. As such, it answers a different question. It does not allow for individual differences in developmental patterns among children who are classified as showing the same LGC. Furthermore, the ability to identify predictors of developmental patterns was compromised by the classification of individual longitudinal data into categorical data in our example. It is not surprising that we had less power to identify predictors of intellectual developmental patterns when the outcome measure is a categorical variable than when the outcome variables are individual growth curve intercepts and slopes.
The HLM and LGC approaches provide the most flexible and precise estimates of individual developmental patterns and identification of predictors of developmental patterns. There are minor differences in how the HLM and LGC approaches estimate individual growth curves using the defaults in commonly used software, but larger differences in how they relate the predictors to patterns of development. Both the HLM and LGC estimate a separate intercept and slope for each child. The LGC approach estimates individual developmental patterns for each longitudinal outcome and predictor and relates levels and rates of change in the outcome and the predictors; whereas the HLM approach links level and rate of change in the outcome variable to co-occurring changes in the predictors.

The HLM and LGC approaches also differ in terms of their power to detect mediators and moderators (Singer & Willett, 2003). Currently, only the HLM approach can be used with easy-to-use software when data are not time structured. The HLM approach is ideal for identifying moderators. It is easy to create interaction terms by crossing either categorical variables representing groups of interest or continuous variables. There is more power to detect interactions involving rates of change when default options are used with the HLM approach in most circumstances because the HLM approach smoothes by estimating individual growth curve parameters from both the group’s and individual’s data whereas the LCG approach estimates those parameters only from the individual’s data. In contrast, the LGC approach provides considerably power to detect mediators. Only the LGC approach provides to estimate the indirect path suggesting mediation within a single analysis model.

Despite the many differences of the selected methods, it is reassuring that similar conclusions were reached when all five approaches were used to analyze the data from a child care intervention study. In all analyses, higher IQ scores were linked to the child care treatment and were related to both maternal IQ and the family environment. The models that allowed for individuals to show different rates of change over time actually provided better identification of predictors of development. The HLM approach clearly suggested that children in the child care treatment group showed less change over time in IQ, whereas the LGC approach provided clearer evidence suggesting that quality of the family environment mediated the association between maternal and child IQ.

In conclusion, growth curve analyses provide developmentalists with powerful statistical methods for describing individual patterns of development and for identifying predictors of individual developmental patterns. Careful selection of the type of method involves considering characteristics of the data and the model. One must consider whether the attribute is measured with sufficient precision to estimate actual developmental functions or approximate developmental patterns, and whether the
measurement resulted in ignorable missing data, time-structured data, time-varying covariates, and error in predictors and outcomes. It is important to carefully select both the model that describes individual developmental patterns and the model that identifies predictors of development. Model selection must also take into account whether the primary questions involve mediators or moderators. Consideration of each of these factors can result in the selection of the growth curve methodology that should provide the most complete description of individual patterns of development and the most powerful approach to identifying predictors of development. The optimal statistical approach represents the theoretical model of change by describing the developmental patterns dictated by that model using the data collected by the developmental project.

ACKNOWLEDGMENT

We would like to thank Frances Campbell for allowing us to use examples from the Abecedarian Project and its follow-up project, Project CARE, and the many investigators for whom we have provided data analysis over the past 25 years.

REFERENCES


Examining Relationships Between Where Students Start and how Rapidly they Progress: Using New Developments in Growth Modeling to Gain Insight into the Distribution of Achievement Within Schools
Michael Seltzer, Kilchan Choi and Yeow Meng Thum
EDUCATIONAL EVALUATION AND POLICY ANALYSIS 2003; 25; 263
DOI: 10.3102/01623737025003263

The online version of this article can be found at: http://epa.sagepub.com/cgi/content/abstract/25/3/263
Examining Relationships Between Where Students Start and how Rapidly they Progress: Using New Developments in Growth Modeling to Gain Insight into the Distribution of Achievement Within Schools

Michael Seltzer
Kilchan Choi
Yeow Meng Thum
University of California, Los Angeles

Studying change in student achievement is of central importance in numerous areas of educational research, including efforts to monitor school performance, investigations of the effects of educational interventions over time, and school effects studies focusing on how differences in school policies and practices relate to differences in student progress. In this article, we argue that in studying patterns of change, it is often important to consider the relationship between where students start (i.e., their initial status) and how rapidly they progress (i.e., their rates of change). Drawing on recent advances in growth modeling methodology, we illustrate the potential value of such an approach in the context of monitoring school performance. In particular, we highlight the ways in which attending to initial status in analyses of student progress can help draw attention to possible concerns regarding the distribution of achievement within schools. To convey the logic of our approach and illustrate various analysis possibilities, we fit a series of growth models to the time series data for students in several schools in the Longitudinal Study of American Youth (LSAY) sample. In a final section, we discuss some of the possibilities that arise in employing a modeling approach of this kind in evaluating educational programs and in conducting school effects research.

Keywords: educational indicators, equity, growth modeling, longitudinal analysis

Studying change, particularly change in student achievement, is of fundamental importance in educational research. It is therefore not surprising that the use of growth modeling techniques in education and related fields has increased tremendously during the last two decades. Growth modeling techniques provide a valuable framework for studying patterns of change (see Bryk & Raudenbush, 1987; Muthen & Khoo, 1998; Willett & Sayer, 1994). For example, in efforts to monitor school performance, growth modeling techniques are often used to assess how rapidly students in a given school or district are progressing on average in particular content areas, and to assess the extent to which rates of progress differ for various demographic groups (see Willms, 1992; Seltzer, Frank, & Bryk, 1994). In studies of the effectiveness of educational programs (e.g., remedial
reading programs), these techniques are used to compare rates of progress for students participating in different programs (see Muthen & Curran, 1997). In addition, many studies in the school effects literature employ growth-modeling techniques in efforts to examine how differences in various school policies, practices and compositional characteristics relate to differences in student achievement over time (see, for example, Bryk & Raudenbush, 1988).

Clearly much can be learned by moving beyond snapshots of student achievement at single points in time, to analyses and summaries of student growth. To be sure, the notion of growth in knowledge and skills lies at the heart of definitions of learning and education (Bryk & Raudenbush, 1988; Willett, 1988, p. 346).

However, rather than focus exclusively on changes in student achievement over time (e.g., estimates of rates of change), we feel that it is also often important to consider levels of student achievement at the start of the time span one is studying (e.g., estimates of initial status). Drawing on recent advances in growth modeling, we wish to highlight various ways in which attending to the relationship between where students start and how rapidly they progress can help broaden the kinds of questions that we are able to address in longitudinal investigations in education.

In this article we focus on the potential value of employing such an approach in efforts to monitor school performance. An important emphasis in monitoring school performance entails computing and examining school mean rates of change for cohorts of students across a series of grades, and comparing mean rates of change for various demographic groups (see, e.g., Sanders & Horn, 1994; Willms, 1992). One of our key aims is to explore how attending to the relationship between where students start (i.e., their initial status) and how rapidly they progress can help bring to light potentially important findings that might be masked if we limit our focus to examining mean rates of change.

First, in some schools differences among students in their initial status may be strongly related to their subsequent rates of change (e.g., those who start high may progress extremely rapidly, while those with relatively low initial status may display markedly slower rates of change). In other schools, however, initial status may be weakly related to subsequent rates of change. Thus, in addition to computing mean rates of change for schools, we consider the value of examining correlation coefficients and slopes that capture how differences in initial status relate to differences in rates of progress within schools.

Secondly, consider settings in which we wish to compare rates of change in achievement for various demographic groups. In situations where initial status and subsequent rates are strongly related, and where the groups we wish to compare differ in terms of their initial status, simple comparisons of rates of change can be misleading. In some respects, such situations are analogous to intervention or program evaluation settings where the groups of students we wish to compare (e.g., students participating in an innovative remedial reading program versus those participating in a more traditional program) differ to some extent in their pretest scores. As such, we illustrate the value of comparing rates of change for demographic groups of interest in a way that adjusts for, or takes into account, differences in initial status.

Third, we call attention to the fact that overall comparisons between demographic groups, whether adjusted or unadjusted for differences in initial status, can be misleading. In an illustrative example, we show that the size and direction of the difference in growth rates between girls and boys in a particular school may differ markedly, depending on whether we are considering girls and boys with relatively low initial status values, or girls and boys with relatively high initial status values. It becomes important, therefore, to explore interactions between initial status and various demographic characteristics on rates of change.

To help illustrate the analysis possibilities outlined above, we will fit a series of growth models to the time series data for students in several schools in the Longitudinal Study of American Youth (LSAY) sample (see Miller, Kimmel, Hoffer & Nelson, 1999). In the course of presenting these examples, we will discuss how such analyses can help draw attention to possible concerns regarding the distribution of growth in achievement within schools, and, it is hoped, help stimulate discussion among teachers and administrators at given school sites regarding these concerns.

As will be seen, the analysis possibilities that we highlight in this article are made possible by recent advances in growth modeling that enable one to employ initial status as a predictor of rates of change (see Muthen & Curran, 1997; Rauden-
bush & Bryk, 2002, chpt. 11; Raudenbush, Bryk, Cheong & Congdon, 2000, pp. 207–211; Seltzer, Choi & Thum, 2001; Thum, 2002). One of our goals is to make the logic of the growth modeling techniques that we employ as accessible as possible. This will entail defining and explicating certain key technical details and terms. Note that we have placed a number of other technical details in endnotes and in the Appendix for the interested reader.

While the main focus of this article is on the value of attending to relationships between initial status and rates of change in efforts to monitor school performance, we also, in a later section, discuss some of the possibilities that arise in employing such an approach in evaluating educational programs and in conducting school effects research.

All analyses presented in this article were conducted using the software package WinBUGS (Spiegelhalter, Thomas, Best, & Gilks, 2000), which is freely available via the Web. A variety of software options (e.g., HLM) are discussed in the final section of our article.

Moving Beyond Mean Rates of Change

Before turning to analyses of the LSAY data, we first wish to introduce certain key concepts. To help illustrate these concepts, suppose that we have measures of reading achievement across grades 2 through 5 for students in three different elementary schools. Suppose further that in the first school there is a positive relationship between where students start and how rapidly they progress. To help convey this pattern, Figure 1 displays the fitted reading achievement trajectories for four students. As can be seen, for the student with relatively low initial status, the slope of his trajectory is fairly flat (i.e., his rate of progress is very low). When we examine the set of four trajectories, we see that as initial status increases, the slopes of the trajectories increase (i.e., rates of change increase). Furthermore, we see that initial differences in student achievement (i.e., grade 2 levels of achievement) become magnified over time.

In school 2, initial status and rates of change are unrelated. Specifically, the four fitted trajectories displayed in Figure 2 reveal that students tend to progress at the same rate regardless of their initial status. Thus initial differences in achievement essentially hold steady over time.

In some schools initial status and rates of change may be unrelated, but the underlying pattern of change may differ from that in school 2. For example, growth may be slow or rapid for students with low initial status, and slow or rapid for students with high initial status. Thus as initial status
FIGURE 2. No relationship between initial reading achievement levels and rates of change.

increases, the relationship between initial status and rates of change is nonsystematic. This pattern will be illustrated in a later section of this article.

In school 3, the relationship between initial status and rates of change is negative. The four fitted trajectories in Figure 3 help convey that rates of change in this school are most rapid for those students with low initial status. As initial status increases, rates of progress decrease. Thus, in this particular school, initial differences in achievement tend to diminish over time.

An important implication is that while the average rates of change may be highly similar for two schools, the underlying relationship between initial status and rates of change within these schools may differ markedly. Thus, for example, in one school, initial differences in achievement may increase over time (Figure 1), while in the other, initial differences may decrease over time, i.e., those students who start low tend to catch up to those with relatively high initial levels of achievement (Figure 3).

An Illustrative Example Using Data from LSAY

To illustrate the above ideas, we now turn to analyses of the LSAY data. Note that the LSAY data set consists of over 50 cohorts of students in school districts throughout the U.S. Students in a given cohort attended the same middle school and then entered the same high school. In our article, we focus on math achievement scores collected at the start of grades 7, 8, 9 and 10 for students in several different cohorts. Note that users of LSAY typically refer to a cohort of students in the sample as being nested within a particular school (e.g., school 308). We will use this terminology as well. But bear in mind that in general a given cohort was first nested within a particular middle school and then, subsequently, within a particular high school.

As noted above, in this article we fit various growth models to the longitudinal data for students in several schools in the LSAY sample. As will be seen, each of the growth models that we employ consists of two models: A Level-1 or within-student model, and a Level-2 or between-student model. Within-student models enable us to capture key features of growth (e.g., initial status, rate of change) for each of the students in a sample. Between-student models enable us to estimate, for example, the mean rate of change for a group of students, assess the extent to which students vary in their rates of change, and identify key correlates of change.

We will first focus on estimating the mean rate of change and the correlation between initial sta-
tus and rates of change for the sample of students in school 308. Figure 4 displays the series of math achievement scores for each of these students. As can be seen, the observed trajectories are roughly linear. In addition, we see that the trajectories for some students are fairly flat, while the trajectories for a number of other students are quite steep, indicating rapid rates of progress.

Note that we are using a recent release of the LSAY data set, which contains achievement scale scores based on the application of Item Response Theory (IRT) models discussed in Bock and Zimowski (1997). In contrast to the use of such scales as percentile ranks and Grade Equivalent (GE) scores, IRT models, when appropriately applied, yield scales that are especially well suited for studying change.1

We now pose the following within-student (Level-1) model for the time-series data for each of $N=54$ students in school 308:

$$Y_{it} = \pi_{0i} + \pi_{1i}(GRADE_{it} - 7) + \varepsilon_{it},$$

where $Y_{it}$ represents the math achievement score for student $i$ at time $t$, and $GRADE_{it}$ represents the grade that student $i$ has entered at time $t$ (i.e., $GRADE$ takes on values ranging from 7–10).

The key parameters in this model are $\pi_{0i}$ and $\pi_{1i}$. $\pi_{1i}$ represents the growth rate (rate of change) for student $i$, and $\pi_{0i}$ is an intercept. Subtracting particular values from predictors in a model, which is termed centering in the parlance of Hierarchical Modeling, provides a way of giving intercepts meaningful interpretations (see, for example, Raudenbush & Bryk, 2001, chpt. 6). By virtue of subtracting a value of 7 from $GRADE$, $\pi_{0i}$ represents the expected math achievement score for student $i$ at the start of grade 7 (i.e., initial status).2 Finally, the $\varepsilon_{it}$ are residuals assumed normally distributed with mean 0 and variance $\sigma^2$.

What is the mean rate of change for students in school 308? Do rates of change for students in this school tend to vary systematically with initial status? To help address such questions, we specify the following between-student (Level-2) model:

$$\pi_{0i} = \beta_{00} + U_{0i},$$
$$\pi_{1i} = \beta_{10} + U_{1i},$$

where $\beta_{00}$ and $\beta_{10}$ represent, respectively, the means for initial status and rates of change for students in school 308. The $U_{0i}$ and $U_{1i}$ are Level-2 residuals commonly termed random effects. $U_{0i}$

FIGURE 3. Negative relationship between initial reading achievement levels and rates of change.
captures the deviation of initial status for student \(i\) from the mean for initial status, and \(U_{1i}\) represents the deviation of the growth rate for student \(i\) from the mean growth rate. Note that the \(U_{0i}\) and \(U_{1i}\) are assumed normally distributed with variance \(\tau_{00}\) and \(\tau_{11}\), respectively. These variances capture the extent to which students vary in their initial status and their rates of change. Thus, for example, if the rates of change for students in school 308 tend to be tightly clustered around the mean rate (i.e., if student rates of change are rather homogeneous), this will be reflected in small estimates for \(\tau_{11}\). In contrast, if some students progress at substantially faster rates than the mean rate, and some progress at substantially slower rates (i.e., if rates of change are very heterogeneous), this will be reflected in large estimates for \(\tau_{11}\). Similarly, the extent to which students vary in their initial status values will be reflected in the resulting estimate for \(\tau_{00}\).

Furthermore, as is commonly done in applications of growth models, we allow for the fact that the initial status values and growth rates for students may covary. This is captured by a covariance parameter (i.e., \(\tau_{01}\)). Thus, for example, a large positive value for \(\tau_{01}\) would imply that relatively slow rates of change accompany low initial status values, and relatively fast rates accompany high initial status values. A large negative value for \(\tau_{01}\) would indicate that as initial status increases, rates of change decrease.

Note also that the inclusion of a covariance term in the between-student model provides a basis for estimating the correlation between initial status and rates of change. Such correlations, analogous to Pearson correlation coefficients, range between values of −1 and 1.

We now fit the growth model defined by Equations 1 and 2 to the data. With respect to the mean initial status for students in school 308, we obtain an estimate of approximately 45 points (see Table 1). We also see that the resulting estimate for the mean rate of change is 3.52, which implies that student math achievement scores are, on average, increasing approximately 3.5 points per grade.

Next, note that the estimated covariance between initial status and rates of change is positive (i.e., 9.43), and that the 95% interval for this estimate contains only positive values. This provides strong evidence of a positive relationship between initial status and rates of change.

Turning to the correlation between initial status and rates of change, which we denote using the symbol ‘\(\rho\)’, we obtain an estimate of .50. In addition, we see that the 95% interval for \(\rho\) contains only positive values.

A positive relationship between where students start and how rapidly they progress can be discerned in Figure 4. Since time series achievement data is often fairly noisy, it is also useful to plot estimates of growth trends (fitted trajectories) for individuals based on the growth model one is employing. Figure 5 displays the fitted growth trajectories for the students in school 308. Based on Figures 4 and 5, we can see that while the relationship between status at the start of grade 7 and rates of change is positive, the relationship is

### TABLE 1
Comparing Patterns of Change for Students in Schools 308 and 143

<table>
<thead>
<tr>
<th></th>
<th>School 308 (N=54)</th>
<th></th>
<th>School 143 (N=54)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate</td>
<td>95% Interval</td>
<td>Estimate</td>
<td>95% Interval</td>
</tr>
<tr>
<td><strong>Fixed Effect</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean initial status</td>
<td>(\beta_{00})</td>
<td>44.72 (42.28, 47.14)</td>
<td>44.60 (42.26, 46.92)</td>
<td></td>
</tr>
<tr>
<td>Mean rate of change</td>
<td>(\beta_{10})</td>
<td>3.52 (2.69, 4.35)</td>
<td>3.33 (2.27, 4.40)</td>
<td></td>
</tr>
<tr>
<td><strong>Variance Component</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within-person error</td>
<td>(\sigma^2)</td>
<td>13.07 (9.99, 17.38)</td>
<td>17.24 (12.58, 24.16)</td>
<td></td>
</tr>
<tr>
<td>Random effects variance for initial status</td>
<td>(\tau_{00})</td>
<td>70.04 (46.57, 108.30)</td>
<td>60.78 (38.87, 96.78)</td>
<td></td>
</tr>
<tr>
<td>Random effects variance for rates of change</td>
<td>(\tau_{11})</td>
<td>5.36 (3.02, 9.33)</td>
<td>7.23 (3.54, 14.26)</td>
<td></td>
</tr>
<tr>
<td>Random effects covariance</td>
<td>(\tau_{01})</td>
<td>9.43 (2.96, 17.68)</td>
<td>0.70 (−9.09, 9.07)</td>
<td></td>
</tr>
<tr>
<td>Correlation between initial status and rates of change</td>
<td>(\tau_{01} / \sqrt{\tau_{00} \times \tau_{11}})</td>
<td>0.50 (0.15, 0.74)</td>
<td>0.03 (−0.35, 0.44)</td>
<td></td>
</tr>
</tbody>
</table>
FIGURE 4. Observed math achievement trajectories for students in school 308. Each observed trajectory is comprised of a student’s series of math achievement scores. The dashed lines represent students whose grade 7 achievement scores lie below the mean initial status estimate for school 308 (44.72 points). The solid lines represent students whose grade 7 achievement scores lie above a value of 44.72.

FIGURE 5. Fitted growth trajectories for school 308. Similar to Figure 4, the dashed lines represent students whose grade 7 achievement scores lie below the mean initial status estimate for school 308 (44.72 points), and the solid lines represent students whose grade 7 achievement scores lie above a value of 44.72.
far from perfect. For example, some students with relatively high grade 7 achievement scores progress fairly slowly over time. In addition, several students with low grade 7 achievement scores progress quite rapidly. However, in general, we see that among students with relatively low grade 7 scores, many have trajectories that are fairly flat, while among students who start out relatively high, many have trajectories that are substantially steeper. In connection with this, we see a fanning out of achievement scores over time.

We now fit the model defined by Equations 1 and 2 to the data for students in school 143. In Table 1, we see that the estimate of the mean growth rate for students in this school is quite similar to the estimated mean rate for students in school 308. In addition, the mean initial status estimates for these schools are nearly identical.

An important difference in results is that while the estimate of the covariance between initial status and rates of change for school 143 is positive, it is substantially smaller than the covariance estimate for school 308, and the corresponding 95% interval comfortably includes a value of 0. Similarly, the estimated correlation between initial status and rates of change is extremely small (0.03), and the corresponding 95% interval contains a value of 0.

These results are consistent with inspections of the time series data (Figure 6a) and fitted trajectories (Figure 6b) for the sample of students in school 143. As can be seen, there is extensive crisscrossing of trajectories. When we consider students with relatively low grade 7 scores and students with relatively high grade 7 scores, we see that for both groups, rates of change are rapid for some students but slow for others. Student progress is, in some sense, more “fluid” in school 143 than in school 308. By this, we mean that over time, an appreciable number of students with lower than average grade 7 scores catch up to or surpass students with higher than average grade 7 scores.

Thus, as can be clearly seen, while the mean rates of change for schools 308 and 143 are extremely similar, the underlying distributions of growth in achievement are markedly different. In school 308, we see strong positive relationship between where students start and their subsequent rates of change, whereas in school 308, the relationship between initial status and rates of change is decidedly weak.

### Regressing Rates of Change on Initial Status

Correlation coefficients provide us with measures of the strength of linear association between two variables (e.g., variables $A$ and $B$). In addition to information of this kind, we are also often interested in information concerning the expected amount of change in one variable (e.g., variable $A$) when the other increases 1 unit.

Thus, we might ask: For students in school 308, how much of a change in rate of growth ($\pi_{0i}$) do we expect when initial status ($\pi_{00}$) increases 1 unit? Addressing this question implies employing initial status as a predictor of rate of change. Thus we expand our between-student model as follows:

$$
\pi_{0i} = \beta_{00} + U_{0i} \tag{3a}
$$

$$
\pi_{1i} = \beta_{10} + b(\pi_{0i}) + U_{1i} \tag{3b}
$$

A key parameter in this model is $b$, which is a coefficient that captures the amount of change that we expect in $\pi_{1i}$ when $\pi_{0i}$ increases 1 unit. Note that regressing one parameter ($\pi_{1i}$) on another ($\pi_{0i}$) is termed a latent variable regression, and that coefficients such as $b$ are termed latent variable regression coefficients (see, for example, Muthen & Curran, 1997; Raudenbush & Bryk, 2002; Seltzer, Choi, & Thum, 2001). As in the previous model, the $U_{1i}$ are assumed normally distributed with mean 0 and variance $\tau_{11}$. However, in this model $\tau_{11}$ now represents the variance that remains in growth rates after taking into account differences in initial status. Note that the above model for initial status (Equation 3a) is identical to the model for initial status in the previous between-student model (Equation 2a).6,7

When we fit the growth model defined by Equations 1 and 3 to the data for school 308, we obtain an estimate of 0.174 for $b$ (see Table 2). As can be seen, the corresponding 95% interval contains only positive values. This estimate implies that when initial status increases by, say, 10 points, we expect an increase in rate of change in mathematics achievement of $10 \times 0.174 = 1.74$ points per grade.

Based on fitted models of this kind, insight into the distribution of growth in achievement in a school can be obtained by computing the expected growth rates for a range of values for initial status. The procedure for computing expected growth rates is analogous to computing fitted values based on various predictor values.
FIGURES 6a, and b. Figure 6a displays the time series data (i.e., the observed trajectories) for students in school 143; Figure 6b displays the fitted growth trajectories. In both figures, the dashed lines represent students whose grade 7 achievement scores lie below the mean initial status estimate for school 143 (44.60 points). The solid lines represent students whose grade 7 achievement scores lie above a value of 44.60.
TABLE 2
Comparing Initial Status/Rate of Change Slopes for Schools 308 and 143

<table>
<thead>
<tr>
<th></th>
<th>School 308 (N=54)</th>
<th>School 143 (N=54)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate</td>
<td>95% Interval</td>
</tr>
<tr>
<td>Fixed effect</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean initial status ($\beta_0$)</td>
<td>44.73 (42.34, 47.15)</td>
<td>44.62 (42.28, 46.94)</td>
</tr>
<tr>
<td>Mean rate of change ($\beta_{10}$)</td>
<td>3.49 (2.66, 4.32)</td>
<td>3.32 (2.23, 4.44)</td>
</tr>
<tr>
<td>Initial status/rate of change slope ($b$)</td>
<td>0.174 (0.067, 0.290)</td>
<td>0.010 (−0.147, 0.179)</td>
</tr>
<tr>
<td>Variance component</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within-person error ($\sigma^2$)</td>
<td>13.44 (10.23, 17.87)</td>
<td>17.07 (12.42, 24.04)</td>
</tr>
<tr>
<td>Random effects variance for initial status ($\tau_{00}$)</td>
<td>67.27 (44.61, 104.40)</td>
<td>60.54 (38.95, 95.91)</td>
</tr>
<tr>
<td>Random effects variance for rates of change ($\tau_{11}$)</td>
<td>3.06 (1.22, 6.65)</td>
<td>7.69 (3.77, 14.90)</td>
</tr>
</tbody>
</table>

of interest in regression settings. Details can be found in the appendix.

For students in school 308 with initial status values equal to the mean initial status value for that school (i.e., 44.73), the expected rate of change is 3.49 points per grade. However, for students with initial status values 10 points above the mean initial status value (i.e., 55.73), the expected rate of change is appreciably faster: 5.23. Conversely, for students with initial status values 10 points below the mean initial status value (i.e., 34.73), the expected rate of change is substantially slower: 1.75. These initial status values and the corresponding expected rates define the achievement trajectories displayed in Figure 7.

In contrast, for school 143, we obtain an estimate for $b$ of 0.010, and it can be seen that the resulting 95% interval comfortably includes a value of 0 (see Table 2). The point estimate of 0.010 implies that when initial status increases by an appreciable amount (e.g., 10 points or 20 points), we expect hardly any increase at all in rate of

![FIGURE 7. Expected growth trajectories for students in school 308 based on the results in Table 2. The mean initial status value for this school is 44.73.](http://eepa.aera.net)
where we pose the following level-2 model to compare our demographic groups of interest, model as in the above sections (see Equation 1). For illustrative purposes, we focus on changes in more detail in the final section of our article.

Comparing Mean Rates of Change for Different Demographic Groups

We now turn to the issue of comparing rates of progress for different demographic groups of students. For illustrative purposes, we focus on differences in rates of change between girls and boys in school 142 in the LSAY sample. For these analyses, we will work with the same level-1 model as in the above sections (see Equation 1). To compare our demographic groups of interest, we pose the following level-2 model

\[
\pi_{ii} = \beta_{i0} + \beta_{i1}GENDER_i + U_{i1},
\]

(4b)

where \(GENDER_i\) takes on a value of 0 if student \(i\) is a boy and a value of 1 if student \(i\) is a girl. By virtue of this coding scheme, \(\beta_{i0}\) represents the expected initial status for boys in school 142, and \(\beta_{i1}\) captures the expected difference in initial status between girls and boys. Similarly, \(\beta_{i0}\) represents the expected rate of change in math achievement for boys, and \(\beta_{i1}\) captures the expected difference in rates of change between girls and boys. Our level-2 model also contains variance terms (i.e., \(\tau_{00}\) and \(\tau_{11}\)) and a covariance term (i.e., \(\tau_{01}\)). In the context of the model, analogous to comparing groups of interest using ANOVA, \(\tau_{00}\) and \(\tau_{11}\) capture, respectively, the amount of within-group variation in initial status and growth rates. In addition, \(\tau_{01}\) captures the extent to which initial status and growth rates covary within groups.

The resulting estimate for \(\beta_{01}\) indicates that initial status is, on average, approximately 2.8 points higher for the girls in our sample, though the lower boundary of the 95% interval for \(\beta_{01}\) includes a value of 0 (see Table 3). In contrast, the estimate for \(\beta_{11}\) (-1.63) suggests that rates of change for girls are, on average, appreciably lower than rates of change for boys. As can be seen, the 95% interval for \(\beta_{11}\) excludes a value of 0, though just barely so. Thus, while the expected rate of change for boys is 5.22 points per year, the expected rate for girls is \(5.22 - 1.63 = 3.59\) points per year.

It is important to note that our comparison of growth rates for girls and boys does not take into account the fact that math achievement at the start of grade 7 is somewhat higher for the girls in our sample. If differences in initial status were inconsequential in terms of how rapidly students progress, this would not be a concern. However, the results that we obtain for \(\tau_{01}\) and the correlation coefficient (\(\rho\)) point to a fairly strong positive relationship between initial status and rates of change for boys and for girls. Note, in particular, that the resulting estimate for \(\rho\) is 0.52.

Analogous to using ANCOVA models to compare groups of interest adjusting for initial pretest differences, we now attempt to obtain an estimate of the difference in growth rates between girls and boys adjusting for differences in initial status. To accomplish this, we pose the following between-student model:

\[
\pi_{0i} = \beta_{00} + \beta_{01}GENDER_i + U_{0i},
\]

(5a)

\[
\pi_{1i} = \beta_{10} + \beta_{11}GENDER_i + U_{1i} + b(\pi_{0i}) + U_{1i}.
\]

(5b)

In contrast to Equation 4b, \(\pi_{0i}\) appears as a covariate in our model for growth rates. The parameter of primary interest in this model is \(\beta_{11}\), which represents an expected difference in growth.
rates between girls and boys that is adjusted for differences in initial status. Put differently, $\beta_{11}$ represents an expected difference in growth rates holding constant initial status. The parameter $b$ is a regression coefficient that relates differences in initial status to rates of change. Note that this model essentially assumes that the slope relating initial status to rates of change is equivalent for girls and boys. As will be seen in a later section of our article, this assumption appears to be extremely reasonable in the case of this school.

Before proceeding, we wish to point out that in comparing groups it is well-known that under-adjustments for pre-existing differences can arise when covariates are measured with appreciable error (see, for example, Reichardt, 1979). Employing initial status ($\pi_0$) as a covariate can help avoid such problems (see also Muthen & Curran, 1997; Khoo, 2001; Raudenbush, Bryk, Cheong, & Congdon, 2000, p. 208; Raudenbush & Bryk, 2002, chpt. 11).

In Table 4, we see that the resulting estimate for $b$ is 0.153 and that the corresponding 95% interval contains only positive values. This suggests a strong, positive relationship between initial status and subsequent rates of progress for boys and for girls. When we take into account the fact that boys, on average, have lower initial status values than girls, we see that the resulting expected difference in growth rates between girls

### TABLE 3
*Comparing Rates of Change for Girls and Boys in School 142, (24 Girls, 36 Boys)*

<table>
<thead>
<tr>
<th>Fixed effect</th>
<th>Estimate</th>
<th>95% Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model for initial status ($\pi_0$):</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Boys ($\beta_{00}$)</td>
<td>50.78</td>
<td>(47.33, 54.21)</td>
</tr>
<tr>
<td>Girls/Boys contrast ($\beta_{01}$)</td>
<td>2.82</td>
<td>(−2.71, 8.27)</td>
</tr>
<tr>
<td>Model for rates of change ($\pi_1$):</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Boys ($\beta_{10}$)</td>
<td>5.22</td>
<td>(4.18, 6.27)</td>
</tr>
<tr>
<td>Girls/Boys contrast ($\beta_{11}$)</td>
<td>−1.63</td>
<td>(−3.27, −0.01)</td>
</tr>
<tr>
<td>Variance component</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within-person error ($\sigma^2$)</td>
<td>18.69</td>
<td>(14.55, 24.27)</td>
</tr>
<tr>
<td>Random effects variance for initial status ($\tau_{00}$)</td>
<td>94.58</td>
<td>(64.00, 144.10)</td>
</tr>
<tr>
<td>Random effects variance for rates of change ($\tau_{11}$)</td>
<td>5.10</td>
<td>(2.72, 9.18)</td>
</tr>
<tr>
<td>Covariance between initial status and rates of change ($\tau_{01}$)</td>
<td>11.06</td>
<td>(3.86, 20.22)</td>
</tr>
<tr>
<td>Correlation between initial status and rates of change</td>
<td>0.52</td>
<td>(0.18, 0.76)</td>
</tr>
</tbody>
</table>

### TABLE 4
*Comparing Rates of Change for Girls and Boys in School 142 Adjusting for Differences in Initial Status, (24 Girls, 36 Boys)*

<table>
<thead>
<tr>
<th>Fixed effect</th>
<th>Estimate</th>
<th>95% Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model for initial status ($\pi_0$):</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean initial status ($\beta_{00}$)</td>
<td>51.91</td>
<td>(49.28, 54.54)</td>
</tr>
<tr>
<td>Girls/Boys contrast ($\beta_{01}$)</td>
<td>2.80</td>
<td>(−2.57, 8.12)</td>
</tr>
<tr>
<td>Model for rates of change ($\pi_1$):</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Boys ($\beta_{10}$)</td>
<td>5.37</td>
<td>(4.37, 6.41)</td>
</tr>
<tr>
<td>Girls/Boys contrast ($\beta_{11}$)</td>
<td>−2.06</td>
<td>(−3.67, −0.45)</td>
</tr>
<tr>
<td>Initial status/rate of change slope ($b$)</td>
<td>0.153</td>
<td>(0.066, 0.249)</td>
</tr>
<tr>
<td>Variance component</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within-person error ($\sigma^2$)</td>
<td>19.39</td>
<td>(15.07, 25.15)</td>
</tr>
<tr>
<td>Random effects variance for initial status ($\tau_{00}$)</td>
<td>90.27</td>
<td>(60.84, 137.80)</td>
</tr>
<tr>
<td>Random effects variance for rates of change ($\tau_{11}$)</td>
<td>2.37</td>
<td>(0.83, 6.07)</td>
</tr>
</tbody>
</table>
and boys is appreciably larger than the unadjusted difference (i.e., −2.06 versus −1.63), and that the upper boundary of the resulting 95% interval lies well below a value of 0. This result implies that for girls and boys with similar levels of achievement at the start of grade 7 (i.e., holding constant initial status), the expected rate of growth in achievement for boys is approximately 2 points per year faster than the expected rate for girls. This difference in rates translates into a difference in expected achievement scores after three years of schooling (i.e., at the start of grade 10) of approximately 6 points.

Thus, as in the case of the analyses in the previous section, our attention is again drawn to issues concerning the distribution of achievement within schools. What processes or factors likely underlie this pattern of results for girls and boys? What might be done to try to promote more rapid rates of progress among girls in this school? These are some of the questions that the analyses in this section encourage us to consider. Before moving on to the next section, we wish to point out that the samples of girls and boys in school 142 are extremely similar in terms of such potentially important intake characteristics as home resources and educational aspirations. Thus the differences in rates of change that we see cannot be accounted for by these factors.

Overall Comparisons Can Be Misleading: Examining Interactions Between Initial Status and Demographic Characteristics

While comparing rates of change for various demographic groups of interest can be extremely useful, such comparisons can potentially be misleading, even when we have taken into account differences in initial status. To help illustrate this point, we will focus on patterns of growth for students in school 302 in the LSAY sample.

Initial status values and rates of change differ very little for girls and boys in this school. As can be seen in Table 5, initial status values for girls are, on average, approximately 0.60 points lower for girls than boys, and rates of change, holding constant initial status, are slightly faster for girls (0.20). (Though not displayed in a table, we wish to point out that when we conducted an analysis in which we did not adjust for differences in initial status, the resulting expected difference in rates slightly favored girls as well (i.e., 0.15).)

In the case of the analysis presented in Table 5, a key assumption is that the slope relating initial status to rates of change (b) is equivalent for girls and boys. This is analogous to the assumption of parallel within-group slopes in classic ANCOVA analyses. An implication of this assumption is that the expected difference in rates of change between girls and boys is 0.20 regardless of whether we are considering boys and girls with relatively low initial achievement values, or whether we are considering boys and girls with high initial values. As will become clear, this assumption is extremely reasonable in the case of school 142, but highly questionable in the case of school 302.

We now pose a model that allows for the possibility that the slope relating initial status to rates of change may differ for girls and boys. That is, we include an interaction term (i.e., GENDERi × π0i) in our model for rates of change:

<table>
<thead>
<tr>
<th>TABLE 5</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Comparing Rates of Change for Girls and Boys in School 302 Adjusting for Differences in Initial Status, (38 Girls, 41 Boys)</strong></td>
</tr>
<tr>
<td>Fixed effect</td>
</tr>
<tr>
<td>Model for initial status (π0i):</td>
</tr>
<tr>
<td>Mean initial status (β00)</td>
</tr>
<tr>
<td>Girls/Boys contrast (β01)</td>
</tr>
<tr>
<td>Model for rates of change (π1i):</td>
</tr>
<tr>
<td>Boys (β10)</td>
</tr>
<tr>
<td>Girls/Boys contrast (β11)</td>
</tr>
<tr>
<td>Initial status/rate of change slope (b)</td>
</tr>
<tr>
<td>Variance component</td>
</tr>
<tr>
<td>Within-person error (σ²)</td>
</tr>
<tr>
<td>Random effects variance for initial status (τ00)</td>
</tr>
<tr>
<td>Random effects variance for rates of change (τ11)</td>
</tr>
</tbody>
</table>
It is instructive to focus on the portions of Equation 6b that involve \( \pi_{0i} \). As can be seen, \( \pi_{0i} \) appears as a predictor by itself and in the interaction term \( GENDER_i \times \pi_{0i} \); the corresponding coefficients are \( b_1 \) and \( b_2 \). Thus we have:

\[
b_1(\pi_{0i}) + b_2(GENDER_i \times \pi_{0i}).
\]

(7)

Note that for boys (i.e., when \( GENDER_i = 0 \)), the interaction term takes on a value of \((0 \times \pi_{0i}) = 0\), and so Equation 7 reduces to:

\[
b_1(\pi_{0i}).
\]

Thus, we see that \( b_1 \) is the slope capturing the relationship between initial status and rates of change for boys.

For girls (i.e., \( GENDER_i = 1 \)), the interaction term takes on a value of \((1 \times \pi_{0i}) = \pi_{0i}\), and so Equation 7 reduces to:

\[
b_1(\pi_{0i}) + b_2(\pi_{0i}) = (b_1 + b_2)(\pi_{0i}).
\]

As can be seen, the relationship between initial status and rates of change for girls is captured by \( b_1 \) (i.e., the slope for boys) plus \( b_2 \). Thus \( b_2 \) captures the difference between the initial status/rate of change slopes for girls and boys.

In Table 6, we see that the resulting estimate of \( b_1 \) is 0.142, and that the lower boundary of the corresponding 95\% interval is well above a value of 0. This suggests a strong positive relationship between initial status and rates of change for boys. That is, for boys, differences in initial status appear to be very consequential with respect to subsequent rates of change. In contrast, we obtain a negative estimate for \( b_2 \) (i.e., \(-0.121\)). It can also be seen that the corresponding 95\% interval for \( b_2 \) contains only negative values. This suggests that the initial status/rate of change slopes for girls and boys differ substantially. In particular, summing the point estimates for \( b_1 \) and \( b_2 \) we obtain a value of 0.021. Thus for girls, differences in initial status appear to be rather inconsequential with respect to subsequent rates of change. (Note that in fitting the above interaction model to the data for students in school 142, we obtain a point estimate for \( b_2 \) that is extremely close to 0. This suggests that in school 142, initial status/rate of change slopes appear to be similar for girls and boys.)

Analogous to ANCOVA analyses, such differences in initial status/rate of change slopes have important implications for drawing conclusions concerning expected differences in rates of change between girls and boys. Based on our fitted model, it is instructive to compute expected rates of change for girls and boys across a range of

<table>
<thead>
<tr>
<th>TABLE 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Examining the Interaction Between Initial Status and Gender on Rates of Change in School 302, (38 Girls, 41 Boys)^9</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Fixed effect</th>
<th>Estimate</th>
<th>95% Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model for initial status (( \pi_{0i} )):</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean initial status (( \beta_{00} ))</td>
<td>52.54</td>
<td>(50.41, 54.66)</td>
</tr>
<tr>
<td>Girls/Boys contrast (( \beta_{01} ))</td>
<td>-0.61</td>
<td>(-4.89, 3.70)</td>
</tr>
<tr>
<td>Model for rates of change (( \pi_{1i} )):</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Boys (( \beta_{10} ))</td>
<td>4.26</td>
<td>(3.59, 4.93)</td>
</tr>
<tr>
<td>Girls/Boys contrast (( \beta_{11} ))</td>
<td>0.20</td>
<td>(-0.76, 1.15)</td>
</tr>
<tr>
<td>Initial status/rate of change slope (( b_1 ))</td>
<td>0.14</td>
<td>(0.080, 0.214)</td>
</tr>
<tr>
<td>Interaction between gender and initial status (( b_2 ))</td>
<td>-0.12</td>
<td>(-0.226, -0.018)</td>
</tr>
<tr>
<td>Variance component</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within-person error (( \sigma^2 ))</td>
<td>13.12</td>
<td>(10.86, 15.96)</td>
</tr>
<tr>
<td>Random effects variance for initial status (( \tau_{00} ))</td>
<td>82.38</td>
<td>(59.15, 117.10)</td>
</tr>
<tr>
<td>Random effects variance for rates of change</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Boys (( \tau_{11B} ))</td>
<td>0.37</td>
<td>(0.11, 1.72)</td>
</tr>
<tr>
<td>Girls (( \tau_{11G} ))</td>
<td>1.23</td>
<td>(0.49, 3.02)</td>
</tr>
</tbody>
</table>

9
values for initial status. As can be seen in Table 6, the estimate of the mean initial status value for school 302 is 52.54 points. Table 7 displays the expected rates of change for girls and boys corresponding to initial status values that are equal to, 12 points above, and 12 points below 52.54. (The procedure for computing these expected rates is outlined in the appendix.) As can be seen the expected rates for girls and boys with initial status values equal to the school mean are extremely similar (i.e., 4.46 points per grade for girls versus a value of 4.26 for boys). We also see that for initial status values 12 points above or below the school mean, the expected rates for girls differ from a value of 4.46 by only a small amount. This reflects the weak relationship between initial status and rates of change for girls in this school noted above.

For boys, the picture is vastly different. In contrast to an expected rate of 4.26 points per grade for boys with initial status values equal to the school mean, we see that for boys with initial status values 12 points below the school mean, the expected rate is substantially slower (i.e., 2.56). However, for boys with initial status values 12 points above the school mean, the expected growth rate is appreciably faster (i.e., 5.96). The substantial differences in expected rates that we see for boys across the range of initial status values reflects the strong relationship between initial status and rates of change for boys in school 302.

The trajectories defined by the sets of initial status values and expected rates presented in Table 7 are displayed in Figure 8. The plotted trajectories help make vivid the fact that our conclusions about the size and direction of differences in growth rates between girls and boys in school 302 depend very much on the particular initial status value we are considering. For girls and boys whose initial status values are equal to the school mean, the difference in expected rates is minute. However, for girls and boys with relatively low initial status, the expected rate for girls is substantially faster. This pattern is then reversed when we consider girls and boys with relatively high initial status.

An implication of the above analyses is that estimates of overall differences in rates of change for different demographic groups may be misleading. In particular, they may mask the fact that the size and direction of the difference may vary markedly depending upon whether we are focusing on students with relatively low or high initial status values.

**Discussion**

Attending to mean rates of change, and overall differences in rates of change for various demographic groups is of central importance in monitoring the performance of schools and districts. In this article, we have argued for the need to expand this focus by also considering the relationship between initial status and rates of change. In particular, we have focused on various summaries and analyses that can help draw our attention to possible concerns regarding the distribution of achievement within a school (e.g., why does initial status appear to be very consequential with respect to subsequent rates of progress?; among students with high initial status values, why are rates of progress appreciably more rapid for boys than girls?).

**TABLE 7**

*Expected Growth rates for Girls and Boys in School 302 Based on the Initial Status × Gender Interaction Analysis*

<table>
<thead>
<tr>
<th>Student initial status values</th>
<th>Expected Growth Rates</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Girls</td>
</tr>
<tr>
<td>12 points above the school mean:</td>
<td>64.54</td>
</tr>
<tr>
<td>Equal to the school mean:</td>
<td>52.54</td>
</tr>
<tr>
<td>12 points below the school mean:</td>
<td>40.54</td>
</tr>
</tbody>
</table>

*Note.* The mean initial status estimate for school 302 is 52.54 points. The expected rates that we have computed correspond to initial status values that are (a) 12 points above the school mean; (b) equal to the school mean; and (c) 12 points below the school mean.
We saw that estimates of correlations between initial status and rates of change, and estimates of initial status/rate of change slopes, provide us with two key measures of within-school relationships between initial status and rates of change. We also focused on the potential value of comparing growth rates for various demographic groups of interest adjusting for differences in initial status. Note that we feel it is important to consider both unadjusted and adjusted differences in rates. Each of these measures provides us with useful information. Regardless of whether one group of students differs from another in terms of initial status, it is important to know whether these groups tend to be progressing at fairly similar rates or not. Unadjusted differences provide us with information of this kind. However, if initial status in a school of interest appears to be very consequential with respect to subsequent rates of achievement, and if the groups that we wish to compare differ appreciably in terms of initial status, then it is also valuable to ask: What is the expected difference in rates of change holding constant initial status?

An extremely important point, however, is that comparisons of growth rates can be misleading regardless of whether we adjust for differences in initial status or not. Specifically, as we saw above, the size and direction of expected differences in rates of change between demographic groups of interest can vary substantially across various initial levels of achievement. In the case of our example, we saw that among students with relatively low initial status, rates of change were appreciably faster for girls than boys. However, among students with relatively high initial status, rates of progress tended to be more rapid for boys than girls.

Our hope is that the kinds of summaries, analyses and plots (i.e., plots of time series data; plots of fitted trajectories) presented above can, in conjunction with various measures of school performance (see, e.g., Bryk, Thum, Easton, & Luppescu, 1998; Sanders & Horn, 1994; Thum, in press; Willms, 1992), help stimulate fruitful discussion among school personnel regarding possible areas of concern. For example, suppose...
that in a particular school there is evidence of a strong positive relationship between initial status and rates of change across a series of grades of substantive interest. Drawing on their extensive experience and contextual knowledge, it would be extremely valuable for the teachers and administrators in this school to discuss the factors likely underlying this relationship. For example, to what extent might this be due to the school’s policies regarding tracking? To what extent might this be due to policies regarding the number and kinds of mathematics courses that students are required to take? In these discussions, it would be important to consider those factors that are likely contributing to differences in achievement among students observed at the start of the series of grades under consideration (e.g., the nature and extent of previous academic difficulties; differences in the quality of prior instruction received by students). The particular factors underlying the strong relationship between initial status and rates of change in this school may be quite clear, or it may be the case that further study and discussion are necessary.

Note that with the help of research personnel, various hunches could be assessed more formally through additional growth modeling analyses if the necessary data are available. To continue with the above example, suppose we wish to explore whether the strong positive relationship between initial status and rates of change for students in this school (i.e., a large, positive estimate for $b$) is in large part due to differences in the quality of pre-algebra instruction received by students. If this is so, fitting a model in which rates of change are modeled as a function of initial status and measures of this explanatory factor would result in a substantially smaller estimate for $b$.

**The fundamental importance of studying factors related to change**

In inspecting plots of growth trajectories for the students in a school or district, we will almost always see some crisscrossing of trajectories. It is important to explore some of the implications of this.

First, consider again the pattern of change in school 143, a school in which initial status and rates of change are uncorrelated. As noted earlier, plots of the time series data and fitted trajectories for the students in this school reveal an extensive amount of crisscrossing of trajectories (see Figures 6a and b). The heart of the matter is that over time, many students with relatively fast growth rates surpass many whose rates are slower. In particular, in Figure 6b it can be seen that an appreciable number of students with low initial levels of achievement progress quite rapidly, and attain relatively high levels of achievement by the start of grade 10; in fact, the grade-10 levels of achievement for these students rank well above many students with higher initial levels of achievement, but who made very little progress over time.

Thus, although initial status and rates of change are uncorrelated in school 143, students in this school in fact differ substantially in their rates of change, and this, in turn, has important implications for the levels of achievement attained by students at later points in time. Specifically, the differences in attainment that we see among students at the start of grade 10 are related to some extent to the differences that we see in students’ rates of progress. As can be seen in Figure 6b, those students with slow rates tend to have relatively low or moderate levels of achievement at the start of grade 10 (though of course there are some exceptions), and those with fast rates tend to have relatively high grade-10 levels of achievement (see Endnote 10). Thus, it would be important for the teachers and administrators at such a site to try to identify those factors that underlie the appreciable differences that we see in students’ rates of change: Why do some students with low initial levels of achievement progress very rapidly, while others with low initial levels hardly progress at all? Why do some students with relatively high initial levels of achievement make very little progress?

It is also instructive to consider again the pattern of change in school 308. While the correlation between status at the start of grade 7 and rate of change is positive, the relationship is far from perfect. For example, in Figure 5 it can be seen that several students with relatively low scores in grade 7 progress quite rapidly. These students, over time, surpass and attain higher levels of achievement than many students who start higher but whose rates of change are slower. An implication of this is that in trying to understand the factors underlying a strong positive relationship between initial status and rates of change for the students in a school, it is crucial to consider the exceptions to this pattern: Why is it that certain students with relatively low initial levels of achievement progress so rapidly? Why is it that
some students with relatively high initial levels of achievement appear to make little progress over time?

The previous examples serve as a reminder of the fundamental importance of trying to understand the factors underlying differences in student progress. While we have highlighted several key ways in which attending to where students start can help us gain insight into the distribution of growth in student achievement in schools, we do not want this to overshadow the importance of attending to the wide array of factors (e.g., school experiences, policies and practices; home environmental factors) potentially underlying important differences in student progress—differences that have implications for where students wind up at later points in time.

Implications for longitudinal studies of programs and interventions

In longitudinal studies of the effectiveness of educational programs (e.g., remedial reading interventions, school-based prevention programs, innovative curricula in various subject areas), interest often centers on comparing rates of change in achievement, or in certain behaviors or attitudes, for students participating in one program with those assigned to another program (or to control conditions) (see, e.g., Muthen & Curran, 1997). However, when random assignment is not possible, or in situations where random assignment is possible but sample sizes are small, it is likely that the groups we wish to compare will differ to some extent in their baseline measures of the skills or behaviors of primary interest. If where students start is strongly related to how rapidly they progress, then analogous to an ANCOVA analysis, it would be important to compare the rates of change for students in different programs in a way that takes into account differences in initial status. In such situations, using an approach similar to the one used above in comparing the growth rates for girls and boys in school 142, one can employ initial status as a covariate, and thus obtain an estimate of differences in rates of change holding constant initial status.

Secondly, consider a longitudinal study that seeks to compare the effectiveness of an innovative remedial program with one that is more traditional. It may be the case that among students with extreme reading difficulties, rates of progress tend to be more rapid for students in the traditional program. We term phenomena of this kind Initial Status × Program interactions. Analogous to our last example (i.e., the analyses for school 302) in which there was evidence of an interaction between gender and initial status, such interactions could be studied by incorporating Initial Status × Program interaction terms in one’s models for growth rates. See Muthen & Curran (1997), Khoo (2001), Muthen, Jo, & Brown (2003), and Seltzer, Choi & Thum (2001) for illustrative examples. Note that in the spirit of our remarks in the previous section, it is important not to focus exclusively on interactions involving initial status, but rather to consider a variety of variables (e.g., key aspects of a student’s prior educational experiences) that might possibly interact with the effectiveness of programs of interest.

Implications for school effects research

Employing initial status as a predictor of rates of change can also help broaden the kinds of questions that we are able to address in school effects research based on analyses of large-scale longitudinal data sets (e.g., LSAY; NELS). An important point illustrated above is that in some schools, there may be a strong, positive relationship between initial status and subsequent rates of change, such that the gap in achievement between those who start high and those who start low widens appreciably over time. In contrast, in other schools where students start and their subsequent rates of change may be weakly related. By expanding the growth models in this article to include a third level (i.e., a between-school or level-3 model), we can examine how differences in various school policies, practices and intake characteristics relate to differences between schools in their mean rates of change and in the magnitude of their initial status/rate of change slopes. Through applications of growth models of this kind one can attempt to identify, for example, those school factors that appear to eventuate in high mean rates of progress, and in relatively weak relationships between initial status and rates of change (Choi, 2002; Choi & Seltzer, 2003).

Settings in which growth is nonlinear

In our analyses, we employed growth models in which change in math achievement scores was
modeled as a linear function of grade. For each of the cohorts whose time-series data we analyzed (i.e., the time-series data for students in schools 308, 143, 142 and 302), various exploratory analyses and formal tests that we conducted pointed to this being a reasonable and parsimonious representation of growth. This is an important point since the adequacy of the models that we specify for individual growth can have implications for the adequacy of the estimates we obtain for initial status, and, more generally, for the results we obtain regarding relationships between initial status and subsequent patterns of change (see, e.g., Adler, Adam, & Arenberg, 1990). In this regard, it is important to note that the various kinds of analyses presented above can be extended to settings in which patterns of change are nonlinear. For example, consider a school in which growth in student achievement tends to accelerate over time. Suppose further that acceleration is more pronounced for those students with relatively high initial levels of achievement. Patterns of this kind can be detected, summarized and studied more formally by using the modeling framework illustrated above to fit models in which student differences in acceleration are modeled as a function of initial status. For an example involving the use of models for nonlinear growth in intervention settings, see Muthen and Curran (1997).

Software options

A wide variety of software options are available for fitting the kinds of growth models presented above. All results reported in this article were obtained using the software package WinBUGS, which was developed by members of the MRC Biostatistics Unit in Cambridge, UK (see Spiegelhalter, Thomas, Best, & Gilks, 2000). (Note that BUGS is a near acronym for “Bayesian analysis using the Gibbs sampler.”) WinBUGS opens up a variety of modeling possibilities that have proved to be extremely valuable in our growth-modeling work. For example, using WinBUGS, it is possible to obtain robust estimates of parameters of interest (i.e., estimates that are not unduly influenced by outliers) (see, e.g., Seltzer & Choi, 2002). In addition, WinBUGS can be used to conduct the kinds of school effects applications described above. Note that annotated copies of the programs used to conduct the analyses of the LSAY data presented in this article, along with various details regarding their use, are available upon request.

One can also estimate all of the previous growth models using widely available programs for structural equation modeling such as Mplus (Muthen & Muthen, 1998), EQS (Bentler, 2002), LISREL (Joreskog & Sorbom, 2000) and AMOS (Arbuckle & Wothke, 1999). In addition, all but the very last of the two-level models presented in this article can be estimated using HLM5 (Raudenbush, Bryk, Cheong, & Congdon, 2000.)

In short, a number of software options exist for formulating and fitting models that enable us to employ initial status (i.e., the status of students at the start of a time period of substantive interest) as a predictor of change. This, in turn, helps to enrich the kinds of questions that we can address in efforts to monitor school performance, in studying the effects of educational programs and interventions, and in conducting school effects research.

Notes

1 To help grasp the logic of IRT models, it is useful to focus on the Rasch model, which is a relatively basic, but fundamentally important IRT model. As Seltzer, Frank and Bryk (1994) note, it is helpful to view a Rasch-based IRT scale that measures a particular skill as a line or ruler defined by test items arrayed from least difficult to most difficult. In constructing such scales, the ordering of items defining the scale, and the spacing between items, essentially depend on differences in the probability of students correctly answering the various items. As such, the ordering and spacing of items that form the basis of the scale reflect “differences in the amount of [skill] that is required for mastery” (Seltzer, Frank & Bryk, 1994, p. 43). Changes in a student’s score over time—that is, “changes in the location [or placement] of a student along the scale”—give us a sense of how rapidly or slowly the student is mastering the particular skill measured by the test, whether the student experienced a spurt in growth during a particular time period, and the like (Seltzer, Frank & Bryk, 1994, p. 48). Note that this is in marked contrast to the use of commonly reported scales based on norm-referenced tests, such as percentile ranks or grade equivalents (GE). Consider, for example, a set of students in a longitudinal study who score at the 50th percentile on a standardized test of reading comprehension in grades 3, 4 and 5. Though the actual reading comprehension skill of such students is surely increasing across this series of grades, a time series plot based on the use of percentile ranks would yield a flat trajectory. From the time-series of percentile ranks we do have a sense of the standing or ranking of these students in relation to the individuals in a particular norming sample. But how much of an actual increase in reading comprehension skills are such students experiencing?
Are the actual amounts of improvement in comprehension that they tend to experience greater between grades 3 and 4 than between grades 4 and 5? Percentile ranks are not particularly useful in answering such questions. Problems of this nature also arise in connection with GE scores and NCE scores (see, for example, Seltzer, Frank & Bryk, 1994).

2 Based on the model for student growth specified in Equation 1, the expected score for student \(i\) at the start of grade 7 (i.e., \(\text{GRADE}_i = 7\)) is: \(\pi_{0i} + \pi_{1i}(7 - 7) = \pi_{0i}\). Had we not centered \(\text{GRADE}\), then \(\pi_{0i}\) would have represented the expected achievement score for student \(i\) at the start of grade 0, which clearly is not very useful for our purposes.

3 A correlation coefficient, as Hays (1988, p. 555), notes is the standardized covariance between two variables. It can be computed by dividing the covariance between two variables (e.g., \(\text{Cov}(A, B)\)) by the square root of the product of their variances (e.g., \(\sqrt{\text{Var}(A) \times \text{Var}(B)}\)). Thus, the correlation between initial status and rates of change can be expressed as a function of the covariance between initial status and rates of change (\(\tau_{10}\)), the variance in initial status (\(\tau_{00}\)), and the variance in rates of change (\(\tau_{11}\)):

\[
\rho = \frac{\tau_{10}}{\sqrt{\tau_{00} \times \tau_{11}}}.
\]

4 The parameter estimates and 95% intervals that we report in our tables are posterior medians and .025 and .975 quantiles obtained via WinBUGS. These can be viewed as Bayesian analogues of point estimates and confidence intervals. Note that BUGS stands for Bayesian analysis using the Gibbs sampler. For discussions of the logic of Bayesian inference, its similarities to classical approaches to estimation and inference, and its advantages, see Raudenbush and Bryk (2002, chpt. 13), Gelman, Carlin, Stern, and Rubin (1995), and Carlin and Louis (1996). The latter references also contain discussions of the Gibbs sampler; see also Gelfand, Hills, Racine-Poon, and Smith (1990), Seltzer (1993), Seltzer & Choi (2002), Tanner (1996), and Gilks, Richardson and Spiegelhalter (1996). Note that the Gibbs sampler is an estimation tool that makes it possible to obtain estimates and intervals for parameters of interest in a wide-range of complex modeling settings. Annotated copies of the WinBUGS programs that we used in our analyses, along with details regarding their implementation, are available upon request.

5 To obtain fitted trajectories, the model for individual growth specified in Equation 1 was fit to each student’s time series data using Ordinary Least Squares (OLS). That is, the fitted trajectory for a given student was obtained by regressing the student’s time series of math achievement scores on \(\text{GRADE}\).

6 Because \(\pi_{0i}\) is included as a predictor of \(\pi_{1i}\), the residuals \(U_{0i}\) and \(U_{1i}\) are assumed to be uncorrelated, i.e., \(\text{Cov}(U_{0i}, U_{1i}) = 0\).

7 Note that in the between-student model specified in Equation 3, and in subsequent between-student models, we employ various centerings to help give the intercepts in these models (e.g., \(\beta_{10}\)) more useful interpretations. We feel that presenting these centerings in the main body of the text would interfere with conveying key points and modeling possibilities. Therefore, we detail and discuss the value of these centerings in the appendix.

8 Note that our model differs from more standard ANCOVA models in a crucially important way. In ANCOVA analyses, we often employ student pretest scores as covariates (e.g., \(X_i\)). Such scores contain measurement error. A potential problem is that measurement error contained in \(X_i\) can attenuate estimates of the regression coefficient for \(X_i\). This, in turn, could result in underadjustments for initial group differences (see Reichardt, 1979 for a valuable explanation of this phenomenon). One way of overcoming such problems is to employ latent variables as covariates. Thus, rather than employing observed (e.g., \(Y\)) grade 7 math achievement scores as a covariate, we employ the latent variable \(\pi_{0i}\)—a parameter capturing initial status—as our covariate.

9 In comparing growth rates for different demographic groups, one should bear in mind that the amount of variance in initial status and growth rates for one group may differ from the amount of variance in initial status and growth rates for another group. In allowing for an interaction between initial status and gender in modeling growth rates for students in school 302, we also allowed for the possibility that the amount of remaining variance in growth rates for boys may differ appreciably from the amount of remaining variance for girls. This was accomplished by specifying separate residual variances for boys (\(\tau_{11B}\)) and girls (\(\tau_{11G}\)). As seen in Table 6, the estimated residual variance for boys is substantially smaller than the estimated residual variance for girls. This finding is sensible in light of the fact that the estimate of the initial status/rate of change slope for boys takes on a fairly large positive value, while the estimate of the initial status/rate of change slope for girls is close to a value of 0. As such, employing initial status as a predictor of growth rates accounts for substantially more variance in growth rates for boys than for girls.

10 It is useful to think in terms of a rank ordering of students based on their rates, and rank orderings of students based on their levels of achievement at different points in time. In the case of school 143, a rank ordering of students based on their status at grade 7 will correlate weakly with a rank ordering based on their rates. However, with many students with faster rates surpassing and attaining higher levels of achievement than students with slower rates, a rank ordering of students based on their status at grade 10 will be associ-
Exchanging Relationships Between Where Students Start and How Rapidly they Progress

In fact, while we have focused on the relationship between status at grade 7 and rates of change in our article, it is also possible to examine the relationship between status at another point in time (e.g., grade 10) and rates of change. This simply requires changing the centering employed in one’s within-student (Level-1) model. For example, in the illustrative examples in our article, by virtue of centering GRADE in Equation 1 around a value of 7, \( \pi_i \) represents the expected achievement (status) for student \( i \) at the start of grade 7. If we were to center GRADE around a value of 10, \( \pi_i \) would represent the status for student \( i \) at the start of grade 10; in the context of the series of grades we are focusing on in our examples, status at the start of grade 10 would be termed “final status.” By modifying Equation 1 in this way, and employing a between-student model such as the one defined in Equations 2a and 2b, we would be able to obtain an estimate of the correlation between status at the start of grade 10 and rates of change.

11 Longitudinal studies of programs and interventions are often characterized by designs in which data are collected at a single point in time prior to the start of a study’s treatment phase, and at several time points during a study’s treatment phase. In such cases, an estimate of initial status for a given person essentially derives from fitting a trajectory of a suitable form (e.g., a line or curve) to a series of observations for that person consisting of a single baseline observation and the observations collected during the study’s treatment phase. When feasible, we would encourage the use of designs in which data are collected at several points in time prior to the start of a treatment, and then at a series of time points during the treatment phase (see, e.g., Muthen & Curran, 1997, p. 379). We could then employ models for individual growth that in essence enable us to fit separate trajectories to a person’s pretreatment phase data and treatment phase data. As such, we would be able to obtain estimates of pretreatment status and, for example, pretreatment growth rates, that are rooted strictly in the data collected prior to the start of treatment. Moreover, we could compare growth for treatment and comparison group members during the treatment phase of a study, controlling for possible differences in, for example, status at the end of the pretreatment phase and growth rates in the pretreatment phase. We encourage the collection of follow-up data as well.

12 For checking purposes, we re-ran the analyses reported in Tables 1–5 using HLM. The results that we obtained via HLM for key fixed effects (e.g., \( \beta_{10} \) in Equation 2b; \( \beta_{1} \) in Equation 5b) and latent variable regression coefficients (e.g., \( b \) in Equations 3b and 5b) are extremely similar to those that we obtained using WinBUGS. Finally, we re-ran the analysis reported in Table 6 using Mplus (Muthen & Muthen, 1998). The Mplus results for key latent variable regression parameters (i.e., \( b_1 \) and \( b_2 \) in Equation 7) are extremely close to those reported in Table 6.

The estimation strategy used by the HLM program to estimate latent variable regressions (see Raudenbush & Sampson, 1999) can in principle be extended to settings in which one wishes to specify interactions between initial status and various dichotomous predictors. This would involve computing separate Level-2 variance and covariance estimates for the different groups (e.g., girls and boys) one wishes to compare.

Appendix

In each of the between-student models in which we have specified initial status as a predictor of rates of change, we have employed centerings that help give intercept terms in these models (e.g., \( \beta_0 \) in Equation 3b) more meaningful interpretations. In the interest of trying to make the logic of these models as accessible as possible, we did not present these centerings in the main body of the article. The purpose of this appendix is to detail these centerings.

1. In modeling rates of change as a function of initial status for the students in school 308, we employed the following between-student model:

\[
\pi_{0i} = \beta_{10} + U_{0i} \tag{8}
\]

\[
\pi_{1i} = \beta_{10} + b(\pi_{0i} - \beta_{00}) + U_{1i} \tag{9}
\]

In this model, \( \beta_{00} \) represents the mean initial status for students in school 308. Note that in the model for rates of change (Equation 9), we have centered \( \pi_{0i} \) around \( \beta_{00} \) (cf. Equation 3b). This helps give \( \beta_{10} \) a more useful interpretation. Specifically, \( \beta_{10} \) represents the expected growth rate for a student whose initial status value is equal to the mean initial status value for school 308. Had we not centered \( \pi_{0i} \), \( \beta_{10} \) would represent the expected growth rate for a student whose initial status value is equal to 0. This clearly would not be very meaningful since the minimum grade 7 mathematics achievement score for students in this school is approximately 30 points. Note that we employed a similar centering in our analysis of the data for school 143.

Upon fitting the above model to the data for school 308, for example, we can compute expected rates of change for different values of initial status by substituting the estimates for \( \beta_{10} \),
\( b \) and \( \beta_{00} \), shown in Table 2 into Equation 9, and setting the residual term (i.e., \( U_{0i} \)) to a value of 0. This gives us the following equation for computing expected rates: 3.49 + .174 (\( \pi_{0i} - 44.73 \)). For students with initial status values equal to the mean initial status value (i.e., \( [\pi_{0i} - 44.73] = 0 \)), the expected rate of change is: 3.49 + .174 (0) = 3.49. For students with initial status values 10 points above the mean initial status value (i.e., \( [\pi_{0i} - 44.73] = 10 \)), the expected rate is: 3.49 + .174 (10) = 5.23. Finally, for students with initial status values 10 points below the mean initial status value (i.e., \( [\pi_{0i} - 44.73] = -10 \)), the expected rate is: 3.49 + .174 (-10) = 1.75. These initial status values and corresponding expected rates define the plotted trajectories displayed in Figure 7. Note that using WinBUGS one can readily obtain 95% intervals as well as point estimates for expected rates.

2. In the case of the between-student model used to compare rates of change for girls and boys in school 142 holding constant initial status, we employed the following centerings:

\[
\pi_{0i} = \beta_{00} + \beta_{01}(GENDER - \overline{GENDER}) + U_{0i} \tag{10}
\]

\[
\pi_{1i} = \beta_{10} + \beta_{11}GENDER + b(\pi_{0i} - \beta_{00}) + U_{1i} \tag{11}
\]

In modeling initial status as a function of \( \text{GENDER} \) (Equation 10), we have centered \( \text{GENDER} \) around its grand mean. If \( \text{GENDER} \) were uncentered, \( \beta_{00} \) would represent the expected initial status for boys in school 142, and \( \beta_{00} \) would capture the expected difference in initial status between girls and boys. As a result of the centering employed in Equation 10, \( \beta_{01} \) still represents the expected difference in initial status between girls and boys. However, \( \beta_{00} \) now takes on a more useful interpretation for our purposes, i.e., \( \beta_{00} \) represents the grand mean initial status value for students in school 142 (see, e.g., Raudenbush & Bryk (2002) for a discussion of centering level-2 predictors). In the case of classic ANCOVA models, covariates such as pretest scores (e.g., \( X \)) are typically deviated from their grand means (\( X_i - \overline{X} \)) (see, e.g., Reichardt, 1979). Similarly, in the model for growth rates in Equation 6, our covariate (\( \pi_{0i} \)) is deviated from its grand mean (i.e., \( \pi_{0i} - \beta_{00} \)). As a result, \( \beta_{10} \) now represents the expected rate of change for a male student whose initial status value is equal to the grand mean (i.e., \( \pi_{0i} - \beta_{00} = 0 \)). We can also refer to \( \beta_{10} \) as an adjusted rate of change for boys. Similarly, \( \beta_{10} + \beta_{11} \) represents the expected rate of change for a female student whose initial status value is equal to the grand mean, i.e., it is an adjusted rate of change for girls. These adjusted rates are analogous to adjusted posttest means in ANCOVA settings.

3. In the case of the interaction model employed in our analysis of the data for girls and boys in school 302, we used centerings similar to those in Equations 10 and 11:

\[
\pi_{0i} = \beta_{00} + \beta_{01}(GENDER - \overline{GENDER}) + U_{0i} \tag{12}
\]

\[
\pi_{1i} = \beta_{10} + \beta_{11}GENDER + b(\pi_{0i} - \beta_{00})
+ b_2(GENDER \times (\pi_{0i} - \beta_{00})) + U_{1i}. \tag{13}
\]

In Equation 13, \( \beta_{10} \) represents an adjusted rate of change for boys in school 302 (i.e., an expected rate for a boy whose initial status is equal to the mean initial status value for school 302), and \( \beta_{10} + \beta_{11} \) represents an adjusted rate for girls. Note that analogous to adjusted means in ANCOVA models, adjusted rates based on models that allow for different initial status/rate of change slopes for the groups we wish to compare, can differ from adjusted rates based on models in which initial status/rate of change slopes are assumed to be equivalent.

Based on Equation 13, expected rates of change for boys (\( GENDER = 0 \)) have the following form: \( \beta_{10} + b_1 (\pi_{0i} - \beta_{00}) \). Employing the estimates for \( \beta_{10}, b_1 \) and \( \beta_{00} \) displayed in Table 6, we can compute expected rates of change for boys given various values of initial status based on the following equation: 4.26 + 0.142 (\( \pi_{0i} - 52.54 \)). For boys with initial status values equal to the mean initial status value for school 302 (i.e., \( \pi_{0i} - 52.54 = 0 \)), the expected rate of change is: 4.26 + .142 (0) = 4.26. For boys with initial status values 12 points below the mean initial status value (i.e., \( \pi_{0i} - 52.54 = -12 \)), the expected rate is: 4.26 + .142 (-12) = 2.56. Finally, for boys with initial status values 12 points above the mean initial status value (i.e., \( \pi_{0i} - 52.54 = 12 \)), the expected rate is: 4.26 + .142(12) = 5.96.

Expected rates of change for girls based on Equation 13 have the following form: \( (\beta_{10} + \beta_{11}) + (b_1 + b_2) (\pi_{0i} - \beta_{00}) \). Substituting the estimates of \( \beta_{10}, \beta_{11}, b_1, b_2 \) and \( \beta_{00} \) in Table 6 into this equation, gives rise to the following equation for computing expected rates of change for girls:
(4.26 + .20) + (0.142 − 0.121) (π_{0i} − 52.54). This reduces to: 4.46 + 0.021 (π_{0i} − 52.54). Thus, for example, the expected rate of change for girls with initial status values 12 points below the mean initial status value for school 302 is: 4.46 + 0.021 (−12) = 4.21.

References


Seltzer, M., Choi, K., & Thum, Y. M. (2001). Latent variable modeling in the hierarchical modeling

285
framework: Exploring initial status x treatment interactions in longitudinal studies of interventions.

Authors
MICHAELE SELTZER is Associate Professor, Graduate School of Education and Information Studies, UCLA, 2019-B Moore Hall, Los Angeles, CA 90095-1521; mseltzer@ucla.edu. He specializes in the analysis of data from longitudinal studies and multi-site evaluations.
KILCHAN CHOI is a Senior Researcher at CRESST, UCLA GSE&IS Bldg. Los Angeles, CA 90095-1522; kcchoi@ucla.edu. He specializes in methods for analyzing longitudinal data.
YEOW MENG THUM is Assistant Professor Department of Education, Graduate School of Education and Information Studies, 2019-A Moore Hall, UCLA; thum@ucla.edu. He specializes in models for growth and change.

Manuscript Received December 5, 2001
Revision Received March 31, 2003
Accepted June 20, 2003

Downloaded from http://eepa.aera.net at ACER on September 28, 2008
Using Covariance Structure Analysis to Detect Correlates and Predictors of Individual Change Over Time

John B. Willett and Aline G. Sayer

Recently, methodologists have shown how two disparate conceptual arenas—individual growth modeling and covariance structure analysis—can be integrated. The integration brings the flexibility of covariance analysis to bear on the investigation of systematic interindividual differences in change and provides another powerful data-analytic tool for answering questions about the relationship between individual true change and potential predictors of that change. The individual growth modeling framework uses a pair of hierarchical statistical models to represent (a) within-person true status as a function of time and (b) between-person differences in true change as a function of predictors. This article explains how these models can be reformatted to correspond, respectively, to the measurement and structural components of the general LISREL model with mean structures and illustrates, by means of worked example, how the new method can be applied to a sample of longitudinal panel data.

Questions about correlates and predictors of individual change over time are concerned with the detection of systematic interindividual differences in change, that is, whether individual change in a continuous outcome is related to selected characteristics of a person’s background, environment, treatment, or training. Examples include the following: Do the rates at which psychological adjustment related to health status, gender, and home background?

Questions like these can be answered only when continuous data are available longitudinally on many individuals, that is, when both time points and individuals have been sampled representatively. Traditionally, researchers have sampled individual status at only two points in time, a strategy that has proven largely inadequate because two waves of data contain only minimal information on individual change (Rogosa, Brandt, & Zimowski, 1982; Willett, 1989). When true development follows an interesting trajectory over time, “snapshots” of status taken before and after are unlikely to reveal the intricacies of individual change.

In the last 15 years, the methods of individual growth modeling have capitalized on the richness of continuous multivariate data to provide better methods for answering questions about systematic interindividual differences in change (Bryk, 1977; Bryk & Raudenbush, 1987; Rogosa et al., 1982; Rogosa & Willett, 1985; Willett, 1988, 1994). Under this approach, an individual growth model is chosen to represent the change that each person experiences with time. This is often referred to as the within-person or Level 1 model. All members of a given population are assumed to have trajectories of the same functional form, but different members can have different values of the individual growth parameters present in the Level 1 model. For instance, if individual change is linear with time, interindividual differences in change may be due to heterogeneity in initial status (intercept) and rate of change (slope). Alternatively, if individual change is a quadratic function of time, then interindividual differences in change may also be due to between-person variation in the curvature parameter. Interindividual differences in change are said to be systematic when between-person variation in one or more individual growth parameters is related to variation in the selected predictors of change. The hypothesized link between the individual growth parameters and the predictors of change is described in a between-person or Level 2 statistical model.

A variety of methods have been proposed for estimating the parameters of the Level 1 and Level 2 models in the analysis of change. Rogosa and his colleagues (Rogosa et al., 1982; Rogosa & Willett, 1985; Willett, 1985, 1994; Williamson, 1986; Willett, 1989; Zimowski, 1982).
liamson, Appelbaum, & Epanchin, 1991) described exploratory ordinary least squares regression-based methods to separately estimate the parameters of the Level 1 and Level 2 models, with reliability-based adjustments to the latter based on the marginal maximum likelihood methods of Blomqvist (1977). In an extension of the exploratory approach, Willett (1988, based on Hanushek, 1974) provided weighted least squares methods for obtaining asymptotically efficient estimates of the parameters of the Level 2 model. And, as part of their work on hierarchical linear modeling (HLM), Bryk and Raudenbush (1987, 1992) described strategies for simultaneously estimating the parameters of the Level 1 and Level 2 models using empirical Bayes estimation.

Recently, several pioneering authors have demonstrated how the analysis of change can be conducted very conveniently by the methods of covariance structure analysis. Meredith and Tisak (1984, 1990; see also Tisak & Meredith, 1990), for instance, have provided a technical framework for representing interindividual differences in intraindividual development and examples of how model parameters can be estimated by covariance structure analysis. Their work extends earlier research on longitudinal factor analysis (Rao, 1958; Tucker, 1958) and subsumes more traditional approaches to the analysis of panel data, such as repeated measures analysis of variance (ANOVA) and multivariate analysis of variance (MANOVA; see also Jöreskog & Sörbom, 1989). The Meredith–Tisak approach is very general. It permits the evaluation of the general shape of the individual growth trajectories and provides not only estimates of the individual growth parameters (through the estimation of factor scores) but estimates of the Level 2 means, variances, and covariances of the individual growth parameters across all members of the population. These latter statistics estimate the population average growth curve and provide evidence for the presence of interindividual differences in growth in the population.

In a linked body of applied work, McArdle and his colleagues have extended the covariance structure approach of Meredith and Tisak, demonstrating its flexibility by application to a wide variety of developmental problems in psychology and throughout the social sciences. For instance, they showed how covariance structure methods can be used to estimate average growth curves and to indicate the presence of interindividual differences in change in a single domain and simultaneously in many domains (McArdle, 1986a, 1986b, 1989, 1991; McArdle & Epstein, 1987). They also showed how average growth curves can be compared across groups (McArdle, 1989; McArdle & Epstein, 1987) and described how covariance structure methods can be used to conduct convergence analysis, in which segments of average growth curves estimated in overlapping cohorts are linked into a single continuous trajectory (McArdle & Anderson, 1989; McArdle, Anderson, & Aber, 1987; McArdle & Hamagami, 1991). Furthermore, in an extension that provided the impetus for this article, McArdle and Epstein (1987) demonstrated how the Level 2 relationship between slope and a single predictor of change can be modeled and estimated when individual change is represented by a restricted linear growth model that contains only a slope parameter (and no intercept).

Finally, in a separate but related stream of research, Muthén and his colleagues have also described the technical basis for, and provided data-analytic examples of, the modeling of multilevel data using covariance structure methods (Muthén, 1989; Muthén & Satorra, 1989). Of particular interest are a pair of recent papers (Muthén, 1991, 1992) in which the parameters of a linear individual growth model were allowed to vary across individuals in ways systematically related to selected time-invariant predictors of change. The latter paper also presented an application of an interesting strategy based on multigroup analysis (see Bollen, 1989) for incorporating individuals with incomplete observed growth records into the analyses.

We are convinced that these new methods are another powerful data-analytic weapon in the armory of the empirical researcher. Questions about correlates and predictors of change pervade psychology and the social sciences. In this article, we present a careful exposition of the application of covariance structure methods to the investigation of systematic interindividual differences in change. Analytically, our article is closest to—and derivative of—the work of McArdle and Epstein (1987) and Muthén (1991, 1992). However, we intend our principal contribution to be the provision of a viewpoint that inverts that of our covariance modeling colleagues. Unlike them, we owe our principal allegiance to the field of individual growth modeling, and we have come to understand these recent innovations in the covariance structure analysis of longitudinal data from that perspective. Consequently, in this article, we have tried to link the pioneering contributions of McArdle, Muthén, and Meredith and Tisak directly to recent developments in the measurement of individual change (Bryk, 1977; Bryk & Raudenbush, 1987, 1992; Burchinall & Appelbaum, 1991; Rogosa et al., 1982; Rogosa & Willett, 1985; Willett, 1985, 1988, 1989, 1994; Williamson, 1986; Williamson et al., 1991).

The integration of the individual growth modeling and covariance structure approaches capitalizes on the fundamental mathematical equivalence of two alternative methods of representing the same data structure. The process of formulating population Level 1 and Level 2 models for individual change and for systematic interindividual differences in change is equivalent to postulating a specific structure for the matrix of population covariances among the multiple waves of observed data and the predictors of change. By using the general LISREL model with mean structures (Jöreskog & Sörbom, 1989) to explicitly articulate this latter covariance structure and to fit it to the matrix of sample covariances, we can obtain maximum likelihood estimates of the critical between-person parameters that were specified under the original growth modeling formulation and thereby answer our research questions about potential correlates and predictors of change.

This article has three sections and a concluding discussion. In the first section, we begin our presentation by introducing a simple Level 1 growth model to represent individual change over time. In this model, we hypothesize that true change is a linear function of time, and we assume that the occasion-by-occasion errors of measurement are both homoscedastic and mutually independent. Then we formulate a preliminary "no-predictor" Level 2 model for interindividual differences in change that describes heterogeneity in change across members of the population. In this latter model, the intercepts and slopes of the Level 1 individual growth model are permitted to vary...
joined across people, but interindividual heterogeneity in change is left unexplained by potential predictors of change. In the second section, we illustrate how the "baseline" individual growth modeling perspective maps onto the framework provided by the general LISREL model. Then we demonstrate how the respective Level 2 means, variances, and covariance of the individual growth parameters can be estimated straightforwardly by covariance structure analysis. To close out the section, we extend the baseline analysis to test and modify the critical assumptions made at Level 1 concerning (a) the linearity of the individual true growth trajectory and (b) the homoscedasticity and independence of the measurement error covariance structure. In the third section, we introduce potential predictors of change into the baseline Level 2 specification and describe how the methods of covariance structure analysis can be used to estimate the relationship between interindividual heterogeneity in the growth parameters and the predictors of change. Our rationale for this organization is both substantive and methodological because, logically, individual change must be described before interindividual differences in change can be examined, and interindividual differences in change must be present before one can ask whether any heterogeneity is related to predictors of change.

A data-based example is used to frame our presentation throughout the article, and we provide, in the Appendix, illustrative LISREL programs for conducting the proposed analyses. In our concluding discussion, we comment on the advantages and limitations of the covariance structure approach and note further extensions of the method that are made feasible by the flexibility of the general LISREL model.

Modeling Change Over Time: An Individual Growth Modeling Perspective

To answer research questions about individual change in a continuous variable, a representative sample of individuals must be observed systematically over time, with their status being measured on several temporally spaced occasions. To use the covariance structure approach described here, three or more waves of data must be available on each individual. In addition, the data must be balanced in a particular way. The occasions of measurement need not be equally spaced in time, but both the number and the spacing of assessments must be the same for all individuals, a pattern that Bock (1979) referred to as “time-structured” data.

Thus, the methods we describe here are appropriate for analyzing panel data, in which the number of individuals is large with respect to the number of occasions of measurement. The sample size must be large enough to enable the investigator to detect person-level effects. Other analytic methods are available for time-series data sets, which typically contain many repeated assessments on few individuals. In psychology, studies of physical growth, language development, and dyadic interaction are often of this latter form.

Introducing the Data Example

Throughout this article, we illustrate the covariance structure analysis of change using a data set that contains five waves of data obtained from a sample of 168 adolescents on the continuous dependent variable tolerance of deviant behavior. During each year of the study—at 11, 12, 13, 14, and 15 years of age—each participant completed a nine-item instrument that asked whether it was wrong for someone his or her age to cheat on tests, purposely destroy property of others, use marijuana, steal something worth less than $5, hit or threaten someone without reason, use alcohol, break into a building or vehicle to steal, sell hard drugs, or steal something worth more than $50 (Raudenbush & Chan, 1992). Responses to each item were registered on a 4-point scale ranging from very wrong (1) to not wrong at all (4), and responses were averaged across items to provide a scale score. In Table 1, for illustration, we list scores for 16 randomly selected cases. Inspection of the table—and the full data set—suggests that although there is considerable scatter in the observed scores over time and across individuals, adolescents appear to become gradually more tolerant of deviant behavior as they age.

To answer questions about systematic interindividual differences in change, information must also be available on potential predictors of change. Our data example provides the values of two potential predictors of change: (a) the gender of the adolescent (0 = male, 1 = female), and (b) the adolescent’s reported exposure to deviant behavior in the 1st year of data collection (at age 11). Values of these predictors are also presented for the 16 randomly selected cases in Table 1. In our illustrative data analyses, the broad research question is as follows: Are interindividual differences in change of tolerance toward deviance during adolescence related to respondent gender and initial exposure to deviant behavior?

Modeling Individual Change Over Time

As is well known, classical test theory describes the psychometric properties of scores on a single occasion, distinguishing observed score from true score, the former being a fallible operationalization of the latter. The observed score continues to be distinguished from the true score when change is investigated because change in underlying true score is the focus of research interest. Measurement error is an uninvited guest that randomly obscures the true growth trajectory from view. Consequently, when individual growth is represented by a statistical model, the model must contain a part describing a person’s true growth trajectory over time and a part representing the stochastic effect of measurement error.

---

1 We thank Stephen W. Raudenbush for providing these data.
2 Adolescents’ exposure to deviance was also self-reported on a nine-item instrument in each of the 5 years of the study. Participants were asked how many of their peers engaged in the same nine activities identified in the tolerance of deviant behavior instrument. For each item, ratings were obtained on a 5-point scale ranging from none of them (0) to all of them (4). Scale scores were computed by averaging across items. To simplify the presentation, we have chosen to use only the participant’s initial scale score at age 11 as a predictor of change. Further analyses could capitalize on the potentially time-varying nature of the exposure data to answer more complex questions than we are asking here (e.g., Is change over time in adolescents’ tolerance of deviant behavior related to change in their exposure to that behavior?).
Under the individual growth modeling framework, the "true" part of each person's growth trajectory is represented by an algebraic function of time. Many possible mathematical functions are available, both those that depend linearly on time and those that do not. Choice of an appropriate mathematical function to represent true individual change is an important first step in any project.3

A responsible preliminary strategy for choosing a valid growth model is to inspect each person's empirical growth record by plotting his or her observed status against time (see Willett, 1989). As an aid to inspection, we also find it useful to examine of wave-by-wave univariate descriptive statistics on each individual's plot. Such exploratory analyses permit the investigator to compare the appropriateness of alternative models through descriptive statistics that describe the fit of each model. In our data example, this process—along with examination of wave-by-wave univariate descriptive statistics on the dependent variable—suggested that the distribution of the natural logarithm of tolerance of deviance was less skewed and that, once transformed, a linear (straight-line) model was an appropriate representation of individual change in log tolerance of deviance over the adolescent years.4

As an evocative summary of our data exploration, Figure 1 presents OLS-fitted observed straight-line growth curves for the adolescents in Table 1. Notice that the observed (log) tolerance of deviance of most adolescents increases as time passes and that there is evidence of heterogeneity in observed change across people. This collection of trajectories must be interpreted carefully, however, because the trajectories summarize observed rather than true change. It is possible, for instance, that all differences among the slopes of the 16 fitted trajectories can be attributed to measurement error rather than to heterogeneity in true rate of change. The methods described in this article can distinguish these two eventualities.

The methods described here do not explicitly demand the preliminary inspection of empirical growth records. In subsequent analysis, one can confirm earlier "eyeball" suspicions by testing whether higher order nonlinear terms must be added to the individual growth function. Nonetheless, we recommend that every analysis of change begin with individual-level data exploration. Sound data-analytic practice demands knowledge of the data at the lowest level of aggregation so that anomalous cases can be identified, outlying data points can be detected, and assumptions can be checked. Once data are summarized in a covariance matrix, all individual-level richness is lost.

As we have noted, in our data example, preliminary exploration suggested that a straight-line function was most appropriate for modeling change in log tolerance of deviance over the adolescent years. Therefore, we model true individual change in log tolerance as a linear function of time, with a stochastic term added to account for the influence of measurement error:

\[ Y_{it} = \pi_{0i} + \pi_{1i}t_i + \epsilon_{it}, \]  (1)

3 Ideally, theory will guide the rational choice of model so that the specified individual growth parameters have meaningful substantive interpretations. Often, however, the mechanisms governing the change process are poorly understood and, thus, a well-fitting polynomial is used to approximate the trajectories. Also, in much research, only a restricted portion of the life span is observed with few waves of data collected, and so the selected growth model must contain a small number of individual growth parameters if the model is to be fitted successfully. Thus, the most popular growth model is often a linear or a quadratic function of time.

4 Details are available on request. Raudenbush and Chan (1992) also transformed the dependent variable logarithmically.

---

**Table 1**

*Data on a Subsample of 16 Drawn at Random From the 168 Adolescents in the Example*

<table>
<thead>
<tr>
<th>Subject ID number</th>
<th>Reported tolerance of deviant behavior</th>
<th>Predictors of change</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Age 11</td>
<td>Age 12</td>
</tr>
<tr>
<td>0009</td>
<td>2.23</td>
<td>1.79</td>
</tr>
<tr>
<td>0045</td>
<td>1.12</td>
<td>1.45</td>
</tr>
<tr>
<td>0258</td>
<td>1.45</td>
<td>1.34</td>
</tr>
<tr>
<td>0314</td>
<td>1.22</td>
<td>1.22</td>
</tr>
<tr>
<td>0442</td>
<td>1.45</td>
<td>1.99</td>
</tr>
<tr>
<td>0514</td>
<td>1.34</td>
<td>1.67</td>
</tr>
<tr>
<td>0569</td>
<td>1.79</td>
<td>1.90</td>
</tr>
<tr>
<td>0624</td>
<td>1.12</td>
<td>1.12</td>
</tr>
<tr>
<td>0723</td>
<td>1.22</td>
<td>1.34</td>
</tr>
<tr>
<td>0918</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>0949</td>
<td>1.99</td>
<td>1.55</td>
</tr>
<tr>
<td>0978</td>
<td>1.22</td>
<td>1.34</td>
</tr>
<tr>
<td>1105</td>
<td>1.34</td>
<td>1.90</td>
</tr>
<tr>
<td>1542</td>
<td>1.22</td>
<td>1.22</td>
</tr>
<tr>
<td>1552</td>
<td>1.00</td>
<td>1.12</td>
</tr>
<tr>
<td>1653</td>
<td>1.11</td>
<td>1.11</td>
</tr>
</tbody>
</table>

*On a 4-point scale ranging from very wrong (1) to not wrong at all (4).

*0 = male, 1 = female.
A collection of ordinary least squares (OLS) fitted trajectories summarizing observed linear growth in (log) tolerance for deviant behavior between ages 11 and 15 for the subsample of 16 randomly selected youths whose empirical growth records are displayed in Table 1.

The shape of the trajectory for a particular person depends on the way in which time is parameterized in the model and on the values of the constants—the individual growth parameters—that appear on the right-hand side of the model. The straight-line individual growth model in Equation 1 contains a linear parameterization of time and a pair of individual growth parameters representing the intercept and slope of the true trajectory. The slope parameter \( \pi_\text{ip} \) is easy to interpret. It represents rate of change in true status over time for the \( p \)th person. In our example, with time measured in years, \( \pi_\text{ip} \) represents the yearly rate of true change in log tolerance of deviance. Adolescents whose tolerance increased rapidly with time will have large positive values of this parameter; those whose tolerance increased less rapidly will possess correspondingly smaller values.

Mathematically, the intercept parameter \( \pi_\text{ip} \) is also easy to interpret: It is the true status of person \( p \) when \( t = 0 \). However, in many research projects there is no natural or convenient origin for time; thus, the investigator can control the interpretation of the intercept parameter by defining Time 0 at some interesting or important point in the life course. In our example, for instance, although we possess measurements on each person at 11, 12, 13, 14, and 15 years of age, we chose to define the third occasion of measurement as the origin of time (i.e., we set \( t_3 = 0 \)). Our measurement times are thus \( t_1 = -2 \), \( t_2 = -1 \), \( t_3 = 0 \), \( t_4 = 1 \), and \( t_5 = 2 \) years, and the intercept parameter \( \pi_\text{ip} \) represents the true value of log tolerance of deviance for the \( p \)th adolescent at 13 years of age. A word of caution: Like many other common forms of statistical analysis, the individual growth modeling approach is applicable only if it is intuitively sensible to measure change in the outcome variable. At the very least, this means that the outcome variable must have three properties. First, it must be a continuous variable at either the interval or ratio level. Second, it must be equatable from occasion to occasion (i.e., each scale point on the measure must retain an identical meaning as time passes). Finally, it must remain construct valid for the entire period of observation. If any of these conditions are violated, the methods that we describe here are being inappropriately applied.

A Matrix Representation of the Empirical Growth Record

In our data example, five waves of longitudinal data were collected. Therefore, each person's empirical growth record contains five measurements of observed status: \( Y_{1p} \), \( Y_{2p} \), \( Y_{3p} \), \( Y_{4p} \), and \( Y_{5p} \). Under the individual growth model in Equation 1, these measurements can be represented conveniently as

\[
\begin{bmatrix}
Y_{1p} \\
Y_{2p} \\
Y_{3p} \\
Y_{4p} \\
Y_{5p}
\end{bmatrix} = \begin{bmatrix}
1 & t_1 \\
1 & t_2 \\
1 & t_3 \\
1 & t_4 \\
1 & t_5
\end{bmatrix} \begin{bmatrix}
\pi_{1p} \\
\pi_{2p} \\
\pi_{3p} \\
\pi_{4p} \\
\pi_{5p}
\end{bmatrix} + \begin{bmatrix}
\epsilon_{1p} \\
\epsilon_{2p} \\
\epsilon_{3p} \\
\epsilon_{4p} \\
\epsilon_{5p}
\end{bmatrix}.
\]

(2)

For pedagogical reasons, in Equation 2 and throughout the text, we have retained symbols \( t_1 \) through \( t_5 \) to represent the times at which the panel data were obtained. In any particular research project, each of these symbols will have a known constant value. For instance, in our data example, because of our earlier centering of the time axis at age 13, \( t_1 \) through \( t_5 \) have the values \(-2\), \(-1\), 0, 1, and 2 years, respectively.

Note that, by algebraic manipulation, we have shown that the observed growth record for person \( p \) can be written as a combination of three parts: (a) a matrix of known times and constants identical across all individuals, multiplied by (b) an individual-specific vector of unknown individual growth parameters (which we refer to henceforth as the latent growth vector), and added to (c) an individual-specific vector of unknown errors of measurement. This representation provides a critical conceptual emphasis showing that, for each person, one can view the empirical growth record (the observed score vector on the left of Equation 2) as a weighted linear combination of the elements of an unobserved latent growth vector added to a measurement error vector. It is the latent growth vector that is the focus of a study of interindividual differences in change. In our data example, the two elements of the latent growth vector represent the within-person signal: the individual growth parameters that describe true change over time for person \( p \). The elements of the error vector, on the other hand, describe the within-person noise: the forces that disturb measurement of person \( p \)'s

Figure 1. A collection of ordinary least squares (OLS) fitted trajectories summarizing observed linear growth in (log) tolerance for deviant behavior between ages 11 and 15 for the subsample of 16 randomly selected youths whose empirical growth records are displayed in Table 1.
true change over time. If the latter are large and erratic, one may not be able to detect the former.

**Distribution of the Measurement Errors**

In Equations 1 and 2, we stated that error $\epsilon_{t,p}$ disturbs the true status of the $p$th person on the first occasion of measurement, $\epsilon_{2p}$ does so on the second occasion, and so forth. However, we have made no claims about the shape of the error distribution; the errors may be homoscedastic and independent over time within individuals, they may be heteroscedastic, or they may be autocorrelated. We usually begin by assuming that the errors of measurement obey stringent “classical” assumptions; that is, they are distributed independently and homoscedastically over time with mean zero and homogeneous variance $\sigma^2$. In other words, person $p$ draws his or her measurement error vector from the following distribution:

$$
\begin{pmatrix}
\epsilon_{1p} \\
\epsilon_{2p} \\
\epsilon_{3p} \\
\epsilon_{4p} \\
\epsilon_{5p}
\end{pmatrix} \sim N
\begin{pmatrix}
0 \\
0 \\
0 \\
0
\end{pmatrix},

\begin{pmatrix}
\sigma^2 & 0 & 0 & 0 & 0 \\
0 & \sigma^2 & 0 & 0 & 0 \\
0 & 0 & \sigma^2 & 0 & 0 \\
0 & 0 & 0 & \sigma^2 & 0 \\
0 & 0 & 0 & 0 & \sigma^2
\end{pmatrix}.
$$

(3)

where the mean vector and covariance matrix on the right-hand side are assumed to be identical across people.

A major advantage of the covariance structure approach described here lies in the flexibility with which the error covariance structure can be modeled. Under the method, we assume that each person draws his or her measurement error vector at random from a distribution with mean vector zero and an unknown covariance matrix whose shape can be specified as necessary. This permits us to test the fit of the classical error structure against other, more liberal hypotheses, and we can modify the error covariance structure as necessary. And regardless of the final structure adopted, we can estimate the components of the hypothesized error covariance matrix. This facility is important in a study of change because knowledge of the magnitudes of the error variances and covariances underpins the estimation of measurement reliability and error autocorrelation. We provide examples of this later.

**Modeling Interindividual Differences in Change**

Even though all people share a common functional form for their change, the true growth trajectories may still differ across people because of interindividual variation in the values of the individual growth parameters. Thus, formerly ill-specified questions about vaguely defined interindividual differences in change can be reframed as specific questions about the distribution of individual growth parameters across people in the population. In our data example (in which we believe that straight-line growth is occurring), for instance, we can ask questions about (a) the population mean vector of the individual growth parameters (e.g., Across all members of the population, what is the average value of the true intercept? Of the true slope?) and (b) the population covariance matrix of the individual growth parameters (e.g., Across all members of the population, what is the variance of the true intercept? Of the true slope? What is the population covariance of the true intercept and slope?).

Thus, when one studies interindividual differences in change, one expresses an interest in the population between-person distribution of the individual growth parameters. In our data example, for instance, we assume that each person in the population draws his or her latent growth vector independently from a multivariate normal distribution of the following form:

$$
\begin{pmatrix}
\pi_{0p} \\
\pi_{1p}
\end{pmatrix} \sim N
\begin{pmatrix}
\mu_{\pi} \\
\mu_{\pi}
\end{pmatrix},

\begin{pmatrix}
\sigma_{\pi 0}^2 & \sigma_{\pi 10} \\
\sigma_{\pi 10} & \sigma_{\pi 11}
\end{pmatrix}.
$$

(4)

The hypothesized distribution in Equation 4 is our first between-person, or Level 2, model for interindividual differences in true change. Later we extend this model by introducing predictors of change into the formulation. Even in this simple model, however, there are several between-person parameters worthy of estimation: the population means, variances, and covariance on the right-hand side of Equation 4. These parameters provide baseline information on the average trajectory of true change, the variation of true intercept and slope, and the covariation of true intercept and slope in the population, thereby answering the preliminary questions cited earlier in this subsection.

**Adopting a Covariance-Structure Perspective**

In Table 2, we present the sample mean vector and covariance matrix for the variables introduced in Table 1, estimated with data on all 168 adolescents in the illustrative data set. What statements about change over time do these statistics readily support? Focus, first, on the statistics describing the five waves of observed log tolerance of deviance. Examining the wave-by-wave means shows that observed (log) tolerance tends to increase steadily, on average, over the 5-year period of observation in the sample as a whole. In addition, the magnitudes of the variances along the leading diagonal of the sample covariance matrix suggest that observed log tolerance of deviance becomes more variable over time, perhaps as adolescents' scores "fan out" with age. Finally, inspection of the sample between-wave covariances suggests a generally positive association among observed log tolerance of deviance over the five occasions of measurement.

However, even ignoring the distinction between observed and true scores, it is not easy to reach informed conclusions about interindividual differences in change by inspecting between-wave summary statistics (Rogosa et al., 1982; Rogosa & Willett, 1985; Willett, 1989). Between-wave statistics do not provide an optimal view for easy inference about differences in individual change. To answer questions about change, one must adopt a perspective that emphasizes change. Rather than summarizing data as between-wave variances and covariances, one must use individual growth trajectories. For instance, it is easier to see from Figure 1 that observed change is generally positive, individuals are fanning out over time, and there is heterogeneity in level and rate of change across people. The data are identical in both cases, but the view offered by the summary statistics differs; each view supports a qualitatively different kind of interpretation.

*Provided the hypothesized covariance structure model is identified.*
Table 2
Sample Mean Vector and Covariance Matrix for the Five Waves of Log Tolerance
of Deviant Behavior and Two Predictors of Change

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Age 11</th>
<th>Age 12</th>
<th>Age 13</th>
<th>Age 14</th>
<th>Age 15</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean vector</td>
<td>0.2008</td>
<td>0.2263</td>
<td>0.3255</td>
<td>0.4168</td>
<td>0.4460</td>
</tr>
<tr>
<td>Covariance matrix</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age 11</td>
<td>0.0317</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age 12</td>
<td>0.0133</td>
<td>0.0395</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age 13</td>
<td>0.0175</td>
<td>0.0256</td>
<td>0.0724</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age 14</td>
<td>0.0213</td>
<td>0.0236</td>
<td>0.0531</td>
<td>0.0857</td>
<td></td>
</tr>
<tr>
<td>Age 15</td>
<td>0.0230</td>
<td>0.0233</td>
<td>0.0479</td>
<td>0.0663</td>
<td>0.0873</td>
</tr>
<tr>
<td>Gender</td>
<td>-0.0108</td>
<td>-0.0143</td>
<td>-0.0225</td>
<td>-0.0299</td>
<td>-0.0270</td>
</tr>
<tr>
<td>Log exposure</td>
<td>0.0115</td>
<td>0.0133</td>
<td>0.0089</td>
<td>0.0091</td>
<td>0.0186</td>
</tr>
<tr>
<td>Log exposure</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note.  N = 168.

Does this mean that one cannot recover information about change once data have been collapsed into between-wave means and covariances? No, it does not. One must simply “match up” the between-wave and change perspectives explicitly. If we could, for instance, determine the between-wave implications of the individual growth modeling perspective adopted in Equations 1-4, we could check whether they compared favorably with the data summaries in Table 2. For instance, in Equations 4 and 3, we proposed what we believe are reasonable models for interindividual variation in the individual growth parameters and errors of measurement. If we are correct, then these models must underwrite the between-wave mean and covariance structure evident in Table 2. In other words, although we are dealing with two different perspectives on the problem—a between-wave perspective in Table 2 and a growth perspective in Equations 1-4—the between-wave covariance structure implied by the growth models must resemble the between-wave covariance structure observed in our data if our parameterization of change is correct.

Fortunately, well-developed methods are available for testing our suspicions: the methods of covariance structure analysis. Starting with the sample mean vector and covariance matrix in Table 2 as input we can claim that our hypothesized growth models fit when, having estimated the parameters of Equations 3 and 4, we can accurately predict the between-wave covariance structure of the observed data. As Meredith, Tisak, McArdle, and Muthén have pointed out, the growth formulation that we have posited—the within-person models of Equations 2 and 3 and the between-person model of Equation 4—falls naturally into the framework offered by the LISREL model with mean structures (Jöreskog & Sörbom, 1989). Thus, maximum likelihood estimates of the important parameters in Equations 3 and 4 can be obtained by covariance structure analysis, as we now demonstrate.

Rewriting the Individual Growth Model as the LISREL Measurement Model for Y

When covariance structure analysis is used to examine change over time, the hypothesized individual growth model plays the role of the LISREL measurement model for the vector of endogenous variables Y. For instance, we can rewrite the empirical growth record of the pth person in our illustrative example as

\[
\begin{bmatrix}
Y_{1p} \\
Y_{2p} \\
Y_{3p} \\
Y_{4p} \\
Y_{5p}
\end{bmatrix}
= \begin{bmatrix}
0 \\
1 \\
1 \\
1 \\
1
\end{bmatrix}
\begin{bmatrix}
\xi_{1p} \\
\xi_{2p} \\
\xi_{3p} \\
\xi_{4p} \\
\xi_{5p}
\end{bmatrix}
+ \begin{bmatrix}
\eta_{1p} \\
\eta_{2p} \\
\eta_{3p} \\
\eta_{4p} \\
\eta_{5p}
\end{bmatrix}
\]

(5)

which has the format of the LISREL measurement model for endogenous variables Y:

\[
Y = \tau + \Lambda \eta + \epsilon,
\]

(6)

with LISREL score vectors that contain the empirical growth record, the individual growth parameters, and the errors of measurement, respectively:

\[
Y = \begin{bmatrix}
Y_{1p} \\
Y_{2p} \\
Y_{3p} \\
Y_{4p} \\
Y_{5p}
\end{bmatrix}, \quad \eta = \begin{bmatrix}
\xi_{1p} \\
\xi_{2p} \\
\xi_{3p} \\
\xi_{4p} \\
\xi_{5p}
\end{bmatrix}, \quad \epsilon = \begin{bmatrix}
\eta_{1p} \\
\eta_{2p} \\
\eta_{3p} \\
\eta_{4p} \\
\eta_{5p}
\end{bmatrix}
\]

(7)

Furthermore, unlike the usual practice of covariance structure analysis, the elements of the LISREL \(\tau\) and \(\Lambda\) parameter matrices are entirely constrained to contain only known values,

\[
\tau = \begin{bmatrix}
0 \\
1 \\
1 \\
1 \\
0
\end{bmatrix}, \quad \Lambda = \begin{bmatrix}
1 & t_1 \\
0 & 1 & t_2 \\
0 & 0 & 1 & t_3 \\
0 & 0 & 0 & 1 & t_4
\end{bmatrix}
\]

(8)

and the error vector \(\epsilon\) is distributed with zero mean vector and covariance matrix \(\Theta_\epsilon\), which, under the classical assumptions of Equation 3, is given by

\[
\Theta_\epsilon = \text{Cov}(\epsilon) = \begin{bmatrix}
\sigma_{\epsilon_1}^2 & 0 & 0 & 0 & 0 \\
0 & \sigma_{\epsilon_2}^2 & 0 & 0 & 0 \\
0 & 0 & \sigma_{\epsilon_3}^2 & 0 & 0 \\
0 & 0 & 0 & \sigma_{\epsilon_4}^2 & 0 \\
0 & 0 & 0 & 0 & \sigma_{\epsilon_5}^2
\end{bmatrix}
\]

(9)
Rewriting the Model for Interindividual Differences in Change as the LISREL Structural Model

Note that, unlike in more familiar standard covariance structure analyses, we have chosen to specify the entire LISREL \( \Lambda_p \) parameter matrix in Equation 8 as a matrix of known times and constants rather than as a collection of unknown parameters to be estimated. This specification acts to "pass" the critical Level 1 individual growth parameters (\( \pi_{0p} \) and \( \pi_{1p} \)) from the Level 1 growth model into the LISREL endogenous construct vector \( \eta \), which we have then referred to as the latent growth vector. In other words, our fully constrained specification for \( \Lambda_p \) has forced the \( \eta \) vector to contain the very individual-level parameters whose Level 2 distribution is to become the focus of our subsequent between-person analyses.

These Level 2 analyses are conducted within the structural component of the general LISREL model, which permits the distribution of the \( \eta \) vector to be modeled explicitly in terms of selected population means, variances, and covariances. And, of course, the particular population means, variances, and covariances that we have selected as parameters of the structural model are those that we have hypothesized are the important parameters in the joint distribution of the individual growth parameters in Equation 4. All that is required is to write the latent growth vector as

\[
\begin{bmatrix}
\pi_{0p} \\
\pi_{1p}
\end{bmatrix}
= \begin{bmatrix}
\mu_{0p} & 0 \\
0 & \mu_{1p}
\end{bmatrix}
\begin{bmatrix}
\pi_{0p} \\
\pi_{1p}
\end{bmatrix}
+ \begin{bmatrix}
\pi_{0p} - \mu_{0p} \\
\pi_{1p} - \mu_{1p}
\end{bmatrix},
\]

(10)

which has the form of the reduced LISREL structural model

\[
\eta = \alpha + \beta \eta + \xi
\]

(11)

with a latent residual vector \( \xi \) that contains the deviations of the individual parameters from their respective population means,

\[
\xi = \begin{bmatrix}
\pi_{0p} - \mu_{0p} \\
\pi_{1p} - \mu_{1p}
\end{bmatrix},
\]

(12)

and parameter matrices,

\[
\alpha = \begin{bmatrix}
\mu_{0p} \\
\mu_{1p}
\end{bmatrix}, \quad \beta = \begin{bmatrix}
0 & 0 \\
0 & 0
\end{bmatrix}.
\]

(13)

Note that we have moved the population averages of the individual growth parameters—true intercept and slope, in this case—into the LISREL \( \alpha \) vector. This permits these important mean parameters to be estimated explicitly. The elements of the LISREL latent residual vector, \( \xi \), in Equations 11 and 12 contain deviations of \( \pi_{0p} \) and \( \pi_{1p} \) from their respective population means. The \( \xi \) vector of latent residuals is of special interest here because it is distributed with zero mean vector and covariance matrix \( \Psi \), the matrix containing the very variance and covariance parameters in which one is most interested in an investigation of interindividual differences in change:

\[
\Psi = \text{Cov}(\xi) = \begin{bmatrix}
\sigma_{\pi_{0p}} & \sigma_{\pi_{0p},\pi_{1p}} \\
\sigma_{\pi_{0p},\pi_{1p}} & \sigma_{\pi_{1p}}
\end{bmatrix}.
\]

(14)

To summarize, in Equations 1-4, the individual growth modeling framework provides baseline Level 1 (within-person) and Level 2 (between-person) models that represent our initial hypotheses about the growth structure underlying the five waves of panel data in our data example. Then, in Equations 5-14, we have shown that these models can be rewritten, without loss of generality, in the format and notation of the general LISREL model with mean structures. By carefully choosing our specification of the various standard LISREL parameter matrices, we have forced the LISREL \( \lambda \)-measurement model to become our original Level 1 individual growth model (including all existing assumptions on the distribution of the measurement errors), and we have forced a reduced form of the LISREL structural model to become our Level 2 model for interindividual differences in true change.

Because of this direct and explicit mapping of one model into the other, we can test whether our hypothesized growth formulation underpins the matrix of observed between-wave variances and covariances in Table 2 using the LISREL program. If the implied covariance structure fits the data, then we also obtain, and can interpret, LISREL-provided maximum likelihood estimates of the unknown parameters in our growth models that now reside in the \( \alpha \) vector, the \( \Theta \) matrix, and the \( \Psi \) matrix. In the Appendix, we provide an annotated LISREL program for this analysis (Model 1). The program specifies the \( r, \lambda, \theta, \alpha, B, \) and \( \Psi \) parameter matrices as defined in Equations 8, 9, 13, and 14. All hypothesized zero entries in these parameter matrices are fixed at zero in the program; the values of measurement times \( t_i \) through \( t_5 \) are set to \(-2, -1, 0, 1, \) and \( 2 \) in accordance with our earlier centering decision; and all unknown parameters are free to be estimated.

This baseline "no predictors of change" model fits moderately well (see Bollen, 1989, pp. 256–289, for a discussion of the use of summary statistics in model evaluation). Although the model chi-square statistic (49.74) is slightly large, given its degrees of freedom (14), the values of other goodness-of-fit indices are heartening: Both LISREL goodness-of-fit statistics are greater than .9 (goodness of fit index \( \text{GFI} = .918 \), adjusted goodness of fit index \( \text{AGFI} = .912 \)), and the root mean-square residual (RMSR) is small relative to the absolute magnitude of the elements of the sample covariance matrix in Table 2 (RMSR = .008). Maximum likelihood estimates of the unknown parameters are listed under Model 1 in Table 3, along with approximate \( p \) values.

The entries in the first two rows of Table 3 for Model 1 estimate the population means of true intercept (.3231, \( p < .001 \)) and true slope (.0681, \( p < .001 \)) and describe the average trajectory of true change in the dependent variable. On average, ado-
Table 3
Fitted Models Demonstrating Interindividual Differences in Change in Log Tolerance in the Full Sample

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \mu_{\theta} )</td>
<td>0.3231***</td>
<td>0.3235***</td>
<td>0.3197***</td>
</tr>
<tr>
<td>( \mu_{r} )</td>
<td>0.0681***</td>
<td>0.0664***</td>
<td>0.0643***</td>
</tr>
<tr>
<td>( \sigma_{\theta}^{2} )</td>
<td>0.0328***</td>
<td>0.0326***</td>
<td>0.0295***</td>
</tr>
<tr>
<td>( \sigma_{r}^{2} )</td>
<td>0.0026***</td>
<td>0.0029***</td>
<td>0.0023***</td>
</tr>
<tr>
<td>( \sigma_{\theta r}^{2} )</td>
<td>0.0081***</td>
<td>0.0079***</td>
<td>0.0070***</td>
</tr>
<tr>
<td>( \sigma_{\theta} )</td>
<td>0.0254***</td>
<td>0.0186***</td>
<td>0.0195***</td>
</tr>
<tr>
<td>( \sigma_{r} )</td>
<td>0.0254***</td>
<td>0.0269***</td>
<td>0.0275***</td>
</tr>
<tr>
<td>( \alpha_{\theta} )</td>
<td>0.0254***</td>
<td>0.0340***</td>
<td>0.0382***</td>
</tr>
<tr>
<td>( \alpha_{r} )</td>
<td>0.0254***</td>
<td>0.0242***</td>
<td>0.0350***</td>
</tr>
<tr>
<td>( \alpha_{\theta r} )</td>
<td>0.0254***</td>
<td>0.0178***</td>
<td>0.0212***</td>
</tr>
<tr>
<td>( \alpha_{rr} )</td>
<td>0.00010</td>
<td>0.0052†</td>
<td>0.00113***</td>
</tr>
<tr>
<td>( \alpha_{\theta \theta} )</td>
<td>0.00084</td>
<td>0.00084</td>
<td>0.00084</td>
</tr>
<tr>
<td>( \chi^2 )</td>
<td>49.74</td>
<td>39.82</td>
<td>20.83</td>
</tr>
<tr>
<td>( df )</td>
<td>14</td>
<td>10</td>
<td>6</td>
</tr>
<tr>
<td>Goodness of fit index (GFI)</td>
<td>0.918</td>
<td>0.941</td>
<td>0.981</td>
</tr>
<tr>
<td>Adjusted goodness of fit index (AGFI)</td>
<td>0.912</td>
<td>0.912</td>
<td>0.951</td>
</tr>
<tr>
<td>Root mean-square residual (RMSR)</td>
<td>0.008</td>
<td>0.006</td>
<td>0.004</td>
</tr>
</tbody>
</table>

Note. \( N = 168 \). See text for descriptions of models.
† \( p < .10 \). * \( p < .05 \). ** \( p < .01 \). *** \( p < .001 \) (approximate values).

lescents’ true log tolerance of deviance increases by .0681 per year, having a value of .3231 in Year 3 (13 years of age).

The entries in the third and fourth rows for Model 1—the fitted variances of true intercept (.0328, \( p < .001 \)) and true slope (.0026, \( p < .001 \))—estimate population interindividual heterogeneity in true change. Because both of these variances are non-zero, we know that between-person heterogeneity in the age 13 level and rate of change in true log tolerance exist in the population. The fifth table entry under Model 1 provides an estimate of the covariance of the true intercept and true rate of change in log tolerance of deviance across people in the population (.0081, \( p < .001 \)). Combining this covariance estimate with the estimated variances suggests that the true rate of change has a correlation of .887 with the true intercept; this large value indicates that participants with higher age 13 levels of true log tolerance also have the most rapid rates of increase.

Entries in the 6th through 10th rows for Model 1 provide an estimate of the homoscedastic measurement error variance (.0254, \( p < .001 \)), which—together with the variances of observed score along the leading diagonal of the sample covariance matrix in Table 2—indicates that, except for the first and second panels, the within-wave reliability of measurement was moderate (.198, .357, .649, .703, and .709, respectively, in Years 1 through 5). Finally, substituting the estimated variances of measurement error and true slope into Equation 29 of Willett (1989), we find that the rate of true change in log tolerance has been measured with a reliability of .502. This is higher than usually anticipated in the measurement of change and higher than the reliability with which log tolerance itself was measured on two of the occasions. It illustrates that, if sufficient waves of longitudinal data are collected, the measurement of change can be more reliable than the measurement of status on a single occasion despite the prognostications of Bereiter (1963), Lord (1963), and others.

Testing for Evidence of Nonlinear Individual Change

So far, we have hypothesized that individual change over time in log tolerance is linear. However, in practice, individual change may be curvilinear (e.g., see the research on vocabulary growth in children conducted by Huttenlocher, Haight, Bryk, & Seltzer, 1991).

The covariance structure approach that we have described can easily be modified to include any Level I growth model that is linear in the individual growth parameters. Keats (1983) defined such models as having the property of dynamic consistency; for such models, the curve of the averages (obtained by taking the population of true growth curves and plotting the average of the true scores at each value of time) is identical to the average of the curves (obtained by averaging the individual growth parameters over the population and plotting a curve with parameters equal to these averages). Many common growth functions, including segmented growth curves and polynomial growth of any order, are dynamically consistent. Other, more specialized growth functions (such as the Gompertz, Jennis, and logistic functions) are not; therefore, the character of the individual curves is distorted by group averaging, making it difficult to infer the shape of individual growth from a group growth curve (Boas, 1892; Estes, 1956).

If true individual change is hypothesized to be a quadratic function of time, for instance, the observed status of the \( p \)th person at time \( t \) can be represented by

\[
Y_{ip} = \pi_{0p} + \pi_{1p}t + \pi_{2p}t^2 + \epsilon_{ip},
\]

where the presence of the quadratic parameter \( \pi_{2p} \) permits the trajectory to be curvilinear. If \( \pi_{2p} \) is negative, the trajectory is concave to the time axis; if it is positive, the trajectory is convex to the time axis. Because time was centered on the third occasion of measurement, \( \pi_{0p} \) represents the \( p \)th person’s true log tolerance of deviance at age 13, and \( \pi_{1p} \) is the instantaneous rate of true change in log tolerance of deviance at age 13.

Again, following Equation 5, we can write the empirical growth record for the \( p \)th person as the product of a matrix of known times and constants and an individual-specific latent growth vector containing the unknown individual growth parameters, added to a measurement error vector. As before, this provides the LISREL measurement model for the endogenous variables \( Y \), but now the quadratic individual growth parameter is forced into the latent growth vector \( \eta \) along with true intercept and slope by suitable redefinition of the earlier \( \Lambda \) parameter matrix, as follows:

\[
\begin{bmatrix}
Y_{1p} \\
Y_{2p} \\
Y_{3p} \\
Y_{4p} \\
Y_{5p}
\end{bmatrix} = \begin{bmatrix}
0 \\
0 \\
0 \\
0 \\
0
\end{bmatrix} + \begin{bmatrix}
1 & t_1 & t_1^2 \\
1 & t_2 & t_2^2 \\
1 & t_3 & t_3^2 \\
1 & t_4 & t_4^2 \\
1 & t_5 & t_5^2
\end{bmatrix} \begin{bmatrix}
\pi_{0p} \\
\pi_{1p} \\
\pi_{2p} \\
\epsilon_{0p} \\
\epsilon_{2p}
\end{bmatrix} + \begin{bmatrix}
\epsilon_{1p} \\
\epsilon_{2p} \\
\epsilon_{3p} \\
\epsilon_{4p} \\
\epsilon_{5p}
\end{bmatrix}.
\]
This matrix again has the format of the LISREL $Y$-measurement model in Equation 6 with constituent score vectors,

\[ Y = \begin{bmatrix} Y_{1p} \\ Y_{2p} \\ Y_{3p} \\ Y_{4p} \\ Y_{5p} \end{bmatrix}, \quad \eta = \begin{bmatrix} \pi_{0p} \\ \pi_{1p} \\ \pi_{2p} \end{bmatrix}, \quad \epsilon = \begin{bmatrix} \epsilon_{1p} \\ \epsilon_{2p} \\ \epsilon_{3p} \\ \epsilon_{4p} \\ \epsilon_{5p} \end{bmatrix}, \quad (17) \]

and parameter matrices,

\[ \tau_p = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \Lambda_p = \begin{bmatrix} 1 & t_1 & t_1^2 \\ 1 & t_2 & t_2^2 \\ 1 & t_3 & t_3^2 \\ 1 & t_4 & t_4^2 \\ 1 & t_5 & t_5^2 \end{bmatrix}, \quad (18) \]

\[ \Theta = \begin{bmatrix} \sigma_0^2 & 0 & 0 & 0 & 0 \\ 0 & \sigma_1^2 & 0 & 0 & 0 \\ 0 & 0 & \sigma_2^2 & 0 & 0 \\ 0 & 0 & 0 & \sigma_0^2 & 0 \\ 0 & 0 & 0 & 0 & \sigma_2^2 \end{bmatrix}, \quad (19) \]

where the times $t_1$ through $t_5$ have known numeric value (as before).

Although all members of the population share a common quadratic growth function, their true change trajectories may differ when the values of the individual growth parameters differ from person to person. Thus, when studying interindividual differences in quadratic change, one remains interested in the population distribution of the latent growth vector, which can again be modeled within the LISREL structural model:

\[ \begin{bmatrix} \pi_{0p} \\ \pi_{1p} \\ \pi_{2p} \end{bmatrix} = \begin{bmatrix} \mu_{0p} \\ \mu_{1p} \\ \mu_{2p} \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \pi_{0p} + \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \pi_{1p} + \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \pi_{2p} = \begin{bmatrix} \mu_{0p} \\ \mu_{1p} \\ \mu_{2p} \end{bmatrix} + \begin{bmatrix} \pi_{0p} \\ \pi_{1p} \\ \pi_{2p} \end{bmatrix}, \quad (20) \]

Again, this model has the form of the reduced structural model in Equation 11 but with an expanded definition of the LISREL latent residual vector,

\[ \xi = \begin{bmatrix} \pi_{0p} - \mu_{0p} \\ \pi_{1p} - \mu_{1p} \\ \pi_{2p} - \mu_{2p} \end{bmatrix}, \quad (21) \]

This expanded model for interindividual differences in true quadratic change (Equations 15–21) can be fitted by straightforward modification of the LISREL program for Model 1 in the Appendix. The respective fits of the linear and quadratic formulations can then be compared directly because the linear formulation is nested within the quadratic formulation (setting $\pi_{2p}$ to zero in Equations 15–21 leads to Equations 5–14). Specifically, the null hypothesis that the addition of a quadratic term to the linear individual growth model does not improve the fit of the simpler representation can be evaluated by performing a standard "decrement-to-chi-square" test in which the respective goodness of fits (and degrees of freedom) of the linear and quadratic representations are differentiated. In our data example, this test indicates that the addition of the quadratic parameter does not improve global goodness of fit (change in $\chi^2 = 49.74 - 44.18 = 5.56$, change in $df = 14 - 10 = 4$). We conclude, therefore, that the straight-line individual growth model in Equation 1 is an appropriate representation of change over time in this particular case.

Note, however, that we refer to this comparison of the two nested models as a global test. Its purpose is analogous to the usual inspection of the global $F$ statistic associated with a complex main effect in an ANOVA before conducting multiple comparisons or follow-up contrast analyses. With the addition of the quadratic term to the individual straight-line growth model, four new parameters have been added to the Level 2 model: $\mu_{0p}$, the population mean of the quadratic parameter; $\sigma_{0p}^2$, the population variance of $\pi_{2p}$; and $\sigma_{0x0}$ and $\sigma_{0x1}$, the population covariances of $\pi_{2p}$ with $\pi_{0p}$ and $\pi_{1p}$. Therefore, a comparison of the fits of the linea and quadratic formulations is actually a simultaneous test of the jointly null values of these Level 2 quadratic mean, variance, and covariance parameters, holding Type I error at some manageable level (e.g., an alpha level of .05). Broadly speaking, the global test evaluates whether there is any interest in including a quadratic term in the individual growth model; in our case, there is no such interest.\(^8\)

Of course, the flexibility of the covariance structure approach permits a sequence of more subtle and restrictive hypotheses to be tested. If we had rejected the null hypothesis under the global test, for instance, we would have been entitled to follow up with tests of subhypotheses in which the four Level 2 parameters were constrained to be zero, either singly or in interesting combinations. For example, we could test whether everyone does, in fact, experience quadratic growth but with a curvature identical for all, that is, the case in which the value of the quadratic growth parameter $\pi_{2p}$ is fixed at its population average for all individuals (i.e., $\pi_{2p} = \mu_{2p}$ for all $p$). This hypothesis is easily tested by comparing the fit of the full quadratic model, as described earlier, with that of a reduced quadratic model in which the population average of the curvature term $\mu_{2p}$ is estimated but the population variance of $\pi_{2p}$ and its covariances with true intercept and slope are constrained to be zero (in other words, everyone is hypothesized to possess an identical nonzero curvature coefficient equal to $\mu_{2p}$). On the other hand, if we wished to test whether the population average of the quadratic growth parameter was, in fact, zero but that individual quadratic terms

\(^8\) A reviewer pointed out that the results of a global test may be misleading if the variance components associated with the quadratic parameter are small. The power to detect a nonzero mean coefficient may be lower for the global test than for a specific test of the quadratic mean parameter. Therefore, the prudent strategy is always to conduct separate tests of each hypothesis in lieu of the omnibus test.
were distributed (randomly) about this value, we would com-
parison of nested models within the general LISREL framework
provide a very flexible strategy for testing the effectiveness of
reasoned modifications to the basic measurement (within-per-
son) and structural (between-person) models.

Testing Whether Measurement Errors Are
Heteroscedastic and Nonindependent

Investigators often assume that measurement errors are ho-
moscedastic and independent within individuals over time, as
we have done so far in this article. Such assumptions are com-
mon in the psychometric literature, in which they are central to
classical test theory, and in the growth literature (e.g., see Ber-
key, 1982a, 1982b). However, with consecutive measurements
on an individual changing over time, these assumptions may be
tenable. There is no reason to believe a priori that the
precision with which an attribute can be measured is identical
at all ages, and so the measurement errors may be heteroscedas-
tic. And when measurements are closely spaced in time, there
may be inadvertent links among their errors.

The covariance structure approach affords great flexibility in
modeling the error covariance structure; one can easily relax
the stringent assumptions of homoscedasticity and zero auto-
correlation on the measurement errors. At present, this facility
is not available under the HLM approach of Bryk and Rauden-
bush (1992); neither has it been demonstrated in other applica-
tions of covariance structure analysis to the measurement of
individual change (McArdle & Epstein, 1987; Muthén, 1991).

When distributional assumptions on the measurement errors
are relaxed, the basic covariance structure approach is un-
changed. The error covariance structure associated with the
Level 1 (Y-measurement) model is simply modified to contain
whatever parameters permit the hypothesized heteroscedastic-
ity or autocorrelation among the errors. The Level 2 (structural)
model for the distribution of the latent growth vector remains
the same. The modified model can then be fit by means of the
approach already described once the LISREL Θ matrix—in
which the newly hypothesized error structure resides—has been
reparameterized appropriately, with suitable constraints on the
equality of parameters being imposed or omitted as required.

For instance, one can hypothesize that the measurement errors
are independent but heteroscedastic. Then, Θ in Equation
9 becomes

$$
\Theta = \begin{bmatrix}
\sigma_1^2 & 0 & 0 & 0 \\
0 & \sigma_2^2 & 0 & 0 \\
0 & 0 & \sigma_3^2 & 0 \\
0 & 0 & 0 & \sigma_4^2 \\
0 & 0 & 0 & 0 & \sigma_5^2
\end{bmatrix}
$$

A program suitable for fitting the modified model is provided in
the Appendix. When the new linear-change heteroscedastic-
error model is fitted (Model 2 in Table 3), there is a statisti-
cally significant improvement in fit over Model 1 (change in χ^2
= 49.74 - 39.82 = 9.92, change in df = 14 - 10 = 4). Now, rather
than remaining constant at .0254 on all occasions, the estimated
measurement error variance is lower on the first and last oc-
casions of measurement and peaks at .0340 when the adolescent
is 13 years old. This, of course, modifies our earlier yearly esti-
mates of the reliability of the dependent variable, which now
becomes .413,.320,.531,.718, and .797, respectively. Other pa-
rameter estimates remain relatively stable from those obtained
when Model 1 was fitted.

We can also relax the assumption of independence across
time that we imposed on the wave-by-wave measurement errors
by further specifying the Θ matrix. The LISREL program
permits considerable flexibility in this regard. For instance, we
can retain the heteroscedasticity that we detected earlier but
also allow temporally adjacent pairs of measurement errors to
be mutually autocorrelated:

$$
\Theta_i = \begin{bmatrix}
0 & \sigma_{i12} & 0 & 0 & 0 \\
\sigma_{i12} & 0 & \sigma_{i23} & 0 & 0 \\
0 & \sigma_{i23} & 0 & \sigma_{i34} & 0 \\
0 & 0 & \sigma_{i34} & 0 & \sigma_{i45} \\
0 & 0 & 0 & \sigma_{i45} & 0
\end{bmatrix}
$$

Model 3, in Table 3 (see the Appendix for the associated
LISREL program), contains this pairwise autocorrelated error
covariance structure. It improves successfully on the fit of
Model 2 (see Table 3; change in χ^2 = 18.99, change in df =
4), with estimated autocorrelations between adjacent pairs of
measurement errors of .044,.159,.309, and .307, respectively.
Note that, despite major changes in the hypothesized links
among the measurement errors across the models in Table 3,
the shape of the average trend line, the heterogeneity in linear
growth, and the correlation between intercept and slope remain
relatively stable. We use Model 3 as the foundation for sub-
sequent investigations of systematic interindividual differences in
change.

A Comment on the Requirement for
Time-Structured Data

As we noted earlier, covariance structure methods for the
analysis of change require that all sample members be observed
on the same set of occasions. If the available data are not time
structured, then the method cannot be used. This requirement
is a potentially serious limitation not shared by other analytic
approaches, such as those of Bryk and Raudenbush (1987) and

In longitudinal research, investigators often intend at the out-
set to collect time-structured data; however, reality intervenes,
and observations on some individuals on some occasions are
omitted as a result of forces beyond the researcher’s control.
This introduces missing values into the data set and destroys
the intended time structuring, leading to difficulties in the esti-
mation of the sample covariance matrix on which subsequent
analyses are based.

Bollen (1989) described two general approaches for dealing
with missing data in covariance structure analysis. The first uses
either listwise or pairwise deletion of incomplete cases, or the
imputation of missing data points, to construct an alternative
estimator of the sample covariance matrix. Estimators obtained
in the reduced sample created by listwise deletion retain the
important property of consistency when data are missing completely at random. The second approach uses all available data by treating individuals with similar patterns of missing data as subgroups of the original sample so that multigroup analysis (Jöreskog & Sörbom, 1989) can be used to obtain estimates of model parameters, which are again consistent providing that data are missing completely at random. Both approaches are potentially applicable in the investigation of change, although the second may be preferable to the first. Muthén (1992) has described ways of applying this latter approach to the covariance structure analysis of incomplete longitudinal data. Recent advances in the handling of missing data may also lead to further innovations in this area (Graham, Hofer, & Piccinin, in press; Little & Rubin, 1987).

Of course, great caution must be exercised if there are individuals in the data set with incomplete or missing data. Data may not be missing completely at random; people with particular values of the predictors may have experienced unique patterns of change over time. Any such systematicity in the pattern of missing values will necessarily undermine the investigator's ability to make inferences back to the population as originally defined and may lead to bias in the interpretation of the findings. For this reason, we recommend that preliminary exploratory data analysis be conducted to examine whether cases with missing or incomplete data are atypical so that any interpretation of study findings can be circumscribed appropriately.

Modeling Systematic Interindividual Differences in Change

A baseline investigation of the no predictor of change model informs the subsequent analysis. If baseline analyses confirm the existence of heterogeneity in true change in the population, then we can ask whether this heterogeneity is related to selected characteristics of the people being observed. In other words, once interindividual differences in true change have been detected, we can ask about the systematic nature of that variation. And, because we have distinguished among the true change of different people in terms of their individual growth parameters, questions about systematic heterogeneity in true change naturally translate into questions about relationships between the individual growth parameters and predictors (Rogosa & Willett, 1985).

In our data example, we know that heterogeneity in true change exists in the population (Table 3), and we have two potential predictors of change to investigate: adolescent gender ($G_p$) and initial (log) exposure to deviant behavior at age 11 ($E_p$). Therefore, we can ask whether the individual growth parameters—intercept and slope—are related to this pair of predictors: Does the age 13 level of true (log) tolerance of deviance differ for boys and girls? Does it differ by initial exposure to deviance at age 11? Does the rate at which true (log) tolerance changes over time also depend on gender and exposure? The methods described here generalize immediately to the case of more than two predictors.

As we stated earlier, such questions are concerned with between-person differences in change. To answer them, we must incorporate potential predictors of change into the Level 2 model (i.e., into the LISREL structural model in which interindividual differences in change are described). Within the general LISREL framework, predictors can be inserted into the structural model indirectly by taking advantage of the LISREL measurement model for exogenous predictors, $X$. This last component of the general LISREL model permits us to pass predictors of change into the LISREL vector of exogenous constructs, $\xi$, which is a natural and so far unused constituent of the LISREL structural model.

In the current case, in which we have single indicators of each predictor, we set up the $X$-measurement model in the following way:

$$
\begin{bmatrix}
G_p \\
E_p
\end{bmatrix}
= \begin{bmatrix}
\mu_G \\
\mu_E
\end{bmatrix}
+ \begin{bmatrix}
1 & 0 \\
0 & 1
\end{bmatrix}
\begin{bmatrix}
G_p - \mu_G \\
E_p - \mu_E
\end{bmatrix}
+ \begin{bmatrix}
0 \\
0
\end{bmatrix}.
$$

(24)

This model has the format of the LISREL measurement model for exogenous variables $X$:

$$
X = \tau_x + \Lambda_x \xi + \delta.
$$

(25)

With constituent predictor, latent exogenous score, and error vectors,

$$
X = \begin{bmatrix}
G_p \\
E_p
\end{bmatrix},
\xi = \begin{bmatrix}
G_p - \mu_G \\
E_p - \mu_E
\end{bmatrix},
\delta = \begin{bmatrix}
0 \\
0
\end{bmatrix},
$$

(26)

and constituent $\tau_x$ and $\Lambda_x$ parameter matrices,

$$
\tau_x = \begin{bmatrix}
\mu_G \\
\mu_E
\end{bmatrix},
\Lambda_x = \begin{bmatrix}
1 & 0 \\
0 & 1
\end{bmatrix}.
$$

(27)

The measurement model for exogenous variables $X$ in Equation 24 can easily be modified in the usual way (Jöreskog & Sörbom, 1989) to accommodate multiple indicators of each predictor construct (if they are available). The parameter matrix $\Lambda_x$ is simply expanded to include the requisite loadings (under the usual requirements for identification; see Bollen, 1989).

In a regular covariance structure analysis, there is typically an explicit statistical purpose for including multiple indicators of a particular construct in any measurement model. Multiple indicators are present so that their covariation can reveal the true variance of the underlying construct that they represent. This ensures that the fallibility with which each indicator has been measured is accounted for in the measurement model and ultimately leads to estimates of their associated error variances and reliabilities. In the case of exogenous variables $X$, this is achieved by permitting the error vector $\delta$ to contain nonzero entries and by estimating the elements of its associated error covariance matrix $O_x$ (Jöreskog & Sörbom, 1989). This same strategy is, of course, available to the investigator when covariance structure analysis is being used to examine systematic interindividual differences in change and is tantamount to applying a maximum likelihood correction for errors in the predictors to the fitted relationship between individual growth parameters and predictors of change (cf. Fuller, 1987). No other currently available software for the analysis of change permits such a correction.

Note that we have defined the LISREL $\tau_x$ parameter vector and $\Lambda_x$ parameter matrix so that the predictors of change, $G_p$ and $E_p$, are centered at their population averages and passed as deviations from their means into the LISREL latent exogenous...
The standard LISREL covariance matrix $\Phi$—which is intended to represent interrelationships among the elements of the latent exogenous vector $\xi$—is used to account for potential intercorrelations among the predictors of change:

$$\Phi = \text{Cov}(\xi) = \begin{bmatrix} \sigma_{G}^2 & \sigma_{GE} \\ \sigma_{EG} & \sigma_{E}^2 \end{bmatrix}. \quad (28)$$

The centering of the predictors of change forces the LISREL $\alpha$ vector in the forthcoming structural model to contain the population values of true intercept and slope at the population average values of $G_p$ and $E_p$ rather than at $G_p = 0$ and $E_p = 0$, respectively. In other words, the centering of the predictors of change in Equation 24 ensures that the LISREL $\alpha$ vector will continue to contain the population means of the individual growth parameters $\pi_{op}$ and $\pi_{ip}$, as it did in the earlier Equations 10 and 19. This has some convenience for later interpretation.  

The general LISREL model permits us to represent the relationship between all individual growth parameters and all predictors of change simultaneously. The Level 1 individual growth model in Equations 5-9 is unchanged. However, to express the interrelationships among the growth parameters and predictors of change, we must modify the existing LISREL structural model in Equations 10-14 so that the newly defined vector of latent exogenous predictors (which now contains the predictors of change as deviations from their means) is introduced on the right-hand side. We can do this by taking advantage of the so-far unused LISREL latent regression-weight matrix $F$ that is present in the general LISREL model for the specific purpose of modeling the relationship between the $\eta$ and $\xi$ vectors. We simply free those elements of the $\Gamma$ matrix that represent the simultaneous linear regression of true intercept and slope on the predictors of change.

A model that includes all possible linear relationships among the individual growth parameters and the predictors of change is:

$$\begin{align*}
\pi_{op} &= [\mu_{x_0}] + \begin{bmatrix} \gamma_{x_0G} & \gamma_{x_0E} \end{bmatrix} [G_p - \mu_G] \\
\pi_{ip} &= [\mu_{x_3}] + \begin{bmatrix} \gamma_{x_3G} & \gamma_{x_3E} \end{bmatrix} [E_p - \mu_E] \\
&+ \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \pi_{op} \\ \pi_{ip} \end{bmatrix} \\
&+ \begin{bmatrix} \pi_{op} - [\mu_{x_0} + \gamma_{x_0G}(G_p - \mu_G) + \gamma_{x_0E}(E_p - \mu_E)] \\ \pi_{ip} - [\mu_{x_3} + \gamma_{x_3G}(G_p - \mu_G) + \gamma_{x_3E}(E_p - \mu_E)] \end{bmatrix}.
\end{align*} \quad (29)$$

This model is the general LISREL structural model

$$\eta = \alpha + \Gamma \xi + B \theta + \xi \quad (30)$$

with constituent parameter matrices,

$$\alpha = \begin{bmatrix} \mu_{x_0} \\ \mu_{x_3} \end{bmatrix}, \Gamma = \begin{bmatrix} \gamma_{x_0G} & \gamma_{x_0E} \\ \gamma_{x_3G} & \gamma_{x_3E} \end{bmatrix}, B = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}, \quad (31)$$

and latent residual vector

$$Z = \begin{bmatrix} \pi_{op} | G, E \\ \pi_{ip} | G, E \end{bmatrix} = \begin{bmatrix} \pi_{op} - [\mu_{x_0} + \gamma_{x_0G}(G_p - \mu_G) + \gamma_{x_0E}(E_p - \mu_E)] \\ \pi_{ip} - [\mu_{x_3} + \gamma_{x_3G}(G_p - \mu_G) + \gamma_{x_3E}(E_p - \mu_E)] \end{bmatrix}. \quad (32)$$

The two elements of the latent residual vector $\xi$ in Equation 32 contain the values of true intercept and slope deviated from their conditional means, based on their linear relationships with the predictors of change. These are the “adjusted” values of true intercept and slope, partialing out the linear effects of the predictors of change (those parts of the true intercept and slope, respectively, that are not linearly related to $G_p$ and $E_p$). The latent residual vector $\xi$ in Equation 32 is therefore distributed with zero mean vector and covariance matrix $\Psi$:

$$\Psi = \begin{bmatrix} \sigma_{\xi_{10}^2}(X=0,G,E) & \sigma_{\xi_{11}^2}(X=0,G,E) \\ \sigma_{\xi_{10}^2}(X=0,G,E) & \sigma_{\xi_{11}^2}(X=0,G,E) \end{bmatrix}. \quad (33)$$

Unlike Equation 14, the $\Psi$ matrix in Equation 33 contains the partial variances and covariances of true intercept and slope, controlling for the linear effects of the predictors of change. If we successfully predict true intercept and slope by $G_p$ and $E_p$, then we would expect these partial variances to be markedly smaller than their unconditional cousins in Equation 14. In fact, the proportional declines in the variances of true intercept and slope on inclusion of $G_p$ and $E_p$ as predictors of change provide pseudo-$R^2$ statistics that can be used to summarize the magnitude of the systematic heterogeneity in change.

Maximum likelihood estimates of the new parameters in regression-weight matrix $T$, along with estimates of the other unknown parameters in $\tau_x$, $\Phi$, $\alpha$, and $\Psi$—which together characterize our hypotheses about the nature of any systematic interindividual differences in change in the population—can again be estimated straightforwardly with LISREL. In the Appendix (Model 4), we again provide a LISREL program that does so, specifying the $A_p$, $T$, $\tau_x$, $\Phi$, $\alpha$, $B$, $\Gamma$, and $\Psi$ matrices as defined in Equations 8, 23, 27, 28, 31, and 33. Fit statistics and maximum likelihood estimates of the unknown parameters in the model are listed under Model 4 in Table 4 with the exception of the measurement error variances and covariances, which have been omitted to conserve space.

Focus first on the column in Table 4 titled Model 4. The first five rows present maximum likelihood estimates of the population means and the unconditional variances and covariances of the individual growth parameters.  

These estimates are very similar in magnitude to those already examined in the investigation of interindividual differences in Model 3 of Table 3; the inclusion of predictors of change has not seriously disturbed our estimates of the heterogeneity in change present in the population. Rows 6 through 8 of Table 4 contain the estimated partial variances and covariances of true intercept and true slope, con-

---

9 Because gender is a dichotomous variable, its introduction into $X$ appears to violate the assumption of multivariate normality on which the maximum likelihood estimation is based. Bollen (1989) examined the consequences of the nonnormality of $X$ on the estimation and concluded that the estimator retains desirable statistical properties of consistency and efficiency when the predictor is measured fallibly and is truly exogenous. The gender variable meets both of these criteria in our model.

10 Even though these unconditional variances and covariances are not an explicit part of the model for systematic interindividual differences in change, they are output automatically by the LISREL program as part of the default Covariance Matrix of $\eta$ and $\xi$. 

---

875
trolling for the linear effects of adolescent gender and initial exposure to deviant behavior. Comparing the estimated conditional variances with their unadjusted cousins in rows 3 and 4 shows that the simultaneous inclusion of the predictors of change successfully reduced the unexplained variance in true intercept and slope by 11% and 4%, respectively, suggesting that boys were more exposed to deviant behavior initially, and it is this difference in exposure that forges a spurious bivariate relationship between true intercept and gender. Once exposure is controlled, however, there are no statistically significant differences in the way that boys and girls differ in true tolerance at age 13 or in the way that their true tolerance of deviant behavior changes over adolescence. This suggests that the more parsimonious representation in Model 6 is more appropriate for summarizing our findings.

Using fitted coefficients from Model 6, along with the estimated means of true intercept, true slope, and of the predictors of change, we can specify a pair of simultaneously fitted Level 2 models that explicitly describe the linear association between true change and the remaining predictor of interest, log exposure. Substituting parameter estimates from Model 6 in Table 4 into the structural model in Equation 29, we have

\[
\tilde{\pi}_p = .3190 + .2003[E_p - (-.0788)]
\]

For interpretive purposes, these fitted models can be reported explicitly in a research account or used to construct fitted true growth trajectories for prototypical adolescents at substantively interesting values of the predictors. We favor this latter approach and illustrate it in Figure 2.

The particular prototypical fitted trajectories that we have plotted in Figure 2 were selected to provide a substantively interesting display of statistically important effects, in this case, the impact of exposure on growth in tolerance. However, recall that we have also detected collinearity between gender and exposure—boys have greater initial exposure to deviant behavior than girls—suggesting that, even though gender no longer predicts growth in tolerance once exposure is controlled, fitted trajectories for prototypical boys and girls will differ because of the underlying difference in exposure. We have captured these linked differences in Figure 2 by plotting prototypical trajectories for boys and girls at the lower and upper quartile values of log exposure, with the requisite quartiles being estimated separately in subsamples of boys and girls. The four values of log exposure that we have used are as follows: (a) girls, Q1 = −.32,
Figure 2. Fitted true growth trajectories in log tolerance of deviant behavior from 11 to 15 years of age for four prototypical adolescents at varying combinations of gender and initial exposure to deviance.

Q3 = -0.09; and (b) boys, Q1 = -0.17, Q3 = 0.14. Thus, Figure 2 displays fitted true growth trajectories for four prototypical adolescents. The two extreme trajectories represent a high-exposure boy and a low-exposure girl. The slopes of all of the trajectories are approximately parallel (exposure is not a statistically significant predictor of rate of change in Equation 34), illustrating that adolescents who were more exposed to deviant behavior at age 11 do not differ in the rate at which they become more tolerant of deviance than those adolescents who were less exposed.

Figure 2 also displays important differences in elevation, illustrating that the level of tolerance during adolescence differs by initial exposure to deviant behavior and, because of the background link between gender and exposure, also by gender. The main effect of initial exposure is indicated by the vertical displacement of the fitted trajectories within gender; for both genders, higher levels of initial exposure to deviance are related to higher levels of tolerance throughout adolescence. The effect of gender is illustrated by the vertical displacement of the trajectories within exposure; for adolescents at the same level of initial exposure, boys display higher tolerance for deviant behavior than girls at all ages.

Discussion

In recent years, pioneering authors (McArdle & Epstein, 1987; Meredith & Tisak, 1990; Muthén, 1991) have demonstrated how notions of individual growth modeling can be accommodated within the general framework offered by covariance structure analysis. Their work has illustrated how the methods of covariance structure analysis—here in the guise of the LISREL program—can provide a straightforward and convenient technique for answering important research questions about the relationship between attributes of individual true change and selected characteristics of the person.

Here, we have explored and reviewed the links between these two formerly distinct conceptual arenas, carefully laying out in detail the mapping of the one onto the other. Specifically, we have reviewed and illustrated how the Level 1 (within-person) and Level 2 (between-person) models of the individual growth modeling framework can be reformatted to correspond, respectively, to the measurement and structural components of the general LISREL model with mean structures. The direct correspondence between these two pairs of models permits the population covariance matrix of the errors of measurement and the relationships among the individual growth parameters and potential predictors of change to be modeled explicitly within a covariance structure framework. Consequently, critical parameters in the investigation of systematic interindividual differences in change can readily be estimated. This innovative application of covariance structure analysis offers several important features to data analysts.

1. The method can accommodate any number of waves of longitudinal data. Willett (1988, 1989) has shown that the collection of additional waves of data leads naturally to higher precision for the estimation of the individual growth trajectory and greater reliability for the measurement of change. In the case of covariance structure analyses of change, extra waves of data act to extend the length of the empirical growth record and expand the dimensions of the sample between-wave covariance matrix (thereby increasing the number of degrees of freedom...
imposed by the requirements of statistical power and the tenets of common sense, multiple predictors of change can be included under the covariance structure approach; the specifications that we have described extend straightforwardly to these more complex situations. Predictors can represent the main effects of important correlates of change, or, by suitable preprocessing of the data set to create crossproducts among interesting combinations of predictors, statistical interactions among potential correlates can be included in the Level 2 model. In our example, we described the configuration of the LISREL parameter matrices for the case in which interindividual differences in change are related to the main effects of two predictors.

6. The method of maximum likelihood is used to provide overall goodness-of-fit statistics, parameter estimates, and asymptotic standard errors for each hypothesized model, including estimates of all Level 2 variance and covariance parameters that are central to the detection of systematic interindividual differences in change. By using the covariance structure method, the investigator benefits from all of the flexibility and utility of a well-documented, popular, and widely disseminated statistical technique. Appropriate computer software is available on many systems, both mainframe and personal computer based. In this article, we have relied on the well known LISREL computer package, but the techniques that we have reviewed can easily be implemented with other well-known packages such as EQS (Bentler, 1985), LISCOMP (Muthén, 1987), and PROC CALIS (SAS Institute, 1991).

7. By comparing the goodness of fit of explicitly specified nested models, the investigator can test complex hypotheses about the nature of interindividual differences in true change. An additional benefit of fitting an explicitly parameterized covariance structure to data using a well-tested and flexible software package such as LISREL is that selected parameters in the model specification can be individually or jointly constrained during analysis to particular values. This allows the investigator to conduct a variety of nested tests on the specific shape of the average growth trajectory and on the variability of the individual growth parameters across people. As with other common analytic approaches such as HLM (Bryk & Raudenbush, 1992), for instance, we can “fix” the value of one growth parameter (e.g., the slope) to a value common across individuals but permit another parameter (the intercept) to be random.

8. The flexibility of the general LISREL model permits the covariance structure analysis of longitudinal data to be extended in several statistically and substantively interesting ways. There are several potential extensions of the covariance structure approach that are facilitated by the flexibility of the general LISREL model. First, multiple indicators can be used to represent each predictor of change, providing a ready maximum likelihood adjustment for errors in the predictors (cf. Fuller, 1987). Second, individual change can be modeled simultaneously in more than one domain (Willett & Sayer, 1993), engendering investigation of profiles of change (see Williamson, 1986; Williamson et al., 1991). This includes the investigation of (a) interrelationships among the several types of change and (b) the simultaneous and joint association of these several changes and selected predictors of change. Third, the method enables the modeling of intervening effects, whereby a predictor may not act directly on change but indirectly through the influence of intervening variables.
Of course, other analytic methods are available for fitting these hierarchical models that describe individual change and individual differences in change. Each method has its own strengths. Nevertheless, one particular method may offer benefits over another in a specific research setting. For instance, the HLM approach of Bryk and Raudenbush (1992) is also well documented and flexible, and software suitable for conducting the analyses is widely available. And, like the covariance structure approach, HLM can handle any number of unequally spaced waves of longitudinal data; individual growth trajectories can be linear or curvilinear; multiple predictors of change can be included in the Level 2 model; a variety of goodness-of-fit statistics, parameter estimates, and standard errors are provided; and complex hypotheses about change over time can be tested through the analysis of contrasts and the program's ability to restrict individual growth parameter variances to zero at the investigator's behest. In addition, HLM offers an advantage that cannot be matched by the covariance structure method as we have described it here: It does not require time-structured data; each individual in the data set can possess an empirical growth record containing different numbers of waves of data with randomly assigned temporal spacing (but see Muthén, 1992).

Nevertheless, we believe that there is no single analytic method that can be declared unilaterally the best, nor have we written this article to declare such a winner. Different empirical settings demand different analytic decisions. When the data are not time structured, either by accident or design, then the investigator may have to set aside the analytic methods that we have described here. However, when equal numbers of waves of data are available on each subject, the covariance structure approach offers great flexibility in the investigation of systematic interindividual differences in change and an unparalleled opportunity to model error covariance structure explicitly. That a variety of analytic tools are available for such an important task is a boon rather than a hindrance, leading us to conclude—to paraphrase Cronbach and Furby (1970)—that we can measure "change"—and we should!

References

Appendix

Sample LISREL VII Programs

In this appendix, we present the LISREL VII programs that were used to fit Models 1 through 4. All have a similar structure. We present the program used to fit Model 1 in its entirety. For the remaining models, we present the title lines and then highlight the code that differentiates each from the other.

In the first part of any complete program, we list several title lines to distinguish the particular model being fitted. Following the title lines is the data definition (DA) line, in which we specify a single group rather than multisample analysis (NG = 1), the number of input variables (NI), and the sample size (NO). The next line identifies the file that contains the input data. Then we label the input variables (LA) and, when necessary, select those that are to be analyzed in the current run (SE).

The model definition (MO) line describes the basic shape of the various hypothesized LISREL variable and parameter matrices. The Y, q, X, and ε score vectors are dimensioned (NY, NE, NX, NE), and the shape and initial contents of the τ, a, B, Ψ, and Ψ parameter matrices are specified (TY, TX, PH, AL, BE, GA, PS) according to their definitions in text. Afterward, the contents of the latent growth record and the vector of latent exogenous predictors are labeled (LE, LK).

Then, in several lines, we specify the various fixed and free parameters that constitute the particular measurement and structural models listed in the text. We completely fix the contents of the A, and A, matrices (in the matrix of stated values following the MA LY and MA LX lines). We free appropriate elements of the 0, vector (FR TE) according to the hypothesized error covariance structure and elements of the Ψ matrix (FR GA) according to the hypothesized structural model. Occasionally, parameter groups are constrained to be equal. For instance, when the errors of measurement are assumed to be homoscedastic, selected diagonal elements of the 0, matrix are set equal (EQ TE). Similar constraints could be used to test specific hypotheses in a nested sequence of models.

Finally, in the “output” line (OU), we indicate that we require the estimation and printing of standard errors (SE) and f values (TV), as well as the printing of a residual analysis (RS), all to six decimal places (ND = 6). We limit the maximum number of iterations during estimation (IT = 10,000).

Model 1: Linear individual growth in log deviance
Origin of time is age 13
No predictors
Independent homoscedastic errors

Model 2: Linear individual growth in log deviance
Origin of time is age 13
No predictors
Independent heteroscedastic errors

Model 3: Linear individual growth in log deviance
Origin of time is age 13
Pairwise autocorrelated heteroscedastic errors

Model 4: Linear individual growth in log deviance
Origin of time is age 13
Predictors are gender and initial log exposure to deviance at age 11
Pairwise autocorrelated heteroscedastic errors

MO NY = 5 TY = ZE NE = 2 TE = SY, FI AL = FR BE = ZE PS = SY, FR
LE
‘PI0’ ‘PI1’
MA LY
1 2
1 1
1 0
1 1
1 2
FR TE(1,1) TE(2,2) TE(3,3) TE(4,4) TE(5,5)
EQ TE(1,1) TE(2,2) TE(3,3) TE(4,4) TE(5,5)
OU SE TV RS ND = 6 IT = 10000

Received February 17, 1993
Revision received September 17, 1993
Accepted December 17, 1993