Questions and Answers  
In the Measurement of Change

JOHN B. WILLETT  
Harvard University  
Graduate School of Education

In some ways, [measuring individual change over time] is the most important topic in educational measurement. The primary object of teaching is to produce learning (that is, change), and the amount and kind of learning that occur can be ascertained only by comparing an individual’s or a group’s status before the learning period with what it is after the learning period.

Frank B. Davis  
*Educational Measurements and Their Interpretation*

It is the increment in achievement the school provides, which should be the measure of the school’s quality. If we had good measures of that increment, as well as good measures of the level of various school resources in the same school, it would be possible to establish a relation between the size of the increment and the level of certain resources, and thus to determine which school resources were most important to learning.

James S. Coleman  
*Methods and Results in the IEA Studies of Effects of School on Learning*

Differences between scores tend to be much more unreliable than the scores themselves.

Frederic M. Lord  
*The Measurement of Growth*

Investigators who ask questions regarding gain scores would ordinarily be better advised to frame their questions in other ways.

Lee J. Cronbach and Lita Furby  
*How We Should Measure “Change”—or Should We?*

345
I. WHY WE SHOULD MEASURE "CHANGE"—AND, CAN WE?

Fundamental questions in education center upon issues of individual learning. For instance, if we are to improve educational practice in our schools, then accurate measurement of individual learning can provide one of the yardsticks by which the effectiveness of pedagogy and the effectiveness of educational innovation can be judged. Similarly, valid and reliable statistical tools are required for the measurement of individual learning if we are to distinguish students with learning handicaps from their unimpaired colleagues, or if we are to select small groups of learners for enrichment or remediation. Students enter schools to learn, to grow, to develop, to change. It is these changes that are created and maintained by the activities and resources of our schools. It is the measurement of these changes, and the investigation of their relationship to supporting activities in the classroom and the resources provided by the school, with which the empirical investigator should ultimately be concerned.

Thus, the very notion of learning implies growth and change. And yet, as a consequence of methodological controversy (e.g., Bereiter, 1963; Linn & Slinie, 1977), influential authors writing on the measurement of change have offered little solace to the empirical researcher interested in assessing individual learning. Rather, it has been suggested that questions of individual learning should be reframed as questions of educational status rather than educational growth (Cronbach & Furby, 1970; Werts & Linn, 1970). Thus, teaching effectiveness has been judged by the influence of pedagogy on student achievement, attitude, and interest at a single point in time, rather than by its impact on their growth over time. Educational innovations have similarly been assessed on the basis of their effect at some given instant, and the influence of learning disability judged from a single "snapshot" of the

---

This work was partially supported by the William F. Milton Fund and the Spencer Foundation. I would like to acknowledge the continuous help and guidance of my close colleagues: Richard J. Murnane, for providing detailed and thoughtful feedback during the writing of this paper, and Judith D. Singer, for participating (often unknowingly) in the conceptual development by virtue of our once-a-week lunchtime meetings at the Ta Chien Restaurant in Cambridge. In addition, many thanks also go to Anthony S. Bryk of the University of Chicago who, in his role as the consulting editor, offered encouragement and advice throughout the writing process.
child. In other words, questions of individual learning and development, questions that quite naturally fall into a longitudinal framework, have frequently been addressed by research studies with a cross-sectional design.

Between the idea of measuring change and the reality of its empirical measurement has fallen the shadow of an unnatural, or at least unhelpful, conceptualization. It is a conceptualization that visualizes change as an “increment,” and change measurement as a comparison of “an individual’s or a group’s status before the learning period with what it is after the learning period” (Davis, 1964, p. 234). It is a conceptualization that views individual learning, not as a process of continuous development over time, but as the quantized acquisition of skills, attitudes, and beliefs. It is as though the individual is delivered of a quantum of learnings in the time period that intervenes between the premeasure and the postmeasure, and that our only concern should be with the size of the acquired “chunk.”

This chapter contends that viewing change in this unnatural way has prevented empirical researchers in the social sciences from entertaining a richer, broader spectrum of research questions, questions that deal with the nature of individual development and that center upon the parameters of the very process of learning. Moreover, the unnatural conceptualization is responsible for creating many of the “problems in measuring change” (Harris, 1963). It has fostered an approach to change measurement that deals exclusively in “two-wave” designs, whose principal measures are the difference score and the residual-change score, whose predominant graphical displays are scatterplots of the posttest on the pretest, whose major group-level statistical summaries are correlations of the posttest and the pretest, whose overriding concerns are unreliability, invalidity, and regression to the mean. It is a conceptualization that, by virtue of its very existence, has retarded the longitudinal investigation of individual development, and it has prevented the generalization of traditional measures of change to more than two waves of data. The attitude predominant among empirical researchers over much of the last half-century seems to have been: If individual change cannot be measured adequately with two waves of data, why bother collecting three (or more) waves? And, anyway, how would the extra wave of data be used if it were collected?

Recent methodological research in the measurement of individual change has been reoriented from “change” to “growth” and has begun to focus directly on the “continuously moving line of each child’s development” (Bock, 1976, p. 76). This new longitudinal approach was born out of a long statistical tradition in the modeling of individual growth over time. Under the new approach, multiple waves of data are collected on a sample of individuals over time, and an explicit model of individual growth is adopted as the basis for the statistical analysis (see Bryk, 1977; Bryk & Weisberg, 1977). Application of the new multiwave approach has illuminated many of the outstanding issues and problems in the measurement of individual change and has led to the
development of methods for a richer, more complete analysis of educational growth.

Nevertheless, to a large extent, empirical research in education has yet to benefit from these methodological advances. Thus, this chapter fosters a longitudinal perspective on individual growth. It serves to reassure empirical researchers that they certainly can measure individual growth and that it is really not too hard to do a pretty good job! It asks, Can individual growth be measured adequately, or even well? And, if so, how? Is it possible to distinguish among individuals on the basis of their learning? Competing methodologies for the empirical measurement of change are chronicled and contrasted. The presentation illustrates how traditional two-wave methods for the measurement of individual change and the newer multiwave methods are related. Problems, both real and imagined, that have beset change measurement for many years are resolved. Finally, a hierarchy of exploratory and easily applied strategies that can be effectively employed in the empirical measurement of individual change is outlined and demonstrated.

II. GATHERING, DISPLAYING, AND SUMMARIZING GROWTH DATA

In Section II, comments are made on the typical structure of empirical growth studies. Concerns of design naturally lead to a consideration of the attributes of the data that might be gathered in a growth study, and strategies for displaying and summarizing those data begin to be explored. Alternative perspectives on the representation of individual growth are discussed in advance of any specific discussion of technical methodology for the measurement and analysis of growth and change. One of these alternative perspectives—based on individual growth trajectories—is shown to be more informative and productive. The various displays created in this section provide a framework for the reinterpretation of problems in the measurement of change in Section III, and underpin the technical presentation in Sections IV and V, where the application of the individual growth-trajectory perspective to empirical growth measurement is expanded.

Phases of a Growth Study

Coleman’s quotation (1975, p. 355), cited at the beginning of the review, makes explicit the notion that the measurement of change is usually only the first phase in a more complete and complex study of growth. He argues that interest is centered more typically on a second-phase data analysis in which differences in learning are related to differences in “resources.” It is during such a second-phase analysis that investigators determine “which school resources were most important to learning.” Although all investigations of learning do not focus their attention at the level of aggregation that interests Coleman, the point is well made. Investigations of growth and change can be conceptualized typically as hierarchical enterprises comprised of two linked
phases. Thus, when individual learning is the principal focus of interest, the complete investigation of that learning includes both within-individual and between-individual analyses.

In the first phase of the investigation, interest centers upon the measurement of the cognitive or psychological growth of each subject in the sample separately. Thus, data are collected on each individual over time and the growth of the individual is characterized by some measure of change (such as the gain score or the residual-change score, or by some other measure). Whatever the measurement strategy applied, this strategy is responsible for generating a within-subject summary of the growth of each individual in the sample over time. Subsequently, the initial within-subject summaries of individual growth become the basis of the second-phase between-subject analysis in which variations in growth from subject to subject are related to interindividual differences in “resources.”

The basic framework for such multilevel analyses was described by Bryk (1977). Thus, “the objectives of longitudinal research are to characterize patterns of individual . . . change over time and to investigate the effects of covariates on these patterns” (Ware, 1985, p. 95). Correlates and predictors of individual growth are located by systematically linking interindividual differences in growth with interindividual heterogeneity in selected background characteristics. Rogosa and Willett (1985) have characterized such investigations as the search for systematic individual differences in growth.

In general, the universe of interesting covariates can be construed quite broadly to include measures of the school, the home, and the peer group, and measures of attitude, interest, personality, and other important behavioral parameters. Typically, in educational research, the particular covariates employed in the second phase of a growth analysis are what have been called “between-subject covariates” (Ware, 1985, p. 96). These are covariates that either do not vary over time (for instance, covariates such as gender, ethnic background, and other demographic variables), or are relatively slow-moving compared to the individual growth processes that are being investigated (for instance, covariates such as the socioeconomic status of the home and some measures of behavioral status). For this reason, and for reasons of simplicity, this chapter focuses primarily upon strategies for the investigation of systematic relationships between individual growth and time-invariant covariates.

Notice that categorical (dummy) variables, such as those representing gender and ethnic background, can be considered legitimate correlates and predictors of individual growth in the between-subject phase of a growth investigation. Similarly, by suitable representation of an educational innovation or intervention by a dummy variable, the methods to be described in this review may be extended to the investigation of individual change as a function of an educational innovation or program, either in a randomized or a nonrandomized setting.
Data Collection in the Study of Growth

In a typical empirical growth investigation, important questions involve both the design of the study and the nature of the data that must be collected. Typically, the researcher asks: How many subjects will my study require? On how many occasions should I measure each of these individuals? What covariates are important when systematic interindividual differences in growth are examined? What type of instruments should be used to measure the outcome variable and the covariates, and what psychometric properties are they required to have? These questions are considered here, and throughout the chapter.

Issues of Design

Typically, the data gathered in a growth study include measurements on the growth of each of several individuals over time, along with measures of a variety of background variables of interest to the investigator. Thus, a sample of individuals must be selected from the population of interest and their growth observed. Although the underlying growth of each subject can be conceived of as a continuous process in which the cognitive, social, or psychological status of each individual changes smoothly with time, it is usually impossible for the investigator to obtain a complete and continuous history of the growth of all the subjects in the sample. Therefore, the investigator must draw a sample of subjects from the population and also must observe and measure these subjects on a sample of occasions.

Interesting and difficult issues in the design of growth studies arise when the investigator begins to ask: How many subjects do I need in my sample? and, On how many occasions do I need to collect data? As in the application of any technique of statistical inference, the precise number of subjects required in any given empirical investigation is a complex function of the effect size that is to be detected, the required statistical power, and the selected level of significance. In general, larger numbers of subjects are required for the detection of smaller effects, all else being equal.

Empirical researchers in education have traditionally adopted a two-wave design for studies of individual growth. Thus, for instance, in a study of growth in the reading ability of elementary school children, tradition has demanded only that a pair of measurements be obtained on each of the subjects—one measurement at the beginning and one at the end of the study. A more informed longitudinal investigation of the same question may require individual observations to be made every other month for a year. Under both of these designs we would say that panel data had been obtained, but that under the traditional design two waves of data were collected while, in the latter design, the number of waves was six.

This chapter contends that growth measurement will always be improved
when panel data are collected over a larger number of occasions because “two
time points provide an inadequate basis for studying change” (Bryk & Rau-
denbush, 1987). Furthermore, when more waves of data are gathered on each
subject in the sample, the details of the individual growth trajectories become
much clearer to the investigator. It is possible to determine whether the indi-
vidual is growing at the same rate for all of time or whether more complex,
nonlinear growth is evident. Thus, the researcher enters an arena in which
richer and more complex research questions can be both asked and answered.
By gathering multiple waves of data, both substantive and methodological
 gains are made in the research.

Issues of Measurement

Types of data gathered in growth research. In social research, it is useful
to distinguish between two different types of outcome measure: qualitative
(categorical) outcome measures, and quantitative (continuous) outcome mea-
 sures. An example of the former is graduation from college, whereas test
scores are usually construed as examples of the latter. In educational research,
as in most psychological and social research, data are obtained by the appli-
cation of both types of measure. This chapter focuses entirely on methods for
the measurement and analysis of individual growth in quantitative outcomes
over time.

Construct validity. Whether instruments are actually measuring the con-
struct they purport to measure is of crucial concern in empirical research.
Typically, investigators will go to great lengths to ensure that construct valid-
ity can be claimed for the particular application of their measuring instru-
ments. The empirical investigation of growth is no different; all of the usual
psychometric technology for the validation of measuring instruments must
also be brought to bear. In this chapter, we will assume that valid measures of
the required variables are available either by virtue of a prior validation study
of those instruments by the investigator or by virtue of the standard and well-
accepted character of the measures.

Construct validity/equatability of the outcome measure overtime. A re-
lated issue more fundamental to the measurement of growth is whether the
selected measure remains construct valid across subsequent occasions of mea-
surement, and whether the scores obtained are equatable from occasion to
occasion. If the measure does not remain construct valid over time then it will
not be appropriate to assume that the obtained data have been measured on
the same metric on subsequent occasions. This creates insurmountable prob-
lems for the interpretation of the obtained data: if the variable being measured
does not retain the same psychological meaning over the occasions of mea-
surement then it is not meaningful to think about change. This has been ex-
pressed by Lord (1958) in a two-wave context when he considers the influ-
ence of instruction on students in the classroom:
It is implicit in the use of the word gain that the initial and final measures are expressed in the same metric. In the case where the variable under consideration is the student’s weight in pounds, there seems to be little question that the initial and final measures are expressed on the same scale. In the case of test scores, this may not be so obvious. Even though the pretest and the post-test consist of the same test questions and are physically identical, it is quite possible to maintain that the student has changed drastically during the course of instruction and that . . . the test no longer measures the same thing when given after instruction as it did before instruction. If this is asserted, then the pretest and the post-test are measuring different dimensions and no amount of statistical manipulation will produce a measure of gain or growth (Lord, 1958, p. 440).

Similar concerns have been expressed by Bereiter (1963), Cronbach and Furby (1970), Linn and Slinde (1977), Lord (1963), and Rogosa, Brandt and Zimowski (1982).

In addition to the problems of equatability of scores over time, empirical investigations of individual change generally make use of tests that have been developed to differentiate among individuals on a single occasion of measurement. Usually, during instrument construction, little consideration is given to the psychometric properties of these measures with regard to differentiating among individuals on the basis of their growth. However, recent developments in Item Response Theory (Bock, 1976; Lord, 1980) seem likely to lead to the development of better tests and measuring instruments for the empirical investigation of growth. In this chapter, it will be assumed that the outcomes being investigated are measured by instruments that maintain their construct validity over time and that the obtained scores remain equitable from occasion to occasion.

Displaying and Summarizing Growth Data

Typically, research questions in education have been framed as pre/post comparisons. Therefore, empirical investigations of educational growth have traditionally been conceived of as two-wave designs. Thus, in a study of academic growth, for instance, measures of mathematics achievement, vocabulary, or reading ability might typically be obtained at the beginning and end of the year, or on two consecutive years for a sample of students. Having opted to gather two-wave panel data, the investigator must also devise techniques for displaying and summarizing those data. There are at least two alternative perspectives that the investigator might adopt when creating such summaries, and each of the perspectives leads to different displays and summaries of the two-wave panel data. One of these perspectives, and the one found more frequently—either implicitly or explicitly—in the educational research literature, relies on a between-wave scatterplot of outcome (i.e., “final” observed status plotted against “initial” observed status) as an appropriate display of the growth information. The second perspective, on the other hand, focuses on plots of the individual growth trajectories over time (i.e., observed status is plotted against time for each individual).
In this section it is argued that, although both types of summary display exactly the same information, it is the individual growth-trajectory perspective that is more appropriate in the context of the empirical investigation of individual growth. Many of the problems that appear to beset the measurement of change are, in fact, artifacts of an inappropriate perspective. Furthermore, the individual growth-trajectory perspective is more easily generalizable to multiwave data, whereas longitudinal generalizations of the between-wave scatterplot perspective are both cumbersome and misleading. Both of the competing perspectives are discussed here in some detail.

**Two-Wave Individual Growth Data**

The growth displays and summaries that are available under either of the two conceptualizations are introduced here by virtue of a data-example. The two-wave panel data displayed in Exhibit 1 are part of an artificial dataset created specifically for analysis in this chapter. Analysis of this same data example continues throughout the paper.

For the purposes of our analysis, Exhibit 1 can be regarded as presenting the observed scores of 35 randomly selected individuals on a proficiency test over two occasions of measurement. For instance, observed growth records like these might have been obtained if, for example, all the subjects had been administered a test of “opposites-naming” on each of two consecutive days. To obtain the scores, each subject could have been presented with a list of words whose opposites had to be named in a fixed amount of time. Individual growth would occur as subjects became more adept at the task (after Chapman, 1914).²

Exhibit 1 also provides an identification number and the value of an appropriate covariate (“correlate of growth”) for each subject at the time of the study. Typical covariates might include measures describing intrinsic time-invariant characteristics of the individuals or features of their environment. In addition to the individual growth records, summary statistics are provided for each of the variables in the exhibit along with simulated estimates of the reliability of the “opposites-naming” test on each of the two occasions of measurement. With this dataset, a question of research interest might be: Are interindividual differences in the growth of “opposites-naming skill” systematically related to the covariate? In particular, do subjects with higher values of the covariate acquire this skill more rapidly than subjects with lower values?

Between-wave scatterplot perspective. Even though principal research interest centers upon individual growth, the structure of the data-matrix in Exhibit 1 invites the representation of “final” observed status as an “outcome” to be “predicted” by “initial” observed status. Thus, methodologists and empirical researchers alike have found it natural to create a between-wave scatterplot that represents final observed status as a function of initial observed
<table>
<thead>
<tr>
<th>Subject ID</th>
<th>Covariate</th>
<th>Observed Scores</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>137</td>
<td>205</td>
<td>217</td>
<td></td>
</tr>
<tr>
<td>02</td>
<td>123</td>
<td>219</td>
<td>243</td>
<td></td>
</tr>
<tr>
<td>03</td>
<td>129</td>
<td>142</td>
<td>212</td>
<td></td>
</tr>
<tr>
<td>04</td>
<td>125</td>
<td>206</td>
<td>230</td>
<td></td>
</tr>
<tr>
<td>05</td>
<td>81</td>
<td>190</td>
<td>220</td>
<td></td>
</tr>
<tr>
<td>06</td>
<td>110</td>
<td>165</td>
<td>205</td>
<td></td>
</tr>
<tr>
<td>07</td>
<td>99</td>
<td>170</td>
<td>182</td>
<td></td>
</tr>
<tr>
<td>08</td>
<td>113</td>
<td>98</td>
<td>131</td>
<td></td>
</tr>
<tr>
<td>09</td>
<td>104</td>
<td>138</td>
<td>156</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>96</td>
<td>216</td>
<td>252</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>125</td>
<td>180</td>
<td>225</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>115</td>
<td>97</td>
<td>138</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>109</td>
<td>145</td>
<td>161</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>95</td>
<td>195</td>
<td>184</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>118</td>
<td>162</td>
<td>138</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>120</td>
<td>119</td>
<td>148</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>118</td>
<td>144</td>
<td>168</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>115</td>
<td>107</td>
<td>165</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>120</td>
<td>167</td>
<td>201</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>118</td>
<td>158</td>
<td>158</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>126</td>
<td>165</td>
<td>228</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>121</td>
<td>197</td>
<td>161</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>108</td>
<td>206</td>
<td>209</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>104</td>
<td>182</td>
<td>198</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>118</td>
<td>174</td>
<td>198</td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>104</td>
<td>199</td>
<td>238</td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>124</td>
<td>160</td>
<td>178</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>130</td>
<td>184</td>
<td>231</td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>87</td>
<td>174</td>
<td>194</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>131</td>
<td>215</td>
<td>226</td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>109</td>
<td>147</td>
<td>168</td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>115</td>
<td>127</td>
<td>172</td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>104</td>
<td>165</td>
<td>217</td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>110</td>
<td>76</td>
<td>139</td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>110</td>
<td>166</td>
<td>197</td>
<td></td>
</tr>
</tbody>
</table>

| Mean       | Observed Scores | 184.5 | 192.0 |
| Std. Dev.  | 12.49          | 36.18 | 32.54 |
| Reliability| .8243          | .8727 |       |

status (see Figure 1). Notice that, although the displayed datapoints are widely scattered, a trend is evident in the plot: subjects with high initial observed scores were more likely to score highly on the second occasion of measurement. This suggests that the empirical bivariate relationship can be captured succinctly in group-level summary statistics such as the estimated correlation of initial and final observed status, or the estimated slope in a
regression of final observed status on initial observed status. The fitted line obtained in this latter regression is displayed in Figure 1. The estimated regression slope is .745 (with a standard error of .096), and the estimated between-wave correlation of “opposites-naming skill” is .804.

Individual growth-trajectory perspective. A simple and obvious alternative strategy for plotting the individual growth data in Exhibit 1 is presented in Figure 2. In this latter plot the vertical and horizontal axes represent observed status and time respectively, rather than final observed status and initial observed status as in Figure 1. Furthermore, a separate observed growth-trajectory has been drawn on the plot for each subject in the sample by simply joining the initial observed status and the final observed status with a straight line for each pair of datapoints. Thus, each line segment represents the linear observed growth of an individual in the group (constituting the within-individual phase of the growth analysis) and all the line segments viewed simultaneously inform as to interindividual differences in growth across the group (constituting part of the between-individual phase of the growth analysis).

Notice that interpretations of individual growth and interindividual differences in growth are easily made from the individual growth-trajectory plots in Figure 2. In particular, although the observed scores of every individual in the sample have changed between the two occasions of measurement, not all the changes are equal. Thus, there is readily recognizable empirical evidence of interindividual heterogeneity in growth. Group-level summary statistics that characterize the between-individual phase of the growth analysis include the estimated variance of the slopes of the individual growth trajectories (to represent the interindividual heterogeneity in growth), and the estimated correlation between the slopes of the individual growth trajectories and important covariates (to represent the systematic interindividual differences in growth). In Figure 2, the estimated variance of the observed slopes is 482.3 and the estimated correlation between the observed slopes and the covariate is .101.

Multiwave Individual Growth Data

The two-wave displays and summaries of the foregoing subsection can be generalized to the multiwave case by virtue of an extension of the data-example introduced in Exhibit 1. The panel data displayed in Exhibit 2 are four waves of data drawn from the same simulation, the first two waves having been already presented in Exhibit 1. Thus, Exhibit 2 can be regarded as presenting the “number of opposites correctly named in 600 seconds” by each subject on each of four consecutive days. Also included are the subject identification codes and values of the time-invariant covariate, as in Exhibit 1.

Between-wave scatterplot perspective. Unfortunately, because a between-wave scatterplot represents observed status on one occasion as a function of observed status on some earlier occasion, it is not readily and easily gener-
FIGURE 1: Between-wave scatterplot of observed status. A plot of observed score on the second occasion of measurement ("final observed status") as a function of observed score on the first occasion of measurement ("initial observed status") for the 35 individuals in the sample of two-wave growth data in Exhibit 1.

FIGURE 2: Collection of fitted (two-wave) individual growth trajectories. A plot of observed score versus occasion of measurement, showing growth trajectories for each of the 35 individuals in the sample of two-wave growth data in Exhibit 1. Each observed growth trajectory was obtained by plotting, and connecting with a straight line, the initial and final observed status of each individual in the sample.
**EXHIBIT 2**  
Values of a Time-Invariant Covariate and Observed Scores for 35 Subjects on Four Occasions of Measurement

<table>
<thead>
<tr>
<th>Subject ID</th>
<th>Covariate</th>
<th>Observed Scores</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>01</td>
<td>137</td>
<td>205</td>
<td>217</td>
<td>268</td>
<td>302</td>
</tr>
<tr>
<td>02</td>
<td>123</td>
<td>219</td>
<td>243</td>
<td>279</td>
<td>302</td>
</tr>
<tr>
<td>03</td>
<td>129</td>
<td>142</td>
<td>212</td>
<td>250</td>
<td>289</td>
</tr>
<tr>
<td>04</td>
<td>125</td>
<td>206</td>
<td>230</td>
<td>248</td>
<td>273</td>
</tr>
<tr>
<td>05</td>
<td>81</td>
<td>190</td>
<td>220</td>
<td>229</td>
<td>220</td>
</tr>
<tr>
<td>06</td>
<td>110</td>
<td>165</td>
<td>205</td>
<td>207</td>
<td>263</td>
</tr>
<tr>
<td>07</td>
<td>99</td>
<td>170</td>
<td>182</td>
<td>214</td>
<td>268</td>
</tr>
<tr>
<td>08</td>
<td>113</td>
<td>96</td>
<td>131</td>
<td>159</td>
<td>213</td>
</tr>
<tr>
<td>09</td>
<td>104</td>
<td>138</td>
<td>156</td>
<td>197</td>
<td>200</td>
</tr>
<tr>
<td>10</td>
<td>98</td>
<td>216</td>
<td>252</td>
<td>274</td>
<td>298</td>
</tr>
<tr>
<td>11</td>
<td>125</td>
<td>180</td>
<td>225</td>
<td>215</td>
<td>249</td>
</tr>
<tr>
<td>12</td>
<td>115</td>
<td>97</td>
<td>138</td>
<td>169</td>
<td>222</td>
</tr>
<tr>
<td>13</td>
<td>109</td>
<td>145</td>
<td>161</td>
<td>151</td>
<td>177</td>
</tr>
<tr>
<td>14</td>
<td>95</td>
<td>195</td>
<td>184</td>
<td>209</td>
<td>213</td>
</tr>
<tr>
<td>15</td>
<td>118</td>
<td>162</td>
<td>138</td>
<td>204</td>
<td>195</td>
</tr>
<tr>
<td>16</td>
<td>120</td>
<td>119</td>
<td>148</td>
<td>164</td>
<td>208</td>
</tr>
<tr>
<td>17</td>
<td>118</td>
<td>144</td>
<td>166</td>
<td>236</td>
<td>261</td>
</tr>
<tr>
<td>18</td>
<td>115</td>
<td>107</td>
<td>165</td>
<td>193</td>
<td>262</td>
</tr>
<tr>
<td>19</td>
<td>120</td>
<td>167</td>
<td>201</td>
<td>233</td>
<td>218</td>
</tr>
<tr>
<td>20</td>
<td>118</td>
<td>156</td>
<td>156</td>
<td>197</td>
<td>246</td>
</tr>
<tr>
<td>21</td>
<td>126</td>
<td>165</td>
<td>228</td>
<td>279</td>
<td>290</td>
</tr>
<tr>
<td>22</td>
<td>121</td>
<td>197</td>
<td>181</td>
<td>185</td>
<td>217</td>
</tr>
<tr>
<td>23</td>
<td>108</td>
<td>206</td>
<td>209</td>
<td>230</td>
<td>255</td>
</tr>
<tr>
<td>24</td>
<td>104</td>
<td>182</td>
<td>196</td>
<td>217</td>
<td>199</td>
</tr>
<tr>
<td>25</td>
<td>118</td>
<td>174</td>
<td>198</td>
<td>229</td>
<td>236</td>
</tr>
<tr>
<td>26</td>
<td>104</td>
<td>199</td>
<td>238</td>
<td>253</td>
<td>282</td>
</tr>
<tr>
<td>27</td>
<td>124</td>
<td>160</td>
<td>178</td>
<td>189</td>
<td>229</td>
</tr>
<tr>
<td>28</td>
<td>130</td>
<td>184</td>
<td>231</td>
<td>260</td>
<td>292</td>
</tr>
<tr>
<td>29</td>
<td>87</td>
<td>174</td>
<td>194</td>
<td>189</td>
<td>168</td>
</tr>
<tr>
<td>30</td>
<td>131</td>
<td>215</td>
<td>228</td>
<td>257</td>
<td>310</td>
</tr>
<tr>
<td>31</td>
<td>109</td>
<td>147</td>
<td>168</td>
<td>197</td>
<td>232</td>
</tr>
<tr>
<td>32</td>
<td>115</td>
<td>127</td>
<td>172</td>
<td>222</td>
<td>273</td>
</tr>
<tr>
<td>33</td>
<td>104</td>
<td>165</td>
<td>217</td>
<td>230</td>
<td>266</td>
</tr>
<tr>
<td>34</td>
<td>110</td>
<td>76</td>
<td>139</td>
<td>150</td>
<td>214</td>
</tr>
<tr>
<td>35</td>
<td>110</td>
<td>166</td>
<td>197</td>
<td>203</td>
<td>233</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Reliability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Covariate</td>
<td>113.5</td>
<td>164.5</td>
<td>.8243</td>
</tr>
<tr>
<td>Observed Scores</td>
<td>192.0</td>
<td>216.7</td>
<td>.8727</td>
</tr>
<tr>
<td>1</td>
<td>268</td>
<td>35.94</td>
<td>.8812</td>
</tr>
<tr>
<td>2</td>
<td>302</td>
<td>37.62</td>
<td>.9331</td>
</tr>
</tbody>
</table>

alizable to the representation of multiwave data on a single plot. Rather, empirical researchers are likely to conceptualize multiwave data as a sequence of discrete two-wave "chunks" in order to display the data by creating all possible pairwise between-wave scatterplots (with appropriate summary statistics attached to each). Thus, with the four waves of data in Exhibit 2, the investigator may examine observed status on the fourth occasion of measure-
TABLE 1

Estimated Between-Wave Correlations of Observed Score for the Sample of four-Wave Data in Exhibit 2

<table>
<thead>
<tr>
<th>Occasion</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>.804</td>
<td></td>
<td>.709</td>
<td>.414</td>
</tr>
<tr>
<td>.804</td>
<td>.678</td>
<td></td>
<td>.847</td>
<td>.678</td>
</tr>
<tr>
<td>.709</td>
<td>.847</td>
<td>1.000</td>
<td></td>
<td>.800</td>
</tr>
<tr>
<td>.414</td>
<td>.678</td>
<td>.800</td>
<td>1.000</td>
<td></td>
</tr>
</tbody>
</table>

...ment as a function of observed status on the third occasion, and as a function of the observed status on the second and on the first occasions of measurement too. Similarly, observed status on occasion three can be regarded as a function of observed status on occasions two and one, and so forth. Each of these pairwise plots can then be examined using whatever statistical and psychometric methodology is considered appropriate in the two-wave case.

Under this approach, however, with T waves of data collected, a total of $T(T - 1)/2$ scatterplots are required to display all of the possible bivariate relationships among the measures of observed status. With as few as 8 occasions of measurement, there are 28 separate between-wave scatterplots to be examined. Therefore, the use of between-wave scatterplots to display multiwave growth data rapidly becomes cumbersome. This being the case, it is common to find empirical researchers dispensing with the plots completely and focusing entirely on the associated summary statistics. Specifically, they rely upon an estimated between-wave correlation matrix of the type shown in Table 1 in the forlorn hope of somehow capturing the essential features of the individual growth in the group.

The matrix of estimated correlations in Table 1 is typical of correlation matrices found with most longitudinal data. As might reasonably be anticipated, correlations estimated between occasions of measurement that are more remote from each other have a smaller magnitude; thus, the estimated correlation between observed status on the first and second occasions is .804, whereas there is an estimated correlation of .709 between the first and third waves. Similarly, the estimated correlation of observed status between waves two and four is smaller than the equivalent estimate obtained between waves two and three, and so forth. It is these inequalities that give the estimated correlation matrix its familiar “band-diagonal” appearance. Notice that little can be inferred about individual growth and about inter-individual differences in growth from the between-wave correlation matrix.

Individual growth-trajectory perspective. A simple, more informative strategy for displaying the growth data in Exhibit 2 is to adopt an individual growth-trajectory perspective, as was done earlier with the two-wave data. As
before, the growth of each individual can be conceptualized as an ongoing, continuous process of which we have obtained "snapshots" of observed status on each of four occasions. The appropriate data display continues to be the individual growth-trajectory plot although, rather than there being two important fiducial marks on the time-axis (as before, corresponding to the two occasions of measurement), now there are four. Thus, for example, the observed growth record of subject #21 is displayed in Figure 3.

In Figure 2, where individual growth trajectories were plotted for all the subjects in the sample on the basis of the two-wave data in Exhibit 1, the pairs of datapoints were simply connected with a straight line to represent the growth of each individual. Of course, in practice, there is absolutely no a priori justification for assuming that individual growth is necessarily linear (although, by choice in the artificial data of Exhibits 1 and 2, the assumption of linear true growth is appropriate). Nevertheless, when there are only two waves of data available on each subject, there is no way to know the exact shape of individual growth over time because it is possible to draw any mathematical curve through two datapoints on a scatterplot. Therefore, in Figure 2, the simplest possible individual growth trajectory—a straight line—was selected for convenience rather than for substantive reasons.

In the case of Figure 3, however, where there are now four consecutive datapoints on the observed growth-trajectory plot, it is possible to make a more informed choice of individual growth model. In practice, if there were empirical or theoretical evidence to support the application of a nonlinear individual growth model, then such a model could have been chosen to represent each of the individual growth records. Section IV presents a variety of more sophisticated mathematical functions that have been used to represent individual growth, and Section V outlines how these models may be used in practice. Thus, if multiwave data are available on each subject then a more realistic empirical representation of individual growth becomes feasible. The utilization of mathematical growth modeling in the representation of individual growth is at the heart of recent methodological advances in the measurement of growth and change.

However, in the current discussion a linear individual growth model is appropriate, and therefore a straight-line trajectory has been fitted to the observed growth record of subject #21 by simple linear regression and is displayed in Figure 3 as a broken line. Notice that, in general, this subject shows improvement in "opposites-naming skill" with each successive measurement occasion. Similar trajectories have been obtained for all subjects in the sample and the entire set of observed growth trajectories is displayed in Figure 4. As with Figure 2, attributions of individual growth and individual differences in growth are easy to make from Figure 4. The sample variance of the fitted slopes (over persons) is 164.3, representing moderate interindividual differences in growth. The sample correlation between the covariate and the fitted
FIGURE 3: Observed growth record of subject #21. A plot of observed score versus occasion of measurement for subject #21 in the sample of four-wave growth data in Exhibit 2. The broken line indicates a straight-line growth trajectory fitted to the observed growth record by the simple linear regression of observed score on occasion of measurement.

FIGURE 4: Collection of fitted (four-wave) individual growth trajectories. A plot of fitted straight-line growth trajectories for each of the 35 individuals in the sample of four-wave growth data in Exhibit 2. The fitted trajectories were obtained by the simple linear regression of observed score on occasion of measurement, for each subject in the sample.
individual growth rates is .422, suggesting the existence of a systematic relationship between the covariate and individual growth.

Choosing a Perspective

In addition to introducing the data-example that forms a basis for the ensuing discussion, this section of the chapter has had a single crucially important point to make: when investigations of individual growth are the principal issue of empirical concern then certain types of display are more appropriate than others. Choosing an appropriate display is tantamount to adopting a particular mindset. Unfortunately, the very way in which raw datasets are typically assembled in computers makes it relatively easy for the empirical researcher to obtain between-wave scatterplots, correlations, and regressions, but it is much less easy to model individual growth and to examine interindividual differences in growth.

From a statistical viewpoint, panel data can be displayed and summarized perfectly adequately by either between-wave scatterplots or individual growth-trajectory plots. However, from a substantive viewpoint, there are clear conceptual differences between the two perspectives. In particular, lucid interpretations of individual growth over time, and of interindividual differences in growth, are extremely difficult to make from between-wave scatterplots. When multiwave data are available, the latter approach divides the individual data streams into two-wave chunks and presents multiple scatterplots in each of which two key pieces of growth information on each individual are represented by a single datapoint. This is not the case for a growth-trajectory plot in which the individual growth records retain their integrity, and which addresses the dual substantive concerns of individual growth and group heterogeneity in growth directly.

Attributions of stability. Furthermore, the between-wave scatterplot perspective has engendered considerable confusion concerning the issues of stability and validity. Rather than being the intended summary of the individual observed growth of the 35 subjects in the sample, the scatterplot of Figure 1 presents a view of final observed status conditional on initial observed status. Furthermore, the eye focuses naturally on the accompanying fitted regression line, and the reader is apt to summarize by concluding that individuals who "do well" initially will continue to "do well" subsequently. The magnitude of the corresponding estimated between-wave correlation (.804) suggests that the observed rank-order of the subjects in the sample has not changed substantially from occasion #1 to occasion #2. It is tempting to state that the between-subject rank-order on observed status is "stable" from occasion #1 to occasion #2. If the magnitude of the estimated correlation had been low, we would have inferred that the empirical between-subject rank-order had been modified considerably between the two occasions of measurement. We
would have been tempted to claim that the between-subject rank-order on observed status was "unstable" over time.

Rogosa, Fodden, and Willett (1984) have pointed out that the use of between-wave correlations as indices of stability may be very misleading. These authors stress that particular care should be taken to ensure that indices of the stability of interindividual rank-order are not incorrectly interpreted as indices of the stability of a psychological construct over time. Despite this warning, there is an unfortunate tendency on the part of empirical researchers to interpret the temporal maintenance of between-subject rank-order on observed status as a confirmation that the same psychological construct is being measured on subsequent occasions, that the measure of observed status has remained equitable over the occasions of measurement (see Bereiter, 1963; Bond, 1979; Linn & Slinte, 1977). These interpretations falsely argue that, if subjects do not maintain the same rank-order from occasion to occasion, then the outcome being measured must somehow be different on the two occasions!

That this may not be the case is patently obvious from an inspection of the collection of observed growth trajectories in Figure 4. The failure of the between-subject rank-order to be maintained over time may simply be the consequence of interindividual differences in the rates of growth. It is only when all the subjects in the sample are growing at approximately the same rate (i.e., when all the individual growth trajectories are more or less parallel), that the between-wave correlation of status will be high. If different subjects in the sample have different rates of growth then the individual growth trajectories will not be parallel and subjects will exchange places in the rank-order. The observed measure itself may be perfectly valid but, as a natural consequence of interindividual differences in growth, the between-wave correlation is reduced. Thus, incorrect interpretations of the between-wave correlation of observed status confound issues of construct validity, instrument equatability over time, and interindividual differences in growth. This has important ramifications for the discussion in Section III.

III. MEASURING INDIVIDUAL CHANGE WITH OBSERVATIONS AT TWO TIMEPOINTS

Rather than conceptualizing change as a process that is continuous over time, the authors of the first four quotations cited at the head of this chapter conceptualize change as an increment accrued between the beginning and the end of a particular period of growth. This orientation is common throughout both the methodological and the empirical research literatures in education. Researchers who adopt this orientation ignore the potentially interesting features of growth that occur continuously throughout the duration of the investigation. Instead, they simply observe the status of each subject in the sample at the beginning and end of the study. In the first phase of the growth analysis,
Willett: Measurement of Change

these pairs of observations are then used to construct a two-wave measure of change for each individual. Subsequently, in the second phase of the analysis interindividual differences in the two-wave measure of change are associated with interindividual variation in selected covariates.

A simple two-wave measure of individual change can be obtained by subtracting the initial observed status from the final observed status for each individual. Such a measure is known as a difference, change, or gain score. It is an intuitively reasonable measure of individual change that is easy to compute. For many years, the difference score has been (incorrectly) perceived as unreliable, invalid, and unfair. However, rather than taking the seemingly obvious step of simply collecting more waves of data to obtain an improved measure of change, a variety of "fix-ups" of the two-wave measure have been proposed. This has led to the creation of several modified two-wave measures of individual change. Recently, some of these latter strategies also have come under attack and they have been shown to be inadequate measures of individual growth. Section III reviews and reexamines these issues from an individual growth-trajectory perspective.

**Distinctions, Assumptions, and Notation**

In Section II of this paper, a variety of figures and plots that can be used to display empirical growth data were examined. In all of these exploratory displays, it was the empirical observations on the educational, psychological, or social status of a sample of subjects that were being displayed. No comments were made concerning the intrinsic fallibility of the empirical observations, no questions were asked about underlying traits that might be driving the observed growth processes, nor was the larger population from which the sample had been drawn explicitly defined. Issues of inference and generalizability were not discussed.

**Distinguishing Between True and Observed Status**

When an individual's status is observed or measured with some test or instrument the influence of measurement error will act to disturb the observation. Thus, in any empirical measurement, it is not the underlying true status of the individual that is being recorded by the investigator but a combination of the subject's true status and the accompanying measurement error. An individual's observed status is a fallible measure of his or her true status. Furthermore, it is not relationships among the fallible observed measures that are of interest to the investigator, but relationships among the true underlying variables. In the measurement of individual growth, we are not interested in seeking out systematic interindividual differences in observed growth but are trying to locate interindividual differences in true growth that are systematically related to selected covariates. Our measures of observed status provide
us with a lens through which we hope to discern the true nature of individual growth and change.

Thus, in common with classical test theory (Lord & Novick, 1968), we will assume throughout the remainder of this chapter that an individual's status observed by some instrument, test, or other measure is a linear combination of two independent components: a systematic ('true status') component, and a stochastic ('error') component that obeys some probability law. Thus, our basic measurement model will be:

\[ X_{ip} = \xi_p(t_i) + \epsilon_{ip} \]  

(1)

where the subscript \( i \) denotes the occasion of measurement, \( t_i \) is the time at which the \( i \)th occasion of measurement occurred, and subscript \( p \) indicates which particular person in the population is being represented. The symbol \( \xi \) denotes true status, and the parenthetical inclusion of the time at which the \( i \)th measurement occurred is intended to indicate that the subject is growing and therefore the true status is a function of time. Thus, notationally, \( \xi_p(t) \) represents a function describing the true status of individual \( p \) at time \( t \). The variable \( X \) is a fallible measure of \( \xi \) that has been obtained in the presence of measurement error, \( \epsilon \). We will assume furthermore that the measurement errors \( \epsilon_{ip} \) are drawn independently from identical normal distributions with zero mean and a constant variance, \( \sigma^2_\epsilon \), for all \( i \) and \( p \).

In this section, two-wave panel data are assumed to be available on a sample of \( n \) individuals and therefore datum \( X_{ip} \) is a fallible observation on the \( p \)th individual at time \( t_i \), where observations are obtained at discrete times \( t_i \) and \( t_j \). Thus, for the purposes of the presentation in this section, the "initial" observed status of individual \( p \) is given by

\[ X_{ip} = \xi_p(t_i) + \epsilon_{ip} \]  

(2)

and the "final" observed status is given by

\[ X_{jp} = \xi_p(t_j) + \epsilon_{jp} \]  

(3)

Reliability

A parameter that enjoys considerable popularity among psychometricians and empirical researchers in education is the reliability of a measure, score, or statistic. The rationale for the notion of reliability has its origins in the fact that:

Two sets of measurements of the same features of the same individuals will never exactly duplicate each other. . . . The fact that repeated sets of measurements never exactly duplicate one another is what is meant by unreliability. At the same time, however, repeated measure-
ments of a series of objects or individuals will ordinarily show some consistency. . . . This tendency toward consistency from one set of measurements to another is called reliability (Stanley, 1971, p. 356).

There is an obvious and important need to obtain measurements that are highly reliable. An unreliable measure may offer a very different value of the same trait on replication of the measurement for each individual, the ordering of the individuals within a group will therefore differ, and the obtained ranks will be inconsistent over replications of the measurement. Thus, reliability is "a measure of interindividual differentiation and can only be defined over a group or population" (Rogosa et al., 1982, p. 730).

Reliability and measurement error are intimately connected (in fact, it may be more appropriate to say "intimately confounded"). If the random errors of measurement are large compared to the interindividual variation in true score, then the measurement of a trait will not be consistent over replications of the measurement and reliability will be low. If the errors of measurement are small compared to the interindividual variation in true score, the reliability of the measure will be high. However, extreme caution should be exercised when the values of particular reliability coefficients are being interpreted. In the case of an instrument demonstrating low reliability, for instance, it is impossible to judge from the magnitude of the reliability alone whether it is the influence of measurement error or the lack of interindividual heterogeneity in true score that is responsible for the reduced reliability. This confounding is particularly important when the reliability of a growth measure is being judged (Rogosa et al., 1982; Rogosa & Willett, 1983).

Estimating the reliability of a measure "reduces to a determination of how much of the [observed] variation . . . is due to certain systematic differences among the individuals in the group and how much to other sources of variation that are considered, for particular purposes, errors of measurement" (Stanley, 1971, p. 359). Thus, the reliability coefficient is defined in terms of the observed true and error variances of the measure in question. For an attribute being evaluated by a test, for instance, the population reliability of the test is "that proportion of the variance in test scores that is due to true differences within that particular population of individuals" (Stanley, 1971, p. 362, italics added). Being a proportion, reliability takes on values in the range 0 to 1. The population reliabilities of the initial and final measures of observed status in Equations 2 and 3 are:

\[
\rho(X_1) = \frac{\sigma^2_{x_1}}{\sigma^2_{t_1}},
\]

(4)

\[
\rho(X_2) = \frac{\sigma^2_{x_2}}{\sigma^2_{t_2}},
\]

(5)
where $\rho(X_0)$ and $\rho(X_2)$ are the population reliabilities of $X_0$ and $X_2$, and the variances of the true and observed scores at $t_1$ and $t_2$ in the numerators and denominators of Equations 4 and 5 are obtained as expectations over persons in the population. A variety of different strategies are available for the empirical estimation of reliability (see, for instance, Lord & Novick, 1968).

**Measuring Individual Change with the Difference Score**

A simple and intuitively appealing measure of observed individual change between the initial and final occasions of measurement can easily be obtained. Subtract the initial observed status from the final observed status for each individual:

$$D_p = X_{t_2} - X_{t_1},$$

and $D_p$ is the observed difference (gain or change) score for individual $p$. Thus, from Exhibit 1, the initial and final observed scores of subject #01 are 205 and 217 respectively and the value of the difference score is 12. Similar observed differences can be computed for each of the subjects in the sample, and are presented in Table 2.

Notice that there is a simple relationship between the observed difference score and the underlying true change that has occurred between the initial and final occasions of measurement. Substituting into Equation 6 from Equations 2 and 3 and simplifying, the observed difference score for each individual can be expressed as the sum of the underlying true change, $\Delta_p = \xi_p(t_2) - \xi_p(t_1)$, and the difference between the two measurement errors, $\epsilon_p^* = \epsilon_{2p} - \epsilon_{1p}$. Thus,

$$D_p = \Delta_p + \epsilon_p^*,$$

where, as a consequence of the assumed independence of $\epsilon_{1p}$ and $\epsilon_{2p}$, $\epsilon_p^*$ is normally distributed with zero mean and variance $2\sigma^2_i$. In a growth study, the investigator is interested in ascertaining not the observed growth that has occurred but the true growth of each individual. Thus it is $\Delta_p$, and not $D_p$, that is the real focus of interest.

Taking expectations through Equation 7 (over imaginary replications of the measurement for individual $p$) reveals that “the difference score for individual $p$ is an unbiased estimate of the quantity $\Delta_p$ regardless of the magnitude of the measurement error” (Rogosa et al., 1982, p. 730, notation altered). Despite this obvious and optimal statistical property, authors in the empirical and methodological literatures have criticized the difference score so thoroughly and continuously over the years that investigators have become wary of its use in their research (see, for instance, Bereiter, 1963; Bohrnstedt, 1969; Kes-
sler, 1977; Linn & Slinde, 1977; O'Connor, 1972). These authors bemoan the difference score for its purported unreliability and the fact that it appears to be negatively correlated with initial status. However, recent methodological research has revealed that these deficiencies are perceived rather than actual, imaginary rather than real (Rogosa et al., 1982; Rogosa & Willett, 1983, 1985; see also Zimmerman, Brotohusodo, & Williams, 1981; Zimmerman & Williams, 1982a).

### TABLE 2

Values of a Time-Invariant Covariate and the Observed Difference Scores for the Sample of Two-Wave Data in Exhibit 1

<table>
<thead>
<tr>
<th>Subject ID</th>
<th>Covariate $\alpha_p$</th>
<th>Difference Score $D_p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>137</td>
<td>12</td>
</tr>
<tr>
<td>02</td>
<td>123</td>
<td>24</td>
</tr>
<tr>
<td>03</td>
<td>129</td>
<td>70</td>
</tr>
<tr>
<td>04</td>
<td>125</td>
<td>24</td>
</tr>
<tr>
<td>05</td>
<td>81</td>
<td>30</td>
</tr>
<tr>
<td>06</td>
<td>110</td>
<td>40</td>
</tr>
<tr>
<td>07</td>
<td>99</td>
<td>12</td>
</tr>
<tr>
<td>08</td>
<td>113</td>
<td>35</td>
</tr>
<tr>
<td>09</td>
<td>104</td>
<td>18</td>
</tr>
<tr>
<td>10</td>
<td>96</td>
<td>36</td>
</tr>
<tr>
<td>11</td>
<td>125</td>
<td>45</td>
</tr>
<tr>
<td>12</td>
<td>115</td>
<td>39</td>
</tr>
<tr>
<td>13</td>
<td>109</td>
<td>16</td>
</tr>
<tr>
<td>14</td>
<td>95</td>
<td>-11</td>
</tr>
<tr>
<td>15</td>
<td>118</td>
<td>-24</td>
</tr>
<tr>
<td>16</td>
<td>120</td>
<td>29</td>
</tr>
<tr>
<td>17</td>
<td>118</td>
<td>22</td>
</tr>
<tr>
<td>18</td>
<td>115</td>
<td>58</td>
</tr>
<tr>
<td>19</td>
<td>120</td>
<td>34</td>
</tr>
<tr>
<td>20</td>
<td>118</td>
<td>0</td>
</tr>
<tr>
<td>21</td>
<td>126</td>
<td>63</td>
</tr>
<tr>
<td>22</td>
<td>121</td>
<td>-16</td>
</tr>
<tr>
<td>23</td>
<td>108</td>
<td>3</td>
</tr>
<tr>
<td>24</td>
<td>104</td>
<td>14</td>
</tr>
<tr>
<td>25</td>
<td>118</td>
<td>24</td>
</tr>
<tr>
<td>26</td>
<td>104</td>
<td>39</td>
</tr>
<tr>
<td>27</td>
<td>124</td>
<td>18</td>
</tr>
<tr>
<td>28</td>
<td>130</td>
<td>47</td>
</tr>
<tr>
<td>29</td>
<td>87</td>
<td>20</td>
</tr>
<tr>
<td>30</td>
<td>131</td>
<td>11</td>
</tr>
<tr>
<td>31</td>
<td>109</td>
<td>41</td>
</tr>
<tr>
<td>32</td>
<td>115</td>
<td>45</td>
</tr>
<tr>
<td>33</td>
<td>104</td>
<td>52</td>
</tr>
<tr>
<td>34</td>
<td>110</td>
<td>63</td>
</tr>
<tr>
<td>35</td>
<td>110</td>
<td>31</td>
</tr>
</tbody>
</table>

- **Mean**: 113.5
- **Std. Dev.**: 27.54

- **Mean**: 12.49
- **Std. Dev.**: 21.96
Reliability of the Difference Score

Many authors have criticized the difference score for its purported unreliability, the Lord (1956) quotation at the beginning of this chapter being a particularly well-known example. Some of these authors have argued that the difference score cannot simultaneously be both a reliable and a valid measure of change, that there is an "invalidity/unreliability dilemma" (Bereiter, 1963; Linn & Slinde, 1977). More recently, authors have begun to suggest that there is, in reality, no such dilemma; that high reliability does not necessarily imply low validity, and that "low reliability does not necessarily mean lack of precision" (Rogosa et al., 1982, p. 744). Furthermore, it has become apparent in recent years that the difference score is not necessarily unreliable anyway (Rogosa & Willett, 1983; Zimmerman & Williams, 1982a).

The population reliability of the difference score, \( \rho(D) \), can be defined as the ratio of the total variances of \( \Delta_x \) and \( D_x \) over all individuals in the population:

\[
\rho(D) = \frac{\sigma^2_2}{\sigma^2_1}
\]

Typically, \( \rho(D) \) is reexpressed in terms of observed quantities—the variances of, reliabilities of, and correlation between initial and final observed status. A variety of such reexpressions can be found in the literature, differing according to whether the measurement errors on the two occasions of measurement are assumed to be uncorrelated, of equal variance, and so forth (see, for instance, Stanley, 1967; Williams & Zimmerman, 1977). The casual reader must ensure that the particular assumptions involved are well understood when examining a particular reexpression. Rogosa et al. (1982, Table 2) present a summary of various expressions for \( \rho(D) \) found in the literature, and document the assumptions that underlie each of them. Under the assumptions of this chapter, the reliability of the difference score can be reexpressed as:

\[
\rho(D) = \frac{\sigma^2_{1}, \rho(X_1) + \sigma^2_{2}, \rho(X_2) - 2\sigma_{X_1}, \sigma_{X_2}, \rho_{X_1X_2}}{\sigma^2_{1} + \sigma^2_{2} - 2\sigma_{X_1}, \sigma_{X_2}, \rho_{X_1X_2}}
\]

where previously introduced notation and definitions have been used, and \( \rho_{X_1X_2} \) is the correlation of initial and final observed status over all individuals in the population.

Equation 9 and its kin have been used in the psychometric literature to demonstrate that "differences between scores tend to be much more unreliable than the scores themselves" (Lord, 1956, p. 429). Note that the term containing the between-wave correlation of initial and final status is subtracted in both the numerator and the denominator. Because the variances in the first
two terms of the numerator are multiplied by their respective reliabilities and are therefore smaller than the equivalent terms in the denominator, the subtraction of the term containing the between-wave correlation in both the numerator and the denominator ensures that when $\rho_{x_1x_2}$ is large and positive then $\rho(D)$ will be low. And, as the correlation of initial and final status is frequently misinterpreted as an index of construct validity, authors are apt to report that the difference score cannot be both reliable and valid simultaneously.

However, as discussed earlier, when subjects are growing it is a mistake to interpret $\rho_{x_1x_2}$ as an index of construct validity. Rather, it is simply an index of the extent to which interindividual rank-order on observed status is maintained over time. If different individuals are growing at different rates, then their trajectories may crisscross and their ranking on observed status will vary naturally as time goes by. This fluctuation in interindividual rank-order (and consequent reduction in $\rho_{x_1x_2}$) is entirely unrelated to the construct validity of the instrument used to measure X. In fact, even with a measure that is perfectly valid (and which maintains this validity over time), it is certainly possible for $\rho_{x_1x_2}$ to be zero, or even negative, when there are considerable interindividual differences in growth.

Misguided by these misinterpretations of validity, authors in the psychometric literature have traditionally examined the reliability of the difference score for situations in which $\rho_{x_1x_2}$ is very close to 1, and they have therefore found its value to be low (see Kessler, 1977, Table 1; Linn & Slinde, 1977, Table 1; Stanley, 1971, Table 13.2). Other authors have shown that, in perfectly reasonable situations, the reliability of the difference score can be quite respectable (Rogosa et al., 1982; Rogosa & Willett, 1983; Zimmerman & Williams, 1982a). In fact, to paraphrase the quote by Lord at the head of this chapter, when interindividual differences in growth are large, it is entirely possible for the reliability of the difference score to be greater than the “reliabilities of the scores themselves.” This should be of considerable comfort to those among us who subscribe to the conventional wisdom that more data means more information and, consequently, a stronger analysis.

An alternative, and more comprehensible, formulation of the reliability of the difference score is:

$$\rho(D) = \frac{\sigma_{\Delta x}^2}{\sigma_{\Delta x}^2 + 2\sigma_{\varepsilon}^2}, \hspace{1cm} (10)$$

where an expression for $\sigma_{\varepsilon}^2$ in the denominator of Equation 8 has been obtained from Equation 7. In Equation 10, the reliability of the difference score depends upon only two quantities: the population variance of true change $\sigma_{\Delta x}^2$.
(a parameter describing the population interindividual heterogeneity in growth), and the measurement error variance \( \sigma^2 \). Figure 5 displays the reliability of the difference score plotted as a function of the population interindividual heterogeneity in true growth, at different levels of measurement error variance. Notice that difference score reliability increases monotonically as individual differences in true change increase, whatever the measurement error variance. And that, for a population in which the heterogeneity in true growth is fixed, the difference score reliability decreases as the measurement error variance increases.

By inspection of Equation 10, the dependence of difference score reliability on interindividual heterogeneity in true growth is revealed. When interindividual differences in true growth are large (i.e., when \( \sigma^2 \) is large and the individual true score growth trajectories are very different from each other), the true growth of different individuals is easily distinguishable in terms of observed growth and consequently, the reliability of the difference score is high. Thus, the greater are the individual differences in true growth, the greater is the reliability of the difference score. When there are no individual differences in true growth to detect (i.e., when \( \sigma^2 \) is equal to zero and all the individual true score growth trajectories are parallel), the reliability of the difference score can only be zero, regardless of the precision with which measurement has been carried out. Thus,

The crucial message is that low reliability does not necessarily imply lack of precision. Although individual differences in growth are necessary for high reliability, the absence of such differences does not preclude meaningful assessment of individual change (Rogosa et al., 1982, p. 731).

**Estimating the reliability of the difference score.** Under the current assumptions on the measurement errors, an estimator of \( \rho(D) \) can be obtained by substituting estimates of appropriate sample variances, reliabilities, and correlation into Equation 9:

\[
\hat{\rho}(D) = \frac{\hat{\sigma}^2_{X_1} \hat{\rho}(X_1) + \hat{\sigma}^2_{X_2} \hat{\rho}(X_2) - 2\hat{\sigma}_{X_1} \hat{\sigma}_{X_2} \hat{\rho}_{X_1X_2}}{\hat{\sigma}^2_{X_1} + \hat{\sigma}^2_{X_2} - 2\hat{\sigma}_{X_1} \hat{\sigma}_{X_2} \hat{\rho}_{X_1X_2}}
\]

(11)

Not unexpectedly, given the relatively low heterogeneity among the observed growth trajectories in Figure 2, the estimated reliability of the difference score for the two-wave data in Exhibit 1 is low (\( \hat{\rho}(D) = .226 \), where estimates of \( \hat{\sigma}^2_{X_1}, \hat{\sigma}^2_{X_2}, \hat{\rho}(X_1), \hat{\rho}(X_2) \) and \( \rho_{X_1X_2} \) have been obtained from Exhibit 1 and Table 1).

Notice that, in order to use the estimator in Equation 11, it is *not sufficient* simply to have available measures of observed status on each subject in the
sample on each of the occasions of measurement. In addition to the sample variances and correlation provided by the two waves of growth data, the investigator must also have available two supplementary pieces of information: the estimated reliability of observed status on each of the two occasions of measurement. Either this supplementary information must be obtained externally to the growth investigation (i.e., from a test manual or a previous empirical reliability study), or duplicate measurements of observed status must be made on each subject at each of the two time points to permit the in situ estimation of the required reliabilities. As will be shown later in this section, a good estimate of the reliability of the difference score is a crucial component of any subsequent data analysis of individual growth. Thus, although the difference score is an appealing and unbiased measure of individual growth, the fact that additional external information is required in order for it to be utilized in a data analysis serves to emphasize that investigations of growth that rely upon a two-wave design are particularly weak approaches to the study of individual change.6

Using the Difference Score to Detect Systematic Interindividual Differences in Growth

As discussed in Section II, questions of systematic interindividual differences in change naturally arise in the empirical study of growth. Thus, true
change may differ from person to person in a way that is related to some background characteristic or covariate. One purpose of studying correlates of growth is to detect this systematic variation.

It is useful to distinguish between two different types of correlate of individual growth: true initial status, \( \xi(t_i) \), or a background characteristic of the individual or of the environment, \( \omega \). This distinction has important ramifications here because initial status intrinsically depends upon the choice of \( t_i \), whereas the variable \( \omega \) can be time-invariant. In this chapter, \( \omega \) is assumed to be measured without error and to be time-invariant. In addition, for simplicity, only a single background characteristic is considered although multiple correlates could easily be accommodated.

**Correlation with initial status.** Individual growth and individual status must necessarily be related to one another. Subjects who are growing very rapidly will automatically have a higher level of status on a subsequent occasion when compared to another subject who is growing less rapidly. An individual's current status is a consequence of the individual's growth history. Previous growth determines current status, current status mediates the effect of future growth. Consequently, in a population of individuals who are growing, it would be extremely unlikely to find that growth and status are unrelated.

Numerous empirical studies have investigated the correlation between growth in academic achievement and initial academic status (e.g., Thorndike, 1966; Werts & Hilton, 1977). Bloom (1964) provided a compendium of such results. However, the reader must exercise great caution in interpreting this research literature because, providing nonzero growth is occurring, the initial, or "starting," status of each individual necessarily varies with the choice of \( t' \).

Even when individuals are growing linearly with time (the simplest functional form for individual growth), for every occasion that is declared to be "initial" there will be a different correlation between growth and the "starting" status. Inspection of Figure 4 supports this claim. By sketching vertical lines at different times on the horizontal axis, the reader will note that between-person rank-order on status can be quite different on different occasions as a natural consequence of the heterogeneity in growth, whereas the rank-order on slope remains fixed (the slopes of straight-line growth trajectories being directly proportional to \( \Delta \)). Thus, the correlation between growth and status cannot remain constant for different choices of the time at which initial status is measured.

In fact, Rogosa and Willett (1985) have demonstrated that, for populations in which individuals are growing (either linearly or nonlinearly), the correlation of growth and initial status bears a smooth mathematical relationship to the particular "initial" time that has been specified. For instance, in a collection of subjects growing linearly with heterogeneous individual rates of growth, the correlation between true growth rate and true initial status is a
monotonically increasing function of the time at which initial status is defined. Consequently, in any such collection of individual growth trajectories, there is necessarily only one occasion on which the correlation of true growth rate and true status can be zero. For nonlinear growth, the functional relationship between $\rho_{t_{1},t_{A}}$ and $t_{i}$ is more complex but similar findings apply (Rogosa & Willett, 1985).

The implication of this finding is that, unless there is some specific value of $t_{i}$, which can, for substantive reasons, be declared the initial time, then there can be no single unique value of the correlation of growth and initial status. Therefore, it makes no sense to obtain a single estimate of this correlation. It is meaningless to ask, What is the correlation between growth and initial status? Rather, the question must be more clearly specified. For example, in the context of linear growth in reading skills, it might be more appropriate to frame the question, What is the relationship between growth in reading achievement and reading achievement at the beginning of third grade? At the beginning of fourth grade? And so on. Thus, a sequence of estimates should be obtained for a series of “initial” times in order to examine the dependency of the correlation on choice of time of initial status. Furthermore, there are absolutely no grounds for pooling the sequence of estimates into an overall “average” estimate because what is being averaged, in this case, are multiple estimates of quite different quantities.

However, even though the population correlation of true growth and true status can inevitably take on any value between $-1$ and $+1$, the difference score has been criticized simply because it is correlated with initial status (Bohrnstedt, 1969; Cohen & Cohen, 1983; Linn & Slinde, 1977; O'Connor, 1972; Plewis, 1985). Although the intimate connection between growth and status is an irrevocable consequence of individual growth history, these authors are apt to claim that the difference score is an inappropriate and unfair measure of individual growth because it gives “an advantage to persons with certain values of the pretest scores” (Linn & Slinde, 1977, p. 125). At first sight, it seems somewhat inconsiderate to condemn an unbiased measure of individual growth simply because of an intrinsic and natural connection between status and growth.

However, those that condemn the difference score in this way are usually confusing the effect of the natural connection between growth and status with problems arising from the vagueness of statistical estimation. Unfortunately, the population correlation of true initial status and true change $\rho_{t_{0},t_{A}}$ is frequently estimated in practice by computing the sample correlation of observed initial status with the observed difference score, $\hat{\rho}_{x_{0},p}$. This is a terrible mistake because this latter sample correlation is a negatively biased estimate of $\rho_{t_{0},t_{A}}$ due to the occurrence (with opposite sign) of the time-1 measurement errors in both $X_{ip}$ and $D_{ip}$. An equation that illustrates the form of this bias is:
\[ \rho_{X_{1}D} = \rho_{(01)\Delta} \sqrt{\rho(X_{1})\rho(D)} - \frac{\sigma_{x_{1}}}{\sigma_{D}} \left[ 1 - \rho(X_{1}) \right] \]  

(12)

(based on Rogosa et al., 1982, Equation 11). Inspection of the right-hand side of Equation 12 reveals that, in the creation of \( \rho_{X_{1}D} \), not only is the magnitude of \( \rho_{(01)\Delta} \) being reduced due to multiplication by a fractional quantity (the square root of the product of a pair of reliabilities) but a positive quantity is also being subtracted from it. Consequently, even if the population correlation \( \rho_{(01)\Delta} \) is both positive and moderate in magnitude, investigators correlating \( X_{1}p \) and \( D_{p} \) in the sample are likely to obtain estimates that are almost zero or negative (an illustration of this has been provided by R.L. Thorndike, 1966). The failure of the sample correlation of \( X_{1}p \) and \( D_{p} \) to be a good estimator of \( \rho_{(01)\Delta} \) is no reason to reject the difference score as an adequate (and unbiased) measure of individual change: "The correlation \( \rho_{(01)\Delta} \) is an interesting fact of life, whereas the correlation \( \rho_{X_{1}D} \) is an uninteresting artifact of errors of measurement" (Rogosa et al., 1982, p. 735, notation altered and italics added).

The poor statistical properties of the sample correlation of \( X_{1}p \) and \( D_{p} \) as an estimate of \( \rho_{(01)\Delta} \) have been well known for a considerable time (for instance, see E.L. Thorndike, 1924), and several authors have proposed techniques for eliminating the bias (Thomson, 1924, 1925; Zieve, 1940). In practice, an unbiased estimate of the population correlation of true growth \( \Delta \) and true initial status \( \xi_{i}(t_{0}) \) over persons can be obtained by substituting appropriate sample estimates of \( \rho_{X_{1}D}, \rho(X_{1}), \rho(D), \sigma_{x_{1}} \) and \( \sigma_{D} \) into Equation 12 and solving for \( \hat{\rho}_{(01)\Delta} \):

\[ \hat{\rho}_{(01)\Delta} = \frac{\hat{\rho}_{X_{1}D} + \frac{\hat{\sigma}_{x_{1}}}{\hat{\sigma}_{D}} \left[ 1 - \hat{\rho}(X_{1}) \right]}{\sqrt{\hat{\rho}(X_{1})\hat{\rho}(D)}} \]  

(13)

(Zieve, 1940, Equation 5). For the sample of two-wave growth data in Exhibit 1, where \( \hat{\rho}_{X_{1}D} \) is \( -0.419 \), an estimate of the correlation of true change and the true initial status obtained in this way is \( -0.300 \).

**Correlation with other covariates.** When the selected covariate is time-invariant, it is meaningful to enquire as to the relationship between the covariate and the individual growth. Examples of background characteristics
used as correlates of individual change in research on intelligence include measures of personality and parental behavior (McCall, Appelbaum, & Hogarty, 1973), socioeconomic status (Rees & Palmer, 1970), home background variables (Harnquist, 1968), environmental measures (Bloom, 1964), and measures of personality and behavior (Bayley, 1968). Wohlwill (1980) provided an overview of such studies. In addition, the "process-product" paradigm in research on teaching provides many examples of correlates of student learning using measures of teachers' classroom behavior as covariate (for instance, Berliner, 1976; Borich, 1979; Medley, 1979; Veldman & Brophy, 1974).7

Thus, for instance, an investigator may wish to examine the relationship between the socioeconomic status of the home and individual growth in academic achievement at school. Or between personality characteristics of the individual and growth in social integration. How are such questions answered empirically? Typically (and most unfortunately!) the investigator will obtain two waves of growth information, compute an observed difference score for each subject in the sample and then obtain the sample correlation between this observed measure of change and the selected covariate. In the case of the sample of two-wave growth data presented in Exhibit 1, .101 is the magnitude of this sample correlation.

However, this rudimentary and obvious strategy suffers from a serious flaw. Recall that the empirical investigation is fundamentally concerned with the examination of systematic interindividual differences in true growth. Therefore, interest should be focused on the population correlation of true change and the selected covariate, \( \rho_{xw} \), whereas the sample correlation purporting to be a summary of systematic interindividual differences in growth in the previous paragraph is an estimate of the population correlation of the observed difference score and the selected covariate, \( \rho_{xw} \). And, of course, this latter correlation is of little interest to the investigator because the presence of measurement error in the difference score attenuates the obtained correlation, leading to an estimate whose magnitude is too small. (Recall that, in the simulated dataset, \( \rho_{xw} \) is .438 whereas the simple estimate obtained above is .101—a very large discrepancy). Thus, the systematic interindividual differences in true growth will be undervalued during estimation.

To counteract the effects of the fallible measurement of individual growth, the data analyst would usually seek to compensate for the influence of the measurement error during the estimation process. Typically, such a compensation demands knowledge of the population reliability of the difference score because, under the current assumptions on the measurement error, the relationship between \( \rho_{xw} \) and \( \rho_{xw} \) is given by:

\[
\rho_{xw} = \frac{\rho_{xw}}{\sqrt{1 - \rho^2}}
\]

(14)
(Notice that $\hat{\rho}_{m}$ will always have a larger magnitude than $\rho_{m}$ due to division of this latter correlation by the square root of the reliability of the difference score, a quantity that necessarily lies between 0 and 1.) Although the reliability of the difference score is not usually known to the investigator it can be estimated, as discussed earlier. Replacing the population correlation $\rho_{m}$ and the population reliability $\rho(D)$ in Equation 14 by sample estimates provides a disattenuated estimate of $\rho_{ma}$, a summary of the population systematic inter-individual differences in true change. For the sample of two-wave growth data presented in Exhibit 1, an estimate of $\rho_{ma}$ obtained in this way is .212.

If the disattenuation of $\hat{\rho}_{ma}$ is to succeed then a high quality estimate of $\rho(D)$ must be available. As noted earlier, when only two waves of growth data have been collected, the investigator has no empirical basis on which to estimate $\rho(D)$ but must necessarily rely upon an estimate obtained from external information through Equation 11. As the nature of this external information is often dubious at best (i.e., lower-bound test reliabilities based on internal-consistency estimation strategies applied in other populations), the investigator may be well advised to mistrust the outcomes of the disattenuation.

In addition, when the reliability of the difference score is small (as is often the case when only two waves of growth data have been collected), the effect of the disattenuation is both large and highly sensitive to small differences in the magnitude of $\hat{\rho}(D)$. Thus, using one particular set of estimated test reliabilities in the disattenuation process, rather than some other set, can lead to vastly different estimates of $\rho_{ma}$, and it is highly possible that the estimated correlation will be gloriously incorrect. (Notice that, in the example cited here, while the disattenuated estimate of .212 is closer to the population value of .438 than the earlier estimate it is still not that good!). Therefore, it is particularly important to caution the reader against the indiscriminate application of the disattenuated estimator suggested beneath Equation 14. The estimator has been referenced here for the sake of completeness, rather than as an appropriate estimator for empirical application. When there are only two waves of observed data available, it may be better to simply regard the obtained sample correlation of $X_{t}$ and $D$ as a ballpark "lower bound" estimate of $\rho_{m}$. Indeed, for a more robust and meaningful examination of individual growth, the investigator is strongly advised to avoid the two-wave design completely and to apply a multiwave methodology instead (see Sections IV and V).

**Growth in Standardized Scores**

Authors in the psychometric literature are frequently prepared to make a variety of "simplifying" assumptions. Principal among these self-imposed
constraints is the assumption of “dynamic equilibrium” (Lord, 1963). Under this assumption, the population variances of initial and final observed status (over persons) are set equal (i.e., \( \sigma_i^2 \) is assumed to be equal to \( \sigma_{i_0}^2 \)). This assumption, in conjunction with the usual psychometric assumptions on the measurement error, also implies that the population variances of true status are constrained to be equal (i.e., \( \sigma_{\theta_i}^2 \) must be equal to \( \sigma_{\theta_{i_0}}^2 \)), and that the population reliabilities of the observed scores must be the same on the two occasions of measurement (i.e., \( \rho(X_i) \) must be equal to \( \rho(X_{i_0}) \)).

Typically, these constraints are imposed upon the system of individual growth trajectories either explicitly to “simplify” the interpretation of complex algebraic formulae rather than for any substantive purpose (see, for instance, the discussion of the reliability of the difference score in Kessler, 1977, Equation 13 and Table 1, or Linn & Slinde, 1977, Table 1), or implicitly in the discussion of other phenomena (such as “regression to the mean” in Furby, 1973; see also Rogosa et al., 1982, Rogosa & Willett, 1985). But is the “simplifying” assumption of dynamic equilibrium merely a mathematical convenience, or does it have wider and more serious ramifications? In particular, is it realistic to have made such an assumption in a population of subjects who are growing? The answers to these questions become apparent when one considers the mathematical and statistical consequences of the assumption.

One important and pervasive ramification of the assumption of dynamic equilibrium is that, under this constraint, the correlation of true change and true initial status cannot be positive (Rogosa et al., 1982, Equation 10). This is an unrealistic finding that does not necessarily apply in practice (see, for instance, Thorndike, 1966).

Furthermore, as no special significance is usually attributed to the particular choice of initial time, the imposition of the assumption of dynamic equilibrium is tantamount to constraining the population variance of observed status (and of true status) to be homoscedastic for all of time. That this is an entirely unrealistic state of affairs is obvious even at the most casual glance. If the assumption were realistic, then inspection of typical samples of individual growth trajectories should display evidence of the equality of variability and, in particular, we would not expect the range of different values of individual status to be markedly different on different occasions. However, inspection of the sample of observed growth trajectories in Figure 4 suggests that, when individuals are growing heterogeneously, the individual growth trajectories cross and overlap as time passes causing the standard deviation of observed (and true) status to depend on the time at which status is defined. For instance, in a population of individuals who are growing linearly with time, those who are growing the most rapidly will eventually “rise to the top” and will continue to leave their more slowly growing colleagues “behind.”
Thus, as a natural consequence of growth, individuals will eventually "fan out" over time and the standard deviation of true (and observed) status will increase dramatically as time passes. In fact, except for the extremely unlikely situation in which all individual growth trajectories are parallel, it is impossible for the population variance of true (and observed) status to be constant over time (Rogosa & Willett, 1985; Willett, 1985).

Of course, when only two waves of data have been gathered on each subject in the sample and these two occasions of measurement are very close together in time, one would expect the sample variances of observed status to be very similar because the intrinsic heterogeneity in individual growth will not have had time to act. Furthermore, investigations of growth that make use of measuring instruments whose obtained scores are implicitly standardized to some national norm (such as IQ tests) cannot demonstrate dramatic changes in observed score variance over time. Perhaps it is such misguided pieces of empirical evidence that have persuaded some investigators that the standardization of observed status to a fixed standard deviation over time is an appropriate data transformation prior to the analysis of individual growth. However, such standardization constitutes a completely artificial and unrealistic restructuring of interindividual heterogeneity in growth, and "the constraint that has been put on the score scale assures distorted results" (Thordike, 1966, p. 126, italics added).

Certainly, researchers who are planning to investigate individual growth would be well advised to avoid any such artificial recalibration of the natural metric along which status has been defined and measured. It may be that, in the measurement of some attributes (e.g., vocabulary), the raw score is the most appropriate measure of observed individual status. This is for the investigator to decide. However, the advice to avoid the rampant and careless standardization of test scores on subsequent occasions of measurement is not intended to suggest that the investigator should avoid instruments whose metric has been carefully scaled (perhaps by the inclusion of common items) to ensure that the measures of observed status are equatable over time. Indeed, in this respect, it may be that the newer statistical technologies of Item Response Theory (Bock, 1976; Lord, 1980) have much to offer the empirical investigation of individual growth.

**Improving the Difference Score as an Estimate of True Change**

Acknowledging implicitly that two waves of data provide only minimal information about individual growth, several authors have suggested modifications of the difference score in order to better estimate individual true change $\Delta$. The proposed "improvements" to the observed difference score have the effect of incorporating group-level information into the estimates of individual change. Thus Webster and Bereiter (1963), based on earlier work by Kelley (1948), have proposed a reliability-weighted measure of individual change, and Lord (1956) and McNemar (1958) have suggested a regression-
based measure of "estimated true change" (see also Cronbach & Furby, 1970; Davis, 1964).

Rogosa et al. (1982, pp. 735–738) discuss the reliability-weighted measure and the estimated true change measure in detail and state that, under the assumptions of the current chapter, these modified measures are identical. For the purposes of this chapter then, a modified measure of change for the \(p\)th individual is simply a weighted linear combination of the observed difference score for that individual and the average observed difference score for the entire population, with weights being determined by the reliability of the difference score itself. Thus, the modified difference score is given by:

\[
[\rho(D)]D_p + [1 - \rho(D)]\mu_D,
\]

where \(\mu_D\) is the population average of the individual observed difference scores.

Notice that, when the reliability of the difference score is high (i.e., when \(\rho(D)\) is close to one), the modified difference score for the \(p\)th individual is essentially identical to \(D_p\). Alternatively, when the reliability of the difference score is low (i.e., when \(\rho(D)\) is close to zero), the modified difference score for the \(p\)th individual is equal to the population mean \(\mu_D\). For values of \(\rho(D)\) between these limits, the modified difference score for the \(p\)th individual will be an appropriately weighted mixture of individual- and group-level information. Essentially, the weighting scheme places emphasis on those aspects of the growth measurement that are the most trustworthy: favoring the difference score when it is reliable but, otherwise, setting each individual’s growth equal to the average growth for everyone when there is reason to believe that there is little real heterogeneity in individual growth to be detected.

Thus, the weighting in Expression 15 acts to "shrink" each \(D_p\) towards the population mean \(\mu_D\). Therefore, unlike the observed difference score, the modified difference score for the \(p\)th individual is no longer an unbiased estimate of the true change for that individual. Rather, the modified difference score has traded off some of its unbiasedness for an improvement in mean-squared error, and consequently its mean-squared error may be considerably smaller than the corresponding mean-squared error for the difference score (Rogosa et al., 1982, pp. 736, 747). However, because the modified difference score is obtained by applying an identical linear transformation to each \(D_p\), the reliability of the modified difference score is identical to the reliability of the original difference score.\(^8\)

In order to obtain an estimate of true change based on Expression 15, the investigator simply replaces \(\mu_D\) and \(\rho(D)\) by the appropriate sample estimates \(\bar{D}\) and \(\hat{\rho}(D)\). Of course, the quality of the obtained estimates of \(\Delta_p\) will then depend upon both the quality of the original measurement and on the quality of the estimate of \(\rho(D)\) that has been incorporated in the new measure.
Because both $\hat{\rho}(D)$ and $\overline{D}$ have the same fixed value for every member of the sample, the sample correlation of a selected covariate with the modified difference score will be identical to the correlation with the observed difference score. Therefore, any empirical investigation of systematic interindividual differences in growth that is based on the modified difference score will inevitably lead to the same conclusions that were obtained under the earlier difference score approach. Furthermore, since the population reliability of the modified difference score is exactly the same as that of the original difference score, the disattenuation strategy outlined beneath Equation 14 still applies. Thus, other than in the separate representation of individual change (in which the modified difference score shrinks its estimate of individual growth meaningfully toward the sample mean), there appears to be little reason to choose the modified measure over the original difference score. Therefore it may be easier, and it is certainly more direct, to apply $D$, directly in the investigation of systematic interindividual differences in growth rather than recomputing its linearly transformed cousin, the modified difference score.

Nevertheless, although modified difference scores are not being promoted here for use in empirical investigations of systematic heterogeneity in growth, I find myself sympathetic toward the intentions of the creators of these measures of growth. In particular, I agree that additional information must be included in the measurement process, if the measurement of change is to be improved. However, rather than attempting to supplement a simple two-wave measure of change by incorporating group-level information, the empirical researcher would be better advised to design the growth study so that the supplementary information required can be gathered in the form of multiwave data, from the outset. In other words, rather than attempting to "fix up" the difference score "after the fact," the investigator could more appropriately apply the strategies of Sections IV and V.

**Measuring Individual Change with the Residual Change Score**

Largely motivated by a desire to obtain a measure of individual change that is uncorrelated with initial status, the residual change score has been proposed. Much effort and energy has been spent in the psychometric literature detailing the psychometric and statistical properties of the various estimators of the residual change score, and considerable controversy has been aroused. There is disagreement as to exactly what is being estimated, how well it is being estimated, and how the outcomes of the estimation can be interpreted. Among methodologists, at least, the residual change score has now been largely discredited as a measure of individual change. In addition to the many technical and practical problems that arise in its application, there are also issues of logic and substance. A brief discussion of some of these issues is presented here for the sake of completeness; for a more detailed account the reader is referred to Rogosa et al. (1982) and Rogosa and Willett (1985).
The true residual change score for individual $p$ is defined as the residual obtained for that individual in the population linear regression of true final status on true initial status (Cronbach & Furby, 1970; Glass, 1968; Rogosa et al., 1982). The statistical model for this regression is:

$$E[\xi_p(t_2) | \xi_p(t_1)] = \mu_{\xi(t_1)} + \beta_{\xi(t_1)\xi(t_2)} [\xi_p(t_1) - \mu_{\xi(t_1)}]$$  \hspace{1cm} (16)$$

where $\mu_{\xi(t_1)}$ and $\mu_{\xi(t_2)}$ are the population means of true initial and final status respectively (over persons), and $\beta_{\xi(t_1)\xi(t_2)}$ is the associated population regression slope. Therefore, the true residual change score for person $p$ is given by:

$$[E(\xi(t_2) - \xi(t_1))]_p = [\xi_p(t_2) - \mu_{\xi(t_2)}] - \beta_{\xi(t_1)\xi(t_2)} [\xi_p(t_1) - \mu_{\xi(t_1)}].$$  \hspace{1cm} (17)$$

where the use of the notation $[E(\xi(t_2) - \xi(t_1))]_p$ is intended to indicate that the influence of true initial status $\xi_p(t_1)$ has been removed from $\xi_p(t_2)$ by a linear regression adjustment.

Cronbach and Furby have defined the residual change score as "primarily a way of singling out individuals who have changed more (or less) than expected" (1970, p. 74), where "expected" in this context has been defined solely in terms of each individual's true initial status. Logically, one might argue that this use of initial score as the single predictor of growth can only provide a very limited "expectation" of that growth. It is as though we are attempting to evaluate true growth while "correcting" for heterogeneity in true initial status, using only a single prior individual attribute as a basis for the "correction."

Notice that there is no way in which true residual change $[E(\xi(t_2) - \xi(t_1))]_p$ can be considered equivalent to true change $\Delta_p$. Indeed, $[E(\xi(t_2) - \xi(t_1))]_p$ is equal to $\Delta_p$ with the influence of true initial status being removed by a linear regression adjustment. As Cronbach and Furby have noted, "one cannot argue that the residualized score is a 'corrected' measure of gain, since in most studies the portion discarded includes some genuine and important change in the person" (1970, p. 74). In fact, from a substantive perspective, it is clear that "Residual change does not attempt to answer the simple question, How much did person $p$ change on the attribute $\xi$? Instead, residual change addresses a far more difficult and arguably intractable question. . . . How much would person $p$ have changed on $\xi$ if all persons had the same true initial status?" (Rogosa et al., 1982, p. 741).

Estimation of true residual change. Despite these well-documented logical and substantive caveats, empirical researchers have been quick to adopt the concept of the residual change score and have sought practical techniques for estimating it. In one of the earliest strategies, Dubois (1957, 1962) proposed estimating $[E(\xi(t_2) - \xi(t_1))]_p$ by the residual from the sample ordinary least-squares (OLS) regression of observed final status $X_{2p}$ on observed initial status $X_{1p}$.
\[ \hat{R}_p = [X_{2p} - \bar{X}_z] - \hat{\beta}_{sy} [X_{1p} - \bar{X}_y], \]  

(18)

where $\bar{X}_1$ and $\bar{X}_2$ are the sample means of observed initial and final status respectively (over persons), and $\hat{\beta}_{sy}$ is the associated estimated slope from the linear regression of $X_{2p}$ on $X_{1p}$ in the sample.

Individual residual change scores are easy to obtain in practice by applying Equation 18 to a sample of two-wave growth data. The estimation strategy itself has considerable superficial appeal. However, on further examination, it is easy to see that the approach is considerably misguided. In fact, rather than being an estimate of the population true residual change $[\xi(t_1) - \xi(t_2)]$, $\hat{R}_p$ is an estimate of the population observed residual change $[X_{2p} - X_{1p}]$. These two residualized measures being mathematically distinct. Just as both $X_{1p}$ and $X_{2p}$ are fallible measures of $\xi_p(t_1)$ and $\xi_p(t_2)$, then $\hat{R}_p$ is obtained in an errors-in-variables regression of $X_{2p}$ on $X_{1p}$ and is therefore an inefficient and inconsistent estimator of true residual change (Kendall & Stuart, 1961). A variety of modifications to $\hat{R}_p$ have been suggested in an attempt to reduce this bias, and this has led to the so-called “base-free measures of change” (Bond, 1979; Messick, 1981; O’Connor, 1972; Tucker, 1979; Tucker, Damarin, & Messick, 1966).

In addition to the various logical, substantive, and technical inadequacies discussed above (and below), the use of $\hat{R}_p$ suffers from a serious additional drawback because its practical computation is based on the statistical modeling of final observed status as a linear function of initial observed status using ordinary least-squares regression on the sample data. Such fits are notoriously sensitive to the presence of outliers in the dataset and yet it is the “distance” of these very outliers from the fitted line upon which subsequent interest centers! Furthermore, the valid application of ordinary least-squares regression demands predictors that have been measured without error and, in the case of $\hat{R}_p$, the only predictor included in the model is admittedly fallible. Finally, there is no a priori reason to believe that initial and final status (either true or observed) are related by a linear model. In fact it is easy to construct counterexamples in which, although the underlying within-individual true growth is linear, the relationship between initial and final status over individuals at two arbitrarily selected timepoints is decidedly nonlinear. Typically, investigators do not check the assumption of linearity prior to estimating residual change scores. 

**Reliability of the residual change score.** There is much confusion in the technical literature as to what are appropriate expressions for the reliability of the various residual change estimators that have been proposed (see, for in-
stance, Glass, 1968). Different authors favor different expressions, the discrepancies arising as a result of disagreement as to the appropriate algebraic form for the measure of residual change. Compare, for instance, the expression for the reliability of the residual change score cited in Linn and Slinde (1977), Stanley (1971), Zimmerman and Williams (1982b) with that cited in Traub (1967) and O'Connor (1972). Rogosa et al. (1982) summarize and clarify this controversy and claim that "no expression in the literature accurately represents the [reliability] of the residual change score $\hat{\mathbf{R}}_p$" (p. 739, notation altered). Nevertheless, inspection of expressions that do appear in the literature and analysis provided by these latter authors suggests that the reliability of the residual change score is unlikely to be much different from the reliability of the simple difference score itself.

Correlation with initial status. Readers will realize from the earlier discussion that the relationship between change and status is a natural consequence of growth history rather than a problem to be attributed to the difference score. Therefore, it seems somewhat unreasonable to replace the difference score by an alternative measure simply because "the correlation between true change and true initial status . . . is an interesting fact of life" (Rogosa et al., 1982). Nevertheless, it was the misguided conviction that the intrinsic connection between change and initial status was somehow "unfair" because it gave an "advantage" to individuals with certain pretest scores that led to the genesis of the residual change score (e.g., Good, Biddle, & Brophy, 1975, pp. 41-42). It seems ironic then that, whereas $\hat{R}_p$ is uncorrelated with observed initial status in the sample, it is not uncorrelated with true initial status either in the sample or in the population (Tucker, Damarin, & Messick, 1966).

Relationship with other covariates. Investigators interested in the detection of systematic interindividual differences in residual change might typically attempt to estimate parameters describing the population relationship between $[\xi(t_2) - \xi(t_1)]_p$ and a selected covariate $\omega_p$, over persons. Thus, for instance, the population part correlation $\rho_{[\xi(t_2) - \xi(t_1)]_p, \omega_p}$ is often a focus of empirical interest (Tucker et al., 1966). Lord (1958), on the other hand, has proposed that a parameter that more appropriately describes systematic interindividual differences in true residual change is the partial correlation $\rho_{[\xi(t_2) - \xi(t_1)]_p, \omega_p | [\xi(t_2) - \xi(t_1)]_p}$, where the influence of true initial status has been removed from both final status and the covariate by a linear regression adjustment. Other authors have argued that is it the partial regression coefficient $\beta_{[\xi(t_2) - \xi(t_1)] | \omega_p}$, corresponding to this latter partial correlation, that is the crucial parameter of interest (Werts & Linn, 1970).

As has been noted, the true residual change $[\xi(t_2) - \xi(t_1)]_p$ cannot be regarded as a replacement for the true change $\Delta_p$. Therefore, if the focus of interest in a study of systematic interindividual differences in growth is actually the relationship between $\Delta_p$ and the selected covariate (over persons), then it seems
extremely doubtful that estimating any of the three parameters described in the previous paragraph will provide the investigator with the appropriate insight. In fact, Rogosa and Willett (1985) have demonstrated algebraically that the population relationship between true change and the selected covariate is not depicted accurately by either \( p_{i(t_2-i(t_1))} \) or \( \beta_{i(t_2-i(t_1))} \). Furthermore all of these latter three parameters are systematic functions of \( t_i \) and therefore, although the influence of true initial status appears to have been removed from consideration, in fact the empirical findings will not be independent of the time at which initial status is measured.

In this brief presentation of the issues involved in the investigation of systematic interindividual differences in true residual change, problems of estimation have been carefully pushed to one side. However, of course, all of the previously discussed caveats associated with the estimation of true residual change still apply. The possibility of poorly estimating an inappropriate parameter should "signal extreme caution in the use and interpretation of residual change measures" (Rogosa et al., 1982, p. 741, italics added).

IV. REPRESENTING INDIVIDUAL GROWTH

In the first part of Section II, it was emphasized that typical empirical investigations of growth are designed to include linked within-individual and between-individual phases. In the first phase, interest centers upon the growth of each individual separately, and the investigator collects data on each subject in the sample over time. Traditionally, because research questions in education have been framed as pre/post comparisons, only two waves of data are collected. Then, using this individual growth information, the first phase of the growth study concludes with the computation of some observed measure of individual growth (such as the difference score) as an estimate of the true individual growth that has occurred during the period of observation. Subsequently, this within-individual estimate of growth becomes the basis of the second-phase between-subject analysis, in which interindividual differences in growth are systematically associated with interindividual heterogeneity in selected covariates.

However, taking a snapshot of individual status on each of two occasions does not permit the investigator to visualize the intricacies of the underlying individual growth with any great certainty. Although the growth may be proceeding smoothly over time with some complex and substantively interesting trajectory, two waves of data provide no evidence to support conclusions about the shape of that trajectory. Indeed, when only two waves of data are available, the investigator is inclined to adopt the simplest possible mathematical form to represent the individual growth because, in fact, any number of arbitrary curves can be drawn through a given pair of datapoints. Thus each pair of datapoints are simply joined by a straight line, and growth with a
constant rate of change becomes the model of individual growth selected. In
this way, the difference score (as a measure of growth proportional to the
slope of the straight-line growth-trajectory) becomes the default estimate of
the underlying true growth.

Certainly, as a measure of individual change, the difference score is easy
to compute, is intuitively appealing, and is an unbiased estimate of the under-
lying true growth between t₁ and t₂. Furthermore, as discussed in Section III,
it can be both reliable and valid, and it does not suffer from the logical, sub-
stantive, technical, and practical difficulties that beset the estimated residual
change score. However, the difference score does suffer from an important
deficiency.

For instance, examine the sample of observed growth-trajectories displayed
in Figure 2. Consider, for a moment, the observed growth-trajectory of one
of the subjects in the sample, for example, individual #1. Although this
single growth-trajectory was constructed by simply joining together the data-
points X₁₁ and X₂₂ on a plot of observed status versus time, it is in fact the
fitted line that would be obtained by regressing X on t for individual #1.
Similar “regression lines” were “fitted” for each subject in the sample. These
fitted growth-trajectories have estimated slopes (or rates of growth) that are
given by D₁(t₂ − t₁) and are therefore proportional to the observed difference
scores, over all of the subjects in the sample. Just as no self-respecting em-
pirical investigator would consider a regression fit of two datapoints ade-
quate, neither should this same investigator consider the difference score to
be an adequate description of individual growth.

Furthermore, in any linear regression based on only two datapoints, the
fitted line passes exactly through both points. Therefore, no residuals can be
estimated during the fitting process and consequently no appropriate empirically
based estimates of precision and goodness-of-fit can be obtained. These
are also the major failings of the difference score. If an estimate of the sam-
ping variance of the difference score is required in the study of growth, then
an estimate of the measurement error variance must be obtained externally to
the empirical investigation. This usually entails the estimation of the reliabil-
ity of the difference score by substitution of sample moments and reliabilities
of initial and final observed status into Equation 11. In other words, for an
effective empirical investigation of individual growth using the difference
score, more information than simply a knowledge of the observed status of
each individual at t₁ and t₂ is actually required. It is this latter theme that
prevailed the discussion in Section III.

Indeed, to measure individual growth adequately, more information on that
growth in the form of multiwave data is required. When multiwave data are
available on each of the subjects in the sample, the investigator can examine
detailed empirical growth-trajectory plots that summarize the observed
growth of each individual over time. Consequently, a variety of more complex
possibilities can be considered; the investigator can check that the growth is linear, or entertain the possibility that it is indeed nonlinear and perhaps is "flattening off" over time. Inspection of individual growth-trajectory plots permits the investigator to select an appropriate mathematical model to represent the underlying individual growth. Then, rather than simply adopting an ad hoc measure of difference as an estimate of the underlying growth, the investigator is able to fit the newly selected growth model to the observed growth records and subsequent between-individual analyses can be based more appropriately on the results of these fits. The empirical fitting of individual growth models to panel data is the subject of Section V, where it is shown that the collection of multiwave data greatly improves the adequacy of the growth measurement. Furthermore, when a selected growth model has been fitted to each observed growth record, it is possible to construct realistic estimates of the measurement error variance from the regression residuals without needing to rely on external supplementary sources of information (such as estimated test reliabilities, as in the case of the two-wave measurement of growth).

The collection of multiwave data, and the prescription of appropriate mathematical models to describe the complexities of the individual growth, permit the growth to be characterized and summarized with greater validity. Consequently, there is a very real possibility that the quality of the research findings will be enhanced. Furthermore, what is particularly exciting about this growth-modeling approach is that the investigator's substantive base may be able to provide some appropriate theory to underpin the growth. Thus, psychological theories of cognitive growth, or theories of moral or personality development, for example, may suggest an appropriate functional form for the individual growth. It may be that certain types of individual growth are theoretically constrained to rise to asymptote, or that certain conditions of learning engender particular models of cognitive growth. Then, by entertaining a particular growth model, the investigator is able to test hypotheses concerning systematic interindividual differences in growth in general and also perhaps to associate particular features of the growth with particular background characteristics of the individual and the environment. For example, with a growth model that incorporates a smooth approach to asymptote, one set of background characteristics may be related to the (instantaneous) rate of growth, whereas other characteristics may predict the ultimate limits on the growth. The investigator has entered an arena of richer, and more rewarding, research questions. Such refinements are not possible when only two waves of data have been collected.

Earlier we recognized that, when an individual's status is observed with some instrument, test, or other measure, the process of measurement acts to disturb the observation. Thus, observed status is a fallible measure of true status. When an individual is growing, it is as though the underlying true
growth is continuing smoothly and unobserved over time, but periodically the investigator observes the growth with some fallible measuring instrument. In this way the individual’s observed growth record is assembled, and it consists of a chronological series of discrete measurements, each of which is an unknown combination of true status and measurement error. What is of fundamental interest to the investigator, of course, is the underlying continuous true growth trajectory; the multiple entries in the observed growth record are simply a fallible lens through which the true individual growth is viewed. Furthermore, in an investigation of systematic interindividual differences in growth, it is the association between heterogeneity in true growth and interindividual differences in the selected covariates that is of principal research interest and not the corresponding relationships for the observed growth. It is for these reasons that, when we speak of adopting a specific individual growth model here, we are considering the selection of a mathematical model to represent the underlying individual growth in true status and not the growth in the measure of observed status.

In the opening part of Section III, \( \xi_0(t) \) and \( \xi(t) \) were used to represent the true status of the \( p \)th individual at times \( t_1 \) and \( t_2 \), and the usual test-theory measurement model was proposed in Equation 1 to link them with observed status and measurement error. In this section, and subsequently, we will assume that this measurement model remains valid but we will lift the restriction on the number of waves of measurement. Thus, we will assume that \( T \) waves of panel data have been collected and that \( X_p \) and \( \varepsilon_p \) are the observed status and measurement error for the \( p \)th individual at time \( t_i \) (where \( i = 1, \ldots, T \)). Then, Equation 1 defines the relationship among these latter quantities and \( \xi(t) \) represents a smooth mathematical function describing the true status of individual \( p \) at time \( t \). It is the functional form of \( \xi(t) \) that is selected when a mathematical model of individual growth is chosen.

But how can appropriate models be chosen to represent the individual true growth? What mathematical functions are available, and which of them are the most appropriate for modeling educational, social, and psychological growth? In the following section, a few mathematical functions suitable for representing individual true growth are presented and described. In the final portion of this section, an approach for conceptualizing systematic interindividual differences in growth is discussed.

Models for Individual True Growth

The tradition of describing individual growth as a function of time originates largely in the field of human biology and dates back at least to the years 1759 to 1777, when the Count Philibert de Montbeillard kept chronological records of the growth of his son (Goldstein, 1979, Figures 4.1, 4.2; see also Kowalski & Guire, 1974). Mead and Pike (1975, Table 3.1) list a variety of algebraic functions that have been used to model biological growth. Many of
these same functions have also been used to model cognitive and academic growth.

Under each of the proposed growth models, the generic shape of the true growth trajectory is determined by specifying true status as a specific algebraic function of time. What distinguishes the true growth of a particular individual from that of another are the values of the various constants that enter into the general algebraic function. Thus, under the straight-line growth model, the true status of all individuals in a particular population is assumed to be changing linearly with time but different individuals can have different starting points for their growth (i.e., different intercepts) and different rates of growth (i.e., different slopes). These constants (i.e., the slope and intercept in the straight-line growth model) are known as the individual growth parameters and it is interindividual differences in these parameters that distinguishes the growth of one individual from another.

Ideally, in any investigation, a growth model is selected so that the individual growth parameters are rationally interpretable "in terms of the physical or biological [or psychological] processes responsible for the observed relationships" (Bock & Thissen, 1980). Of course, the intention is also to seek a parsimonious and well-fitting representation. Many authors have provided reviews of the learning curve literature emphasizing this "rational" approach to the selection of academic growth models (Atkinson, Bower, & Crothers, 1965; Goldstein, 1979; Gulliksen, 1934; Hilgard, 1951; Keats, 1983; Lewis, 1960; Shock, 1951). Most of the proposed learning models are based on a representation of the particular learning process as a first-order differential equation whose structure can be attributed to some particular psychological theory of learning. For instance, Keats (1983) demonstrated that both the negative exponential learning model and the hyperbolic learning model can be derived from a feedback model of cognitive growth. Other examples include Gulliksen's (1934) demonstration that the same negative exponential growth model can be obtained from Thurstone's (1930) rational model of learning, and Hull's (1943) use of the negative exponential growth function to model "habit strength" (see Hilgard, 1951, p. 561). In a similar fashion, other popular individual academic growth models (such as the logistic) can be derived from alternative psychological theories of learning (Lewis, 1960; Singer & Spilerman, 1979).

When little is known of the mechanisms governing the growth process, an "empirical" approach to growth model selection can be applied and a well-fitting member of the class of polynomial functions adopted as an approximation to the actual growth model. Because a higher-order polynomial will always be better or equal in fit to a lower-order model, a trade-off between goodness-of-fit and parsimony is involved in picking the order of the appropriate polynomial. Unfortunately, as a result of algebraic complexity, higher-order polynomial growth models frequently contain parameters whose bio-
logical or cognitive significance is difficult to interpret. Thus lower-order polynomials (such as the linear and quadratic functions) have become the most popular for the empirical modeling of growth.

A number of mathematical functions may be useful for the modeling of educational, social, and psychological growth: the straight-line and quadratic growth models (selected under the empirical strategy), and the negative exponential growth model (selected under the rational strategy). The functional form of each of these models is presented below and possible interpretations of the individual growth parameters are briefly discussed. Subsequently, in Section V, the fitting of individual growth models to data will be discussed when an example of an investigation of systematic interindividual differences in growth is presented.

**Straight-Line Individual Growth Model**

In educational, social, and psychological research, the individual growth model of first choice is frequently a simple linear function of time. There are several reasons for this selection. First, the mechanisms driving the growth are often unknown and therefore the empirical model-selection strategy described above is applied. This usually leads to the adoption of a low-order polynomial to represent the individual growth. And, even though the growth might be curvilinear in the long run, it may be locally linear (suggesting that the straight-line model may be appropriate). Furthermore, when only a small portion of the total life span is being investigated, the observed growth records will usually contain a limited number of datapoints and therefore, only a growth function with a small number of parameters can be fitted successfully. Second, the use of the straight-line model “possesses a degree of built-in robustness against departure from the local linearity assumption” because the slope of a fitted straight-line model is equal to the average slope of the quadratic function over the same interval (Hui & Berger, 1983, p. 754; see also Seigel, 1975). Third, a collection of straight-line growth models (to represent the heterogeneous growth of a group of individuals) has the appealing property of dynamic consistency. For a collection of growth trajectories, dynamic consistency exists if the “curve of the averages” (the curve fitted to the mean \( \xi \)-values at each value of \( t \)) is the same as the “average of the curves” (obtained by averaging the parameter values over the population and plotting the resulting curve). Many popular growth functions (such as the Gompertz, Jøns, and logistic functions) are not dynamically consistent and therefore the character of the individual trajectories is distorted by group averaging (Boas, 1892; Estes, 1956; Keats, 1983).

Under the straight-line (constant rate-of-change) growth model, the true status of individual \( p \) at time \( t \), \( \xi_p(t) \), is a linear function of time. The algebraic representation of the growth model contains two individual growth parame-
ters: the status of individual $p$ at an arbitrarily selected time $t^*$, $\xi_p(t^*)$, and the rate of growth (or slope) $\theta_p$.

$$\xi_p(t) = \xi_p(t^*) + \theta_p(t-t^*),$$

where $t^*$ can be the time of the initial observation for instance, making $\xi_p(t^*)$ the true "initial" status. If, for example, $\xi_p(t)$ represents the true level of achievement of individual $p$ at time $t$ and $t^*$ is the time at which students enter upon a particular program, then $\xi_p(t^*)$ represents the true level of achievement of individual $p$ on entry into the program. The growth rate parameter $\theta_p$ indicates how rapidly the true status of individual $p$ changes per unit time. If $\theta_p$ is positive, then the true status is increasing with time; if it is negative then the true status is decreasing with time. Providing the model is appropriate, the two individual growth parameters $\xi_p(t^*)$ and $\theta_p$ completely specify the growth in $\xi_p(t)$ for the $p^{th}$ individual, and it is these two parameters that would normally be estimated in a study of individual growth. A plot of the straight-line growth model is provided in Figure 6.

Frequently in empirical research, where $T$ discrete observations on the status of each subject in the sample have been made on occasions $t_1$ through $t_r$, the origin of the time axis is chosen to coincide with the onset of the study and, in the growth model of Equation 19, $t^*$ is chosen to be equal to $t_1$. During subsequent data analysis, this latter model is fit to the observed growth record of each individual. Unfortunately, under this strategy, the estimate of $\xi_p(t^*)$ obtained during fitting is estimated at the very beginning of the observed data stream, with consequent reduction in the precision of the estimation. A more thoughtful strategy is to redefine $t^*$ as the average value, $\bar{t}$ of the $T$ times of observation and, in addition, the individual growth parameter $\xi_p(\bar{t})$ can be estimated with the highest precision possible.

**Quadratic Individual Growth Model**

Quite frequently in empirical work (and particularly when data have been gathered over an extended period of time) it may be appropriate to select a curvilinear growth model to represent true individual growth. When no suitable theory is available to suggest that one particular curvilinear function may be substantively more appropriate than another, the investigator may select a low-order polynomial, such as a quadratic, as a reasonable and parsimonious mathematical model for true individual growth. Then, the true educational, social, or psychological status of individual $p$ might be presented as

$$\xi_p(t) = \xi_p(t^*) + \phi_p(t-t^*) + \frac{1}{2} \alpha_p (t-t^*)^2,$$

where there are three individual growth parameters in the model: true status at $t^*$, $\xi_p(t^*)$, and the coefficients of the linear and quadratic terms, $\phi_p$ and $\alpha_p$. 
FIGURE 6: Straight-line individual growth. A plot of true individual status as a function of time, under the straight-line growth model of Equation 19 with $\xi_0(1)=2$ and $\theta_x=1.2$.

FIGURE 7: Quadratic individual growth. A plot of true individual status as a function of time, under the quadratic growth model of Equation 20 with $\xi_0(0)=0$, $\phi_x=2.83$, and $\sigma_x=-.25$. 
(a factor of \( \frac{1}{2} \) has been introduced into the last term in Equation 20 in order to facilitate the interpretation of \( \alpha_p \)—see below). Some or all of these three individual growth parameters would typically be estimated in a growth study. A plot of the quadratic growth function is presented in Figure 7.

But, is there some simple interpretation of the three individual growth parameters? If \( t^* \) has been selected arbitrarily, then a certain arbitrariness accrues to the interpretation of both \( \xi_p(t^*) \) and \( \phi_p \). The former parameter is simply the true status of individual \( p \) at \( t^* \), and the latter parameter is the slope of a tangent to the growth trajectory at \( t^* \). Inspection of Figure 7 shows that the slope of the quadratic model is not the same on all occasions, but changes with the passage of time (unlike the slope of the straight-line growth model in Figure 6). Individual growth trajectories possessing this type of curvature are not atypical in developmental research. When the slope of a growth trajectory is decreasing with time in this fashion, it is said to be a decelerated (or negatively accelerated) curve. In fact, the quadratic growth model in Equation 20 has a constant acceleration over its entire length and this acceleration is given by the individual growth parameter, \( \alpha_p \) (note that, for the decelerated growth trajectory in Figure 7, \( \alpha_p \) is negative).

Empirically, where \( T \) discrete observations on the status of each subject in the sample have been made on occasions \( t \), through \( t_r \), a sensible choice for \( t^* \) is the average value, \( \bar{t} \), of the \( T \) times of observation. Then, \( \xi_p(t^*) \) is the true status of individual \( p \) at \( t \) and \( \phi_p \) is the instantaneous slope of the growth trajectory at the same instant. In addition, providing the occasions of measurement are symmetrically distributed around \( \bar{t} \), \( \phi_p \) is also equal to both the average slope of the growth trajectory between \( t \) and \( t_r \), and also to the slope of the straight line joining true status at \( t \) and \( t_r \) (Seigel, 1975). These additional equalities often permit the empirical researcher to make meaningful statements about the individual growth parameters estimated under the quadratic model although, more often than not, the model has been selected under an empirical, rather than a rational, model-selection strategy.

Negative Exponential Growth Model

Although offering a simple mathematical model for nonlinear growth, the quadratic growth model does not include growth to an asymptote, a growth pattern found frequently in biological and cognitive data (Goldstein, 1979, pp. 74–79; Thurstone & Ackerson, 1929). For this reason, a more reasonable cognitive or psychological growth function may be the negative exponential growth model. The negative exponential growth model represents the true status of each individual as an exponential function of time, and contains three individual growth parameters: true status at \( t^* \), \( \xi_p(t^*) \), the asymptote or ceiling of the growth process \( \lambda_p \), and a rate constant \( \gamma_p \). Thus, for individual \( p \),
\[ \xi_p(t) = \lambda_p - [\lambda_p - \xi_p(t^*)]e^{-\gamma_p t}. \] (21)

A plot of the negative exponential growth model is provided in Figure 8.

In many psychological applications \( \gamma_p \) is referred to as the "learning rate constant" because, given a specific amount \( \lambda_p - \xi_p(t) \) yet to be learned, the instantaneous true growth rate is directly proportional to \( \gamma_p \). If the learning rate constant has a higher value then growth approaches asymptote more rapidly. The negative exponential growth process, by analogy with the growth of daughter products in a radioactive decay, possesses a half-life. Half-life is defined as the time taken to grow from any current status, \( \xi_p(t) \), to a status halfway between \( \xi_p(t) \) and the ultimate limit of growth \( \lambda_p \). For the negative exponential growth function, half-life is equal to \( (\log_2)\gamma_p \). The inverse dependency of the half-life on \( \gamma_p \) emphasizes the role of the latter parameter as a rate constant. These individual growth parameters, describing asymptote and instantaneous rate of growth, would ordinarily be estimated in an empirical study of growth.

Modelling Interindividual Differences in True Growth

If, in a given population, the true growth of each individual can be represented by a common mathematical function, then interindividual heterogeneity in growth will occur when there are differences in the values of the individual growth parameters from person to person. For instance, if individual true growth in a particular population can be represented by a straight-line function, then any possible pattern of interindividual heterogeneity in growth can be attributed solely to differences in \( \xi_p(t^*) \) and \( \theta_p \) among individuals; individuals will either differ in their status at some arbitrary "starting" time or in their rate of growth, or both. In Figure 9 three collections of straight-line individual growth trajectories are shown that illustrate different patterns of interindividual heterogeneity in growth. If individual true growth in the population were more appropriately represented by a negative exponential function, then interindividual heterogeneity in growth would be attributable to between-person variation in "starting" status, asymptote, and rate parameter, or some combination of the three. In this latter instance, plots displaying alternative manifestations of interindividual heterogeneity in growth, similar to those presented for the straight-line growth model in Figure 9, could also be constructed and displayed.

Furthermore, once a suitable mathematical model has been selected to represent the individual growth, vague questions about interindividual differences in growth can be replaced by specific questions about interindividual differences in the individual growth parameters. Thus, in practice, not only will the dimensionality of the empirical problem be reduced (in that each entire observed growth record can be replaced by estimates of a few critical growth parameters), but a considerable refinement of research question can
FIGURE 6: Negative exponential individual growth. A plot of true individual status as a function of time, under the negative exponential growth model of Equation 21 with $\xi(1) = 2$, $\lambda_0 = 9$, and $\gamma_0 = -1.1$. The horizontal broken line indicates the position of the asymptote, $\lambda_p$.

also be achieved. Depending on the particular growth model selected, the researcher may be in a position to formulate questions about specific attributes of the growth process (such as the ultimate limits of growth, and the instantaneous rates of growth), and also about their relationships with important covariates. Thus, by adopting an appropriate mathematical model to represent the individual growth, the research question can be specified more clearly and focused more meaningfully on the attributes of the underlying growth process.

**Questions of Heterogeneity in Growth**

In a comparison of the two systems of individual growth trajectories in Figures 9a and 9b, rather than vaguely struggling to comprehend the global differences between apparently complex patterns of growth, the reader can simply note that the principal difference between the two growth patterns is one of interindividual variability in slope. Or, between the two collections in Figures 9b and 9c, the distinction is one of interindividual variability in "starting" status. What was formerly a distinction almost too difficult to discuss, becomes a distinction that can be clearly and meaningfully characterized in terms of the between-person variation of the individual growth parameters. Thus, when the straight-line function is an appropriate model of individual growth, it becomes informative to ask: What is the between-person variability in true growth rate in the population? and, What is the between-person variability in true "starting" status in the population?
FIGURE 9: Patterns of interindividual heterogeneity in growth. Three collections of ten individual straight-line growth trajectories illustrating different aspects of interindividual heterogeneity in growth.
In other words, whenever a particular mathematical model has been selected to represent individual growth, there are questions of population interindividual heterogeneity in the individual growth parameters. This suggests that, in the second phase of an empirical growth study, specific between-person variabilities should be estimated. For instance, if the straight-line function is the appropriate model for individual growth, then crucial parameters of interest are the between-person variance of true slope $\sigma^2_s$, and the between-person variance of true "starting" status $\sigma^2_{\theta_{01}}$. The former parameter captures information on population interindividual differences in rates of true growth, and is independent of time under the straight-line model. The latter parameter, while appropriately summarizing population interindividual variation in true status at $t^*$, is undoubtedly a function of the choice of $t^*$. Thus, unless there is an empirical or substantive reason for picking one specific value for $t^*$, interpretation of $\sigma^2_{\theta_{01}}$ becomes difficult. It may be more appropriate to estimate $\sigma^2_{\theta_{01}}$ for several values of $t^*$ in order to more fully understand the extent to which interindividual differences in true status fluctuate throughout the duration of the study.

Similarly, if a nonlinear function is required to model the individual growth, then other between-person variabilities become the focus of interest in the second phase of the growth study. Under the quadratic growth model, the population between-person variance of true growth acceleration $\sigma^2_t$ might be a parameter worth estimating while, under the negative exponential growth model, an estimate of the population between-person variance of true asymptote $\sigma^2_t$ might be of overriding importance. It is for the researcher to decide: first, which individual growth model is appropriate for representing the specific growth process under study, and second, which between-person parameters should be estimated in order to answer the particular research questions of interest.

**Questions of Systematic Heterogeneity in Growth**

In addition to questions about the between-person variability of the individual growth parameters, there are questions of systematic heterogeneity in growth that naturally arise in an empirical study of growth. The individual growth parameters may differ from person to person in a way that is related to some background characteristic or covariate, $\omega_s$. For instance, under the straight-line individual growth model, the investigator might ask: Is there some systematic relationship between the individual growth rates and $\omega_s$? Following Rogosa and Willett (1985), a simple representation of systematic interindividual differences in growth is:

$$E[\theta_s | \omega_s] = \mu_s + \beta_{\omega_s} (\omega_s - \mu_\omega),$$

(22)
where \( \mu_\omega \) and \( \mu_\omega \) are the expectations of \( \theta_\omega \) and \( \omega_\omega \), respectively, and \( \theta_\omega \) and \( \omega_\omega \) are assumed to be linearly related, all expectations being obtained over persons.\(^{12}\)

In the conditional expectation of Equation 22, the regression parameter \( \beta_{\omega\omega} \) summarizes population differences in the individual rates of true growth per unit difference in \( \omega_\omega \). If individuals with larger values of the covariate grow more rapidly then \( \beta_{\omega\omega} \) will be positive, if smaller values of \( \omega_\omega \) correspond to more rapid rates of true growth then \( \beta_{\omega\omega} \) will be negative. In either case, a nonzero value of \( \beta_{\omega\omega} \) indicates that \( \omega \) is a predictor of growth. Often, it is convenient to speak in terms of the correlation between true growth and the background characteristic \( \rho_{\omega\omega} \), rather than in terms of the regression coefficient. Both \( \beta_{\omega\omega} \) and \( \rho_{\omega\omega} \) address such questions as, Is an individual's value of \( \omega \) related to the rate at which he or she learns (grows)? Or, Do particular types of individual learn (grow) faster than others? One, or both, of these between-person parameters might be estimated in the second phase of an empirical study of individual growth in order to examine the systematic interindividual differences in growth.

Similarly, if a nonlinear function is selected for the modeling of individual true growth, then corresponding conditional expectations would describe the ensuing systematic heterogeneity in the individual growth parameters. For instance, under the negative exponential growth model, a between-person regression model could be stated which would describe the nature of the systematic relationship between the covariate \( \omega_\omega \) and the upper limit or asymptote on individual growth \( \lambda_\omega \), over persons in the population. In that latter model, between-person parameters such as \( \beta_{\omega\omega} \) and \( \rho_{\omega\omega} \) would be the critical focus of research interest and would be estimated in the second phase of the growth study. Similar population regression models could be written to represent systematic interindividual heterogeneity in the rate parameter of the negative exponential growth model, or the acceleration parameter of the quadratic growth model, and the corresponding between-person regression and correlation coefficients would then become available for estimation.

V. MEASURING CHANGE IN PRACTICE

The purpose of the current section is to illustrate, by example, some simple ways in which the various mathematical and statistical models that have been introduced in earlier sections can be applied in an empirical investigation of growth using multivariate data. For purposes of this illustration, the four-wave sample of individual growth data presented in Exhibit 2 is reexamined.\(^{12}\) A hierarchy of data-analytic strategies of increasing sophistication are illustrated and discussed. Each of the strategies builds, in a straightforward way, on the strategy preceding it. These strategies are illustrative of the general estimation
problem and are useful in simple cases and as exploratory tools in more complex problems.

Assembling the Statistical Models

This chapter has distinguished a pair of linked, but conceptually distinct, phases in the investigation of growth. Underlying the data analysis in the two phases is a pair of statistical models: a model for individual true growth, and a model for systematic interindividual differences in true growth. Examples of both types of model have been presented and discussed in Section IV, but are gathered together here in order to prepare for the illustrative data analyses. It is the parameters of these models that are the focus of interest in an investigation of growth.

In the current reexamination of the data in Exhibit 2, the straight-line function has been adopted here as an appropriate model to represent true individual growth in the population from which the sample in Exhibit 2 was drawn. Thus, from Equation 19 of Section IV, the true score of individual \( p \) at time \( t \) is represented by:

\[
\xi_p(t) = \xi_p(t^*) + \theta_p(t-t^*),
\]

(23)

where previously defined notation has been employed. However, Equation 23 characterizes growth in "opposites-naming skill" as an unseen or latent trait. In order to investigate growth in this skill, discrete fallible observations on \( \xi_p(t) \) must be made over time on a sample of individuals. Thus, from Equation 1 of Section III, the observed score of individual \( p \) at time \( t_i \) (\( i = 1, \ldots, 4 \)) is given by:

\[
X_p = \xi_p(t_i) + \theta_p(t_i - \bar{t}) + \epsilon_p,
\]

(24)

where previously defined notation has been employed, and the intercept of the individual growth trajectories has been defined at the "center" of the occasions of measurement, \( t = 2.5 \), as discussed in Section IV. Assumptions concerning the distribution of the \( \epsilon_p \) are the usual ones for classical test theory and have been established in Section III.

The individual growth parameter \( \theta_p \) represents the rate at which individual true status is incremented over time. The research question of interest in the illustrative data analysis focuses on this parameter and asks, Are interindividual differences in the rate of acquisition of "opposites-naming skill" systematically related to the values of the covariate? To answer this question, a statistical model is required to represent systematic interindividual differences in growth; to relate interindividual differences in rate of acquisition of "opposites-naming skill" to the selected covariate. Such a model has been presented in Equation 22 of Section IV, and is repeated here:
Thus, the data analysis seeks estimates of the population regression coefficient, $\beta_{\omega}$, or the corresponding correlation coefficient $\rho_{\omega}$.

**Using Estimates of the Individual Growth Parameters Directly as Measures of Growth**

Rao (1958), citing an earlier empirical study by Wishart (1938), suggests that, providing an adequate mathematical model has been adopted to characterize the growth of each individual in the population, then ordinary least-squares (OLS) regression analysis can be used to summarize the essential information about individual growth contained in each of the observed growth records. The obtained slopes and intercepts are then unbiased estimates of the individual growth parameters and can simply stand in place of the observed growth record in a rudimentary second phase between-person analysis, in much the same way as the difference score can be used as an observed measure of growth in the two-wave investigation of systematic interindividual differences in growth.  

*Within-person analysis.* Inspection of the observed scores in Exhibit 2 shows a trend to higher scores on subsequent occasions for all subjects. The average score of the group rose monotonically from 164.5 on occasion 1 to 246.1 on occasion 4. Thus, on average, the observed scores increased by approximately 27 between successive occasions of measurement. In order to summarize the individual time courses (under the assumption of a straight-line individual growth function), the model in Equation 24 was fitted separately to each individual growth record by OLS regression. A plot showing the obtained collection of 35 fitted straight-line growth trajectories was presented earlier in Figure 4, and Tables 3 and 4 display selected estimates from these within-individual regressions.

There is considerable evidence of interindividual heterogeneity in estimated growth rate in the sample. The empirical distribution of the estimated slopes ranges from 3.7 additional "opposites" per session to 49.3 additional "opposites" per session, has a median value of 26.8 and an interquartile range of 19.7. There is also considerable variation in the precision of the individual fits: the standard error of the fitted slopes ranges from 0.933 to 11.906. In general, though, the multiple-$R^2$ in each of the individual regressions is high, with 31 out of the 35 coefficients of determination being larger than 50 percent. Finally, the empirical distribution of the residual variance from the individual fits has a high positive skew, as would be expected of a statistic that possesses a chi-square distribution, and ranges from a minimum of 4.35 to a maximum of 708.75 with a median value of 121.15.
### TABLE 3
Selected Estimates from the 35 Within-Individual OLS Regressions of Observed Score on Time

<table>
<thead>
<tr>
<th>ID</th>
<th>$\lambda_p$</th>
<th>$\hat{\theta}_p$</th>
<th>s.e. ($\hat{\theta}_p$)</th>
<th>Coefficient of Determination $R^2_p$</th>
<th>Residual Variance $MSE_p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>137</td>
<td>34.2</td>
<td>5.271</td>
<td>95.47</td>
<td>138.90</td>
</tr>
<tr>
<td>02</td>
<td>123</td>
<td>28.5</td>
<td>1.775</td>
<td>99.23</td>
<td>15.75</td>
</tr>
<tr>
<td>03</td>
<td>129</td>
<td>47.9</td>
<td>5.429</td>
<td>97.50</td>
<td>147.35</td>
</tr>
<tr>
<td>04</td>
<td>125</td>
<td>21.9</td>
<td>0.933</td>
<td>99.64</td>
<td>4.35</td>
</tr>
<tr>
<td>05</td>
<td>81</td>
<td>9.9</td>
<td>6.170</td>
<td>56.28</td>
<td>190.35</td>
</tr>
<tr>
<td>06</td>
<td>110</td>
<td>29.6</td>
<td>6.980</td>
<td>89.99</td>
<td>243.60</td>
</tr>
<tr>
<td>07</td>
<td>99</td>
<td>32.6</td>
<td>8.642</td>
<td>92.33</td>
<td>220.60</td>
</tr>
<tr>
<td>08</td>
<td>113</td>
<td>37.9</td>
<td>3.804</td>
<td>98.03</td>
<td>72.35</td>
</tr>
<tr>
<td>09</td>
<td>104</td>
<td>22.7</td>
<td>4.922</td>
<td>91.40</td>
<td>121.15</td>
</tr>
<tr>
<td>10</td>
<td>96</td>
<td>28.8</td>
<td>2.209</td>
<td>98.68</td>
<td>24.40</td>
</tr>
<tr>
<td>11</td>
<td>125</td>
<td>19.7</td>
<td>7.213</td>
<td>78.80</td>
<td>260.15</td>
</tr>
<tr>
<td>12</td>
<td>115</td>
<td>40.7</td>
<td>3.135</td>
<td>98.83</td>
<td>49.15</td>
</tr>
<tr>
<td>13</td>
<td>109</td>
<td>8.6</td>
<td>4.660</td>
<td>63.00</td>
<td>108.60</td>
</tr>
<tr>
<td>14</td>
<td>95</td>
<td>7.9</td>
<td>4.677</td>
<td>59.79</td>
<td>109.35</td>
</tr>
<tr>
<td>15</td>
<td>118</td>
<td>16.5</td>
<td>11.906</td>
<td>48.99</td>
<td>708.75</td>
</tr>
<tr>
<td>16</td>
<td>120</td>
<td>25.3</td>
<td>3.746</td>
<td>99.61</td>
<td>70.15</td>
</tr>
<tr>
<td>17</td>
<td>118</td>
<td>42.1</td>
<td>8.593</td>
<td>95.32</td>
<td>217.35</td>
</tr>
<tr>
<td>18</td>
<td>115</td>
<td>49.3</td>
<td>5.313</td>
<td>97.73</td>
<td>141.15</td>
</tr>
<tr>
<td>19</td>
<td>120</td>
<td>17.9</td>
<td>8.722</td>
<td>67.80</td>
<td>380.35</td>
</tr>
<tr>
<td>20</td>
<td>118</td>
<td>31.1</td>
<td>8.091</td>
<td>86.00</td>
<td>327.35</td>
</tr>
<tr>
<td>21</td>
<td>126</td>
<td>42.6</td>
<td>8.457</td>
<td>92.69</td>
<td>357.60</td>
</tr>
<tr>
<td>22</td>
<td>121</td>
<td>6.4</td>
<td>7.611</td>
<td>26.12</td>
<td>289.60</td>
</tr>
<tr>
<td>23</td>
<td>108</td>
<td>15.8</td>
<td>3.617</td>
<td>91.52</td>
<td>65.40</td>
</tr>
<tr>
<td>24</td>
<td>104</td>
<td>7.2</td>
<td>6.015</td>
<td>41.74</td>
<td>180.90</td>
</tr>
<tr>
<td>25</td>
<td>118</td>
<td>21.7</td>
<td>3.468</td>
<td>95.14</td>
<td>90.15</td>
</tr>
<tr>
<td>26</td>
<td>104</td>
<td>26.4</td>
<td>3.118</td>
<td>97.29</td>
<td>48.80</td>
</tr>
<tr>
<td>27</td>
<td>124</td>
<td>21.8</td>
<td>4.310</td>
<td>92.75</td>
<td>92.90</td>
</tr>
<tr>
<td>28</td>
<td>130</td>
<td>35.3</td>
<td>2.798</td>
<td>98.76</td>
<td>39.15</td>
</tr>
<tr>
<td>29</td>
<td>87</td>
<td>3.7</td>
<td>3.903</td>
<td>31.01</td>
<td>78.15</td>
</tr>
<tr>
<td>30</td>
<td>131</td>
<td>31.6</td>
<td>6.642</td>
<td>91.88</td>
<td>220.60</td>
</tr>
<tr>
<td>31</td>
<td>109</td>
<td>26.4</td>
<td>4.210</td>
<td>95.16</td>
<td>88.60</td>
</tr>
<tr>
<td>32</td>
<td>115</td>
<td>48.8</td>
<td>0.990</td>
<td>99.92</td>
<td>4.90</td>
</tr>
<tr>
<td>33</td>
<td>104</td>
<td>37.6</td>
<td>5.933</td>
<td>95.41</td>
<td>170.10</td>
</tr>
<tr>
<td>34</td>
<td>110</td>
<td>42.5</td>
<td>7.426</td>
<td>94.24</td>
<td>275.75</td>
</tr>
<tr>
<td>35</td>
<td>110</td>
<td>20.7</td>
<td>3.468</td>
<td>94.68</td>
<td>60.15</td>
</tr>
</tbody>
</table>

**Between-person analysis.** The OLS estimates of true slope obtained above can be used directly as measures of the individual growth rates in the second-phase between-subject analysis. However, it must be realized that $\hat{\theta}_p$, being an estimate of the hypothesized individual true growth rate in Equation 23, is an unbiased yet fallible measure of that growth. Providing the straight-line function is an appropriate model for the individual true growth, the empirical re-
TABLE 4
Stem-Leaf Plots of Selected Estimates from the 35 Within-Individual Growth Trajectory Regression Fits

<table>
<thead>
<tr>
<th>Estimated Rate of Growth, $\hat{\theta}_p$</th>
<th>Residual Variance, MSE$_p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stem</td>
<td>Leaves</td>
</tr>
<tr>
<td>4t</td>
<td>899</td>
</tr>
<tr>
<td>4</td>
<td>1233</td>
</tr>
<tr>
<td>3f</td>
<td>568</td>
</tr>
<tr>
<td>3</td>
<td>01234</td>
</tr>
<tr>
<td>2f</td>
<td>66789</td>
</tr>
<tr>
<td>2</td>
<td>012223</td>
</tr>
<tr>
<td>1f</td>
<td>778</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0f</td>
<td>6789</td>
</tr>
<tr>
<td>0</td>
<td>4</td>
</tr>
</tbody>
</table>

Multiply Stem. Leaf by 10

The slope estimates in Table 3 and the fitted growth trajectories in Figure 4 offer evidence of considerable interindividual variability in the growth-rate measure. The sample variance of $\hat{\theta}_p$ is 164.3, suggesting that there may also be considerable corresponding interindividual heterogeneity in the underlying rates of true growth. By regressing $\hat{\theta}_p$ on $\omega_p$, or correlating the two variables, the investigator can easily estimate the strength of the population relationship between the new observed measure of growth rate and the selected covariate. Sample estimates of these two parameters are .433 and .422 respectively. This latter estimated correlation suggests that there is a moderately strong association between individual growth and the selected covariate.

Correcting for the Influence of Measurement Error

In the previous subsection, the first-phase within-individual regressions led to unbiased estimates of the individual rates of true growth. These estimates
were then used as accessible summaries of individual growth, summaries that could be readily obtained in empirical practice no matter how many waves of data had been collected. These individual growth summaries then became the basis of a second-phase between-individual analysis.

Unfortunately, when the estimated individual growth summaries enter into the second-phase analysis, the second-phase between-individual parameters being estimated are not the parameters of fundamental empirical interest in a study of systematic interindividual differences in true growth! In fact, in the second-phase analysis, rather than the population relationship between the true growth rate \( \theta \) and the selected covariate \( \omega \) being investigated, it is the population relationship between the estimated growth rate \( \hat{\theta} \) and \( \omega \) that is being examined. In other words, the individual growth parameters have been "measured" (i.e., estimated) with error in the first phase of the analysis and they are carrying this built-in fallibility through into the second-phase analysis with them, disturbing and attenuating the findings. This is a serious problem that must be avoided if the systematic nature of the relationship between true growth and the selected covariate is to be examined appropriately.

**Population Variance of the Estimated Growth Rates**

The sample variance of the estimated individual growth rates \( \hat{\theta} \) was proffered as an indicator of population heterogeneity in growth. However, as the estimated individual growth rates are fallible measures of the underlying true growth rates, their variance over persons will be inflated by the presence of the measurement error in the original panel data. Thus, the sample variance of \( \hat{\theta} \) will overestimate the population variance of the true growth rate \( \sigma^2_\theta \) (in exactly the same way as the sample variance of the difference scores overestimated \( \sigma^2_\delta \) in Section III).

What is the nature of the relationship between the population variances of individual true growth rate and estimated growth rate? It is very simple indeed, and is given by:

\[
\sigma^2_\hat{\theta} = \sigma^2_\theta + \frac{\sigma^2_\delta}{\text{SST}},
\]

where SST is the sum of the squared deviations of the observation times about their mean, \( \sum(t_i - \bar{t})^2 \), a measure of the "spread" of the occasions of measurement (Willett, 1985). For the multiwave panel data in Exhibit 2, SST is equal to 5 because the four waves of data were obtained at the times \( t_i = 1, 2, 3, \) and 4. Since both \( \sigma^2_\delta \) and SST are positive, the population variance of the
estimated growth rates over persons is made larger (inflated) by the presence of measurement error and is consequently greater than the total variance of the true growth rates. Therefore if, in the second phase of a growth analysis, the heterogeneity in growth is unthinkingly estimated by the usual sample variance of \( \hat{\theta}_p \), then it is likely to be overestimated.

*Estimating \( \sigma^2 \).* The availability of multiwave panel data permits the legitimate estimation of the measurement error variance, and allows an improved estimate of \( \sigma^2 \) to be obtained. Assuming that the individual growth model has been specified correctly, deviations of individual observed status from the corresponding fitted trajectory estimate the random error introduced into the measurement process at each occasion for each individual.\(^{16}\) The magnitude of the measurement error variability can be found directly from these residuals. Under the assumptions of this paper, an estimate of \( \sigma^2 \) can be found quite simply by summing the squared residuals from all the first-phase regressions over occasions and persons, and then dividing by the total degrees of freedom. By simple algebraic manipulation, this can be shown to be equivalent to:

\[
\hat{\sigma}^2 = \frac{\sum_{p=1}^{n} \text{MSE}_p}{n},
\]

and the estimated measurement error variance is simply the sample average of the estimated residual variances from each of the first-phase individual regressions. This latter quantity is easily obtained by averaging the entries in the final column of Table 3, giving 159.5 as the estimated measurement error variance for the growth data in Exhibit 2.

Once \( \hat{\sigma}^2 \) has been estimated, \( \sigma^2 \) can be estimated by inverting Equation 26 and replacing the parameters by the appropriate sample estimates. Thus,

\[
\sigma^2 = \hat{\sigma}^2 - \frac{\hat{\sigma}^2}{\text{SST}}.
\]

In our example, the sample variance of the estimated growth rates is 164.3, the estimated measurement error variance is 159.5, and SST is 5. Therefore, by substitution into Equation 28, a disattenuated estimate of the population variance of true growth rate \( \sigma^2 \) is 132.4, providing strong evidence of inter-individual differences in true growth rate. This indicates that the individual true growth trajectories are not all parallel in the population and suggests, in addition, that the reliability of the individual growth measure \( \hat{\theta}_p \) could be large.
Population Reliability of the Estimated Growth Rate

Because reliability is the ratio of true to observed (over-person) variance, the population reliability of the estimated individual growth rates \( \rho(\hat{\theta}) \) is:

\[
\rho(\hat{\theta}) = \frac{\sigma^2_\tau}{\sigma^2_\tau + \frac{\sigma^2_\varepsilon}{[\text{SST}]}},
\]

(29)

where the denominator has been reexpressed using Equation 26 (Willett, 1985; cf. Rogosa et al., 1982, Equation 22).

Notice that, all else remaining constant, the reliability of the individual growth measure is larger when the heterogeneity of individual true growth over persons is larger; as \( \sigma^2_\tau \) becomes very large, the reliability of \( \hat{\theta} \) approaches one. Thus, in a population in which there is plenty of criss-crossing of true growth trajectories, considerable reliability is possible in practice. On the other hand, if there are no interindividual differences in the rate of true growth (\( \sigma^2_\tau = 0 \)) then all of the true growth trajectories will be parallel and the growth reliability can only be zero, regardless of the precision with which measurement has been achieved. Reliability, as a measure of interindividual differentiation, can only be high if there are interindividual differences to be detected. When individual growth is being measured, it is as though the chosen growth measure (in this case, \( \hat{\theta} \)) is being asked to distinguish among the true growth of the different individuals in the group. When all the true growth trajectories are parallel (or almost so) the measure is incapable of supporting such fine distinctions and hence reliability is necessarily low, regardless of the precision with which individual observed scores have been obtained.

If the underlying configuration of individual true growth trajectories is fixed, how is it possible then to increase the reliability of the growth measurement? The most obvious approach is to reduce the magnitude of the measurement error variance. Empirically, this usually involves the reconstruction of the measuring instrument being used and is often difficult to achieve in practice. However, a more subtle (and often more effective) approach is based on the manipulation of the occasions of measurement. Notice, in Equation 29, that SST is strategically placed in the denominator as the divisor of the measurement error variance and it therefore acts directly to mediate the influence of the measurement error on the growth reliability.

The quantity SST is a function of both the number of waves of data collected in the growth study and the relative spacing of those waves. Thus, the reliability of the growth measurement can be improved either by collecting more waves of data or by judiciously altering the spacing between the waves, or both. Moreover, because SST is a quadratic function of the observation
times, it increases very rapidly as the occasions of measurement are manipulated. Therefore, the manipulation of SST is a much more effective and convenient method of increasing the growth reliability in practice.\textsuperscript{17}

Choosing the spacing of the waves. In most growth studies, the occasions of measurement are likely to be distributed more or less evenly throughout the duration of the study. However, this may not be the optimal temporal placement for the measurement times. In fact, for fixed $T$ and $i$, SST can be maximized by grouping half the measurement occasions at the beginning of the observation period and half at the end (i.e., grouping the $i$, as far from $i$ as possible). This approach is equivalent to adopting a pre/post design, using the difference score as the observed measure of growth, but improving the growth measurement by obtaining multiple measures of status on each of the two occasions of measurement. However, the ensuing increase in growth reliability is then obtained at the expense of detailed knowledge of the individual growth processes because the functional form of individual growth over time can no longer be examined empirically. Furthermore, obtaining multiple measures on each subject on each of the occasions of measurement is often impractical, unethical, and empirically unsound.

Choosing the number of waves. On the other hand, SST is dramatically increased if additional waves of data are collected, leading to increased reliability for the growth measurement. Furthermore, the extra waves of data contain supplementary information about the nature of the growth process and may permit a more sophisticated individual growth model to be adopted in the current, or subsequent, investigations. Under the simplifying assumption that the occasions of measurement are equally spaced in time, Figure 10 presents population growth reliability plotted as a function of the number of waves of data gathered $T$ (for the parameter configuration underlying Exhibit 2). Notice that the reliability is a monotonic and dramatically increasing function of $T$, so that the collection of an extra wave of data will always cause the reliability of the growth measurement to improve. Increases in reliability of more than 100% or 200% are not unusual when an extra wave of data collection is added to an existing two-wave design, and the improvement in reliability is particularly dramatic when the psychometric properties of the measuring instruments used are poor. It is those designs under which the least number of waves of data have been collected, and whose measurement is the least precise, that are the most susceptible to improvement by the collection of additional waves of data.

Estimating the reliability of the growth measurement. If the population variances in Equation 29 are replaced by sample estimates, then an estimate of the reliability of $\hat{\theta}$, as a measure of individual true growth can be obtained. In the current example, where estimates of $\sigma_2^2$ and $\sigma_2^2$ are 132.4 and 159.5 respectively, an estimate of $\rho(\hat{\theta})$ is .806 indicating excellent reliability for the
FIGURE 10: Reliability of multiwave growth measurement. Under the assumption of linear
growth, the ordinary least-squares estimate of slope \( \hat{\theta}_p \) can act as a measure of individual
true growth rate. This plot presents the population reliability of \( \hat{\theta}_p \) as a function of the number
of waves of data gathered (for the parameter configuration underlying the artificial data of
Exhibit 2).

![Reliability of Growth](image)

growth measurement and demonstrating the considerable superiority of the
multiwave measurement of growth over the two-wave approach outlined in
Section III.

**Systematic Interindividual Differences in True Growth**

During the investigation of systematic interindividual differences in
growth, it is the population correlation \( \rho_{bw} \) that describes the systematic nature
of the linear relationship between the rates of true individual growth and
the selected covariate. However, in the second part of Section V, the correla-
tion being estimated was the population correlation of \( \omega_p \) and the (fallible)
estimate of \( \theta_p \), rather than the population correlation of \( \omega_p \) and the (true) indi-
guinal growth parameter \( \theta_p \). If an estimate of the latter correlation is to be
obtained from the former, then knowledge of the population relationship be-
tween the two is required. Under the assumptions of this paper, by application
of covariance algebra and the definition of reliability,

\[
\rho_{bw} = \frac{\rho_{bw}}{\sqrt{\rho(\theta)}}
\]  

(30)
Estimating the population correlation of $\theta$ and $\omega$. As a consequence of the fallibility to the first-phase estimation of individual growth parameters, the sample correlation of $\hat{\theta}_p$ and $\omega_p$ (over persons) is an attenuated estimate of $\rho_{\theta\omega}$. Replacing the parameters (population correlation and reliability) in Equation 30 by appropriate sample estimates permits the disattenuated estimation of $\rho_{\theta\omega}$. In our example, where the estimated reliability of the growth measurement is .806 and the sample correlation of $\hat{\theta}_p$ and $\omega_p$ (over persons) is .422, the disattenuated estimate of $\rho_{\theta\omega}$ is .470 (slightly larger than the value of $\rho_{\theta\omega}$ of .438, which underpins the simulated dataset). This suggests that the hypothesized relationship between true growth and the covariate is strong.18

The important role played by SST in the functional form of $\rho(\hat{\theta})$ suggests that, by collecting sufficient waves of data, the measurement of individual growth can be rendered highly reliable and the investigation of systematic interindividual differences in growth can be very precise. In fact, in extended longitudinal datasets, not only can the parameters of complex and substantively interesting individual growth models be well estimated, but also the reliability of the individual growth measurement may be so high that the disattenuation provided by Equation 30 becomes unnecessary. In which case, the rudimentary approach described in the second part of Section V can be applied directly and the subtleties of the superior strategy outlined here may be largely ignored.

Bock (1983, pp. 109–11) showed that marginal maximum-likelihood estimation could be used to estimate the parameters of the models in the first part of Section V. However, Bock’s results are more general than those presented here. He assumed that each individual growth function was "some nonlinear function describing personal developmental trend over the range of ages at which the behavior in question is observed" (Bock, 1983, p. 109). His results are complex as a consequence of the generality of his assumptions, and the estimates themselves were obtained iteratively. Under the more restrictive assumptions of this chapter, the maximization need not be performed iteratively, and closed-form estimators can be obtained. The required closed-form estimators have been obtained by Blomqvist (1977; see also Blomqvist & Svärdssudd, 1978; Wu, Ware, & Feinleib, 1980) and are identical to the method-of-moments estimators presented here.

Estimating the Parameters of Interest More Efficiently

The well-known adage “garbage in, garbage out” has considerable import for data analysis. Whether the research is focused upon individual growth or not, the outcomes of an analysis can only be as good as the sample information that constitutes the input to the analysis. If the data analysts can feed
more information into the analysis then they are likely to learn more from what comes our! Throughout this chapter, we have seen traces of this principle in action. We have seen it in action when the two-wave investigation of systematic interindividual differences in growth was improved by the inclusion of external knowledge of test reliability to disattenuate the outcomes of the analysis. We have seen that the collection of multiwave, rather than two-wave, data leads to superior estimation of more complex and substantively interesting individual growth trajectories. We have noted that the fitting of individual growth models permits the in situ estimation of growth reliability from the residuals of the growth trajectory fits, and that this estimated reliability can be used to estimate the parameters of interest in the subsequent growth analysis more appropriately.

The process that led to the improved estimation described beneath Equation 30 is worthy of a second glance. Under these strategies, the first (within-individual) phase of the growth analysis is no longer providing simply a summary of the individual growth for subsequent use, it is also passing an estimate of the success of that summarization through to the second-phase analysis. Estimating growth reliability from the within-individual regression residuals is tantamount to transporting more information from the within-individual analysis to the between-individual analysis. It is the availability of this additional information that results in the superior estimation. Therefore, one might ask: Might there be better and more appropriate ways of achieving this passage of additional information between the two phases of the analysis? In particular, is the estimation of growth reliability the best way to characterize the precision of the first-phase analysis? Are there alternative approaches which could do a more effective job?

Certainly, our earlier discussions of the reliability of individual growth measurement should have convinced us that the very concept of reliability confounds measurement precision with interindividual heterogeneity in growth. Consequently, using the estimated growth reliability to transport information about the precision of the first-phase growth modeling through to the second-phase analyses may not be the best way to go. Furthermore, the construction of the reliability estimate required the very strong assumption that all of the (unknown) measurement errors, on every occasion and for every person, were drawn from the very same underlying distribution (an assumption which holds for the simulated data of Exhibit 2). However, in the real world of individual growth, is this likely to be a realistic assumption? And, if not, do improved analytic strategies exist which do not require the assumption and yet are still capable of notifying the second-phase analysis of the precision of the first-phase individual growth measurement?

Inspection of the selected within-individual regression estimates in Table 3 reveals that there are considerable interindividual differences in the precision with which the individual growth rates were estimated. In particular, the standard error of the individual growth-rate estimate varies appreciably from sub-
ject to subject. In the sample, s.e. \((\hat{\theta}_p)\) ranges from a minimum of 0.933 (subject #04) to a maximum of 11.906 (subject #15) with a median value of 4.922 and an interquartile range of 3.174. It is these standard errors that most appropriately capture the precision with which the slopes of the individual growth trajectories have been fit during the first-phase estimation. Therefore, it is these same standard errors which may provide a more sensible basis for the better estimation of the second-phase parameters describing systematic interindividual differences in growth.

One obvious strategy for including estimates of growth and estimates of precision in the second-phase analysis would be to use weighted least-squares (WLS) regression to fit the between-individual model in Equation 25 and to base the construction of the required weights on the standard errors obtained during the within-individual analyses. For instance, a weight matrix with elements of the form (s.e. \((\hat{\theta}_p)\)^{-2} might be appropriate for such an ad hoc procedure because, in the subsequent second-phase analysis, the more precisely determined individual growth rates would have the largest weights. Thus, this strategy would ensure that it is those slopes with the smallest standard errors (i.e., the greatest precision) that would play the most important role in the fitting of Equation 25 and in the estimation of \(\beta_{\theta_p}\). While other functions of the standard errors may suggest themselves to the reader as appropriate empirical weights, Hanushek (1974) has suggested one particular function of the

\[\text{s.e.} \,(\hat{\theta}_p)\] that provides optimal weights (in the sense that the obtained estimate of \(\beta_{\theta_p}\) is then asymptotically efficient).

The strategy outlined below is an approximation to the Hanushek procedure (call it “pseudo-Hanushek”) that is intuitively appealing and easy to apply in practice. Furthermore, under the conditions that typically prevail in a growth study, the pseudo-Hanushek procedure generates outcomes that do not differ to any great extent from the outcomes of the parent procedure. Thus, the data analyst should:

- Fit the within-individual growth models, retaining the estimated individual growth rates \(\hat{\theta}_p\) and their standard errors s.e. \((\hat{\theta}_p)\) for subsequent between-person analysis (Table 3).

- Regress the estimated individual growth rates \(\hat{\theta}_p\) on the selected covariate \(\omega_p\) by ordinary least-squares regression. From this latter OLS regression analysis, retain the estimated residual variance \(s^2\) as a global measure of the “scatter” present in the second-phase analysis. In the current example, the sample OLS regression of \(\hat{\theta}_p\) on \(\omega_p\) has a coefficient of determination of 17.8% and the estimated residual variance \(s^2\) is 139.145.

- Compute a weight matrix with elements of the form given in Equation 31 and refit the between-individual model in Equation 25 using weighted least-squares regression.
\[
W_{p} = \frac{s^2}{s^2 + [\text{s.e.}(\hat{\beta}_{m})]^2}.
\] (31)

The estimated coefficient of the covariate \( \omega_{p} \) obtained in this latter regression is an (approximately) asymptotically efficient estimate of \( \beta_{ma} \), the second-phase parameter describing the systematic nature of the relationship between the individual true growth rates and the selected covariate.

In the current example, the regression weights are listed in Table 5. The weighted least-squares estimate of \( \beta_{ma} \) is .445 (with a standard error of 0.159) suggesting that the null hypothesis of no population relationship between the true growth rate and the covariate can be strongly rejected (\( p=0.009 \)). In addition, in the weighted regression, the estimated correlation between the true growth rate and the covariate is .437 (very close to the value of \( \rho_{ma} \) of .438 that underpins the simulated dataset). Therefore we can conclude that in the population it is likely that individual true growth (in "opposites-naming") is greatest for those persons with high values on the covariate.

**Generalizing to the case of nonlinear individual growth.** An additional advantage of this proposed WLS strategy for the investigation of systematic interindividual differences in growth is that it is easily and intuitively generalizable to the case of nonlinear individual growth. Suppose, for instance, that the investigator were to adopt a negative exponential function as an appropriate model for the representation of individual true growth. In this latter case, empirical interest may, for example, center upon the individual growth asymptotes \( \lambda_{p} \) and how they are related, over individuals in the population, to some selected background characteristic \( \omega_{p} \). In this case, the regression parameter \( \beta_{ma} \) (and its associated correlation coefficient \( \rho_{ma} \)) would be estimated in the second phase of the data analysis. The WLS strategy proposed above could easily be modified to accommodate such changes. For instance, during the first phase of the data analysis, the investigator would fit the negative exponential growth function separately to each of the individual growth records using a subroutine from one of the standard statistical packages (e.g., PROC NLIN in SAS). From these nonlinear within-individual regression fits, estimates of \( \lambda_{p} \) and its associated standard error would be retained and entered into the second-phase growth analysis as in steps #2 and #3 above.

**Implications for experimental design.** Table 5 reveals that many of the empirical weights applied in the WLS regression are very similar to one another (the median weight in Table 5 is .851 and the interquartile range of the weights is .161). Furthermore, the majority of the weights have a magnitude which is very close to 1 (more than 77% of the weights are greater than .75). Of course, if all the weights had the same magnitude then the results of the OLS and WLS regressions would be identical (in fact, in the example presented here, the results of both the OLS and WLS regressions are very similar as a consequence of the interindividual homogeneity of the weights).
### TABLE 5
Construction of Weights for Application in the Second-Phase Between-individual

#### Weighted Least-Squares Regression of \( \hat{\theta}_p \) on \( \omega_p \)

<table>
<thead>
<tr>
<th>ID</th>
<th>( \omega_p )</th>
<th>( \hat{\theta}_p )</th>
<th>s.e. (( \hat{\theta}_p ))</th>
<th>( W_p )</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>137</td>
<td>34.2</td>
<td>5.271</td>
<td>.834</td>
</tr>
<tr>
<td>02</td>
<td>123</td>
<td>28.5</td>
<td>1.775</td>
<td>.978</td>
</tr>
<tr>
<td>03</td>
<td>129</td>
<td>47.9</td>
<td>5.429</td>
<td>.825</td>
</tr>
<tr>
<td>04</td>
<td>125</td>
<td>21.9</td>
<td>0.933</td>
<td>.994</td>
</tr>
<tr>
<td>05</td>
<td>81</td>
<td>9.9</td>
<td>6.170</td>
<td>.785</td>
</tr>
<tr>
<td>06</td>
<td>29.0</td>
<td>6.990</td>
<td>7.41</td>
<td></td>
</tr>
<tr>
<td>07</td>
<td>99</td>
<td>32.6</td>
<td>6.642</td>
<td>.759</td>
</tr>
<tr>
<td>08</td>
<td>113</td>
<td>37.9</td>
<td>3.804</td>
<td>.906</td>
</tr>
<tr>
<td>09</td>
<td>104</td>
<td>22.7</td>
<td>4.922</td>
<td>.852</td>
</tr>
<tr>
<td>10</td>
<td>98</td>
<td>26.8</td>
<td>2.209</td>
<td>.966</td>
</tr>
<tr>
<td>11</td>
<td>125</td>
<td>19.7</td>
<td>7.213</td>
<td>.728</td>
</tr>
<tr>
<td>12</td>
<td>115</td>
<td>40.7</td>
<td>3.135</td>
<td>.934</td>
</tr>
<tr>
<td>13</td>
<td>109</td>
<td>8.8</td>
<td>4.660</td>
<td>.865</td>
</tr>
<tr>
<td>14</td>
<td>95</td>
<td>7.9</td>
<td>6.577</td>
<td>.864</td>
</tr>
<tr>
<td>15</td>
<td>118</td>
<td>16.5</td>
<td>11.906</td>
<td>.495</td>
</tr>
<tr>
<td>16</td>
<td>120</td>
<td>28.3</td>
<td>3.746</td>
<td>.908</td>
</tr>
<tr>
<td>17</td>
<td>118</td>
<td>42.1</td>
<td>6.593</td>
<td>.762</td>
</tr>
<tr>
<td>18</td>
<td>115</td>
<td>49.3</td>
<td>5.313</td>
<td>.931</td>
</tr>
<tr>
<td>19</td>
<td>120</td>
<td>17.9</td>
<td>8.722</td>
<td>.647</td>
</tr>
<tr>
<td>20</td>
<td>118</td>
<td>31.1</td>
<td>8.091</td>
<td>.680</td>
</tr>
<tr>
<td>21</td>
<td>126</td>
<td>42.6</td>
<td>8.457</td>
<td>.661</td>
</tr>
<tr>
<td>22</td>
<td>121</td>
<td>6.4</td>
<td>7.611</td>
<td>.706</td>
</tr>
<tr>
<td>23</td>
<td>108</td>
<td>18.8</td>
<td>3.817</td>
<td>.914</td>
</tr>
<tr>
<td>24</td>
<td>104</td>
<td>7.2</td>
<td>6.015</td>
<td>.794</td>
</tr>
<tr>
<td>25</td>
<td>118</td>
<td>21.7</td>
<td>3.468</td>
<td>.920</td>
</tr>
<tr>
<td>26</td>
<td>104</td>
<td>26.4</td>
<td>3.118</td>
<td>.935</td>
</tr>
<tr>
<td>27</td>
<td>124</td>
<td>21.8</td>
<td>4.310</td>
<td>.882</td>
</tr>
<tr>
<td>28</td>
<td>130</td>
<td>35.3</td>
<td>2.798</td>
<td>.947</td>
</tr>
<tr>
<td>29</td>
<td>87</td>
<td>3.7</td>
<td>3.903</td>
<td>.901</td>
</tr>
<tr>
<td>30</td>
<td>131</td>
<td>31.6</td>
<td>6.642</td>
<td>.759</td>
</tr>
<tr>
<td>32</td>
<td>115</td>
<td>48.8</td>
<td>0.900</td>
<td>.993</td>
</tr>
<tr>
<td>33</td>
<td>104</td>
<td>37.6</td>
<td>5.833</td>
<td>.804</td>
</tr>
<tr>
<td>34</td>
<td>110</td>
<td>42.5</td>
<td>7.426</td>
<td>.716</td>
</tr>
<tr>
<td>35</td>
<td>110</td>
<td>20.7</td>
<td>3.468</td>
<td>.920</td>
</tr>
</tbody>
</table>

Since OLS regression is much easier to apply in practice than WLS regression, it would be convenient if the investigation of systematic interindividual differences in growth required the use of the former, rather than the latter, technique. Therefore, it is informative to ask: Under what conditions will all the weights in Equation 31 be (almost) equal? Is there a simple strategy that can be adopted that will render the weights equal and make redundant the use of weighted least-squares regression? Examination of the functional form of
the weights in Equation 31 readily provides the answer to this question: the weights will all be (almost) identical when the individual growth rates have been determined very precisely in the first phase of the growth analysis. In this case, all the s.e. (\(\hat{\Theta}_i\)) will be much smaller than the residual standard deviation \(s\) and all the \(W_r\) will tend to 1.

Empirically, this is a very reassuring conclusion because the first-phase individual growth-trajectory fits can easily be made extremely precise simply by collecting additional waves of data on each individual in the sample. The standard error of each of the fitted individual growth rates in Table 3 is inversely related to the quantity SST that was introduced earlier to describe the temporal distribution of the occasions of measurement. Just as the reliability of the individual growth measurement in Figure 10 is highly dependent on the magnitude of SST, the standard errors of the fitted individual growth rates will decrease dramatically as additional waves of data are collected. Thus, in a very important sense, the empirical researcher has complete control over the precision of the first-phase within-individual estimation: the investigator can simply continue to “add waves” to the panel data until the desired precision is achieved and consequently the application of WLS regression in the between-individual analysis can be avoided.

VI. SUMMARY AND DISCUSSION

To engender cognitive, psychological, and social growth in our students is a major educational enterprise of our schools. To observe and measure the individual differences in growth that exist among learners is an important empirical task. Finding characteristics of the individual and of the environment that are related to interindividual differences in growth is a crucial first step in the manipulation of the educational environment for the benefit of learners. Thus, it is essential that reliable statistical tools are available to empirical researchers for the investigation of systematic interindividual differences in growth.

This chapter has explored the measurement of growth in some detail. It has reiterated an important conceptual and mathematical framework for the collection and analysis of growth data that requires the specification of a pair of linked statistical models: a model for individual growth, and a model for interindividual differences in growth (Bryk, 1977; Rogosa & Willett, 1985). Some new multiwave strategies for the investigation of systematic interindividual differences in growth have been proposed. These exploratory techniques are relatively easy to apply in practice, and contribute greatly to the improvement of statistical estimation in the analysis of growth. Furthermore, these newer analytic strategies engender an important longitudinal perspective. They focus research interest on the statistical modeling of individual
growth trajectories, an arena in which the empirical enterprise can make the greatest and most radical gains.

In addition, some popular two-wave strategies for the measurement of individual growth (such as the use of the difference and residual-change scores) were reexamined from a growth-modeling perspective. The considerable logical, substantive, and technical flaws of the residual-change score were documented and the reader was advised to avoid this measure of individual growth. On the other hand, despite serious drawbacks, the observed difference score was shown to be a moderately successful measure of individual growth.

Contrary to popular mythology, this chapter illustrated that the difference score was not necessarily unreliable and that, when there were interindividuall differences in growth to be detected, its reliability could be very high. Furthermore it was demonstrated that, independent of reliability, the difference score certainly could be a valid measure of individual growth. However, the utility of the difference score was acknowledged to be based upon unsupported assumptions about the functional form of the underlying individual true growth trajectories. Furthermore, its successful application in the analysis of empirical growth data requires, in addition to the sample panel data, external information in the form of estimated test reliabilities. Quite frequently in the measurement of growth, the quality of this latter supplementary information is sorely lacking. Consequently, it was shown that multiwave statistical methods, which made specific assumptions about the shape of the individual growth trajectories and which permitted the in situ empirical estimation of measurement precision, were considerably superior to the usual two-wave methods for the measurement of change.

Under the individual growth modeling approach, although the entire individual time course is the actual outcome of interest, the representation of the individual growth record by a suitable growth model smoothes and summarizes the record, and permits a parsimonious, lower-dimensional representation of individual growth (Bock, 1983; Bock & Thissen, 1980). In the current chapter, true status is assumed to be a specific function of time, and each individual time course is characterized by one or two individual growth parameters that have some substantive or theoretical significance. Individuals may differ in their growth over time. Under the assumptions of this chapter, these differences are interpreted as interindividual heterogeneity in the individual growth parameters. The reason for studying correlates of growth is to detect interindividual differences in the growth parameters that are systematically related to some selected background characteristic of interest (Rogosa & Willett, 1985).

The hierarchical nature of this theoretical formulation for correlates of growth suggests that a two-tier estimation strategy may be appropriate in an empirical growth analysis. In the initial ("within-individual") phase of such
an analysis, the fitting of a suitable growth model to each observed growth record generates summary statistics that capture the salient features of the individual time courses. These summary statistics then become the basis of a second-phase ("between-individual") analysis that examines the interindividual differences in growth as a function of selected background characteristics of the individual and of the environment.

In this chapter, to demonstrate and expand upon this position, ordinary least-squares (OLS) estimates of the individual growth parameters were adopted as reasonable and intuitive summaries of the individual observed growth. In particular, under the straight-line growth model, the OLS estimate of the individual growth rate was shown to be a natural extension of the observed difference score to multiwave data. However, unlike the application of the observed difference score with two waves of panel data, the fitting of individual growth trajectories using multiwave data not only permits the more precise estimation of individual growth but also allows the deviations of the observed datapoints from the corresponding fitted trajectories to be interpreted as estimates of the random errors introduced by fallible measurement. Thus, an empirical estimate of the measurement error variance can be obtained. In this chapter, a hierarchy of improved estimators for describing the systematic nature of interindividual differences in growth were demonstrated. These new estimators incorporate adjustments to mediate the influence of measurement error on the estimation process.

Furthermore, it was shown that the precision and the reliability of the growth measurement is a rapidly and monotonically increasing function of the number of waves of data collected. Thus, very considerable improvements in growth measurement can be achieved by the collection of additional waves of data, supporting the well-worn contention that "you will always find more out, if you gather more data." In addition, the obtained improvements are particularly impressive when individual status itself cannot be measured very well.

This is not intended to suggest that the statistical techniques proposed in this chapter are the very last word in methods for the measurement of change. Superior, and more complex, strategies have recently become available. The methods that are discussed here have been kept computationally simple in order to focus more specifically on the conceptual underpinnings of the measurement of growth. Nevertheless, as a consequence of their correct conceptual focus, these methods do represent a considerable improvement over traditional methods based on the collection of two waves of data. For empirical researchers with only a moderate level of statistical training, the exploratory strategies discussed in this paper offer relatively cheap and simple ways of dramatically improving the properties of the empirical measurement of individual change.

There has been rapid, recent development in the application of multilevel
modeling to the study of growth (Bock, 1983; Stenio, Weisberg, & Bryk, 1983; Ware, 1985; see also Raudenbush & Bryk in this volume). Bryk and Raudenbush (1987) provide a clear and thorough discussion of the investigation of systematic interindividual differences in growth using the techniques of empirical Bayes estimation. They and their associates have also made a proprietary piece of computer software (called HLM—for "Hierarchical Linear Models") available to the empirical researcher (Bryk, Raudenbush, Seltzer, & Congdon, 1986).

Although much of the development in the current chapter has been limited to individual growth with a constant rate of change, multiwave strategies in the measurement of growth are easily generalized to cases of curvilinear individual growth. Thus, multiwave approaches can be particularly appealing in that they permit individual growth to be represented in a more substantively meaningful manner. Essentially, this allows the empirical researcher to enter an arena of richer and more plentiful research questions. For instance, if a negative-exponential function is selected as an appropriate individual growth model, the investigator is able to ask questions concerned with, not only the (instantaneous) rate at which an individual is growing or learning, but perhaps also questions about the ultimate limits (or asymptotes) of growth. If a quadratic individual growth model has been fit then systematic interindividual differences in the growth acceleration, as well as the average rate of growth, can be examined. Indeed, providing sufficient waves of data have been gathered to compensate for the effect of fallible measurement, the crucial ingredient in the analysis becomes, not the statistical sophistication of the estimation process, but the extent to which key features of the individual growth have been captured by successful and thoughtful selection of a suitable within-individual growth model. This message may be somewhat reassuring to the empirical researcher who does not have the technical sophistication to venture beyond the confines of the major statistical computer packages, but who has both the substantive background and logical skills required to make intelligent decisions about the nature of the individual growth.

NOTES

1 At this stage of the presentation, we are concerned only with observed measures of individual growth. In subsequent sections, observed measures of generic status are distinguished from the underlying true, or latent, abilities that are being estimated by these observations.

2 The artificial dataset listed in Exhibits 1 and 2 was created to simulate the typical attributes of a sample of educational growth data. Straight-line true growth (true growth with a constant rate of change) was assumed for each subject in the sample. The values of the covariate, the true growth rates, and true status on each of the occasions of measurement were drawn randomly from a multivariate normal distribution. The underlying correlation over persons between true growth rate and the
covariate is .438. Random errors of measurement, drawn independently from a normal distribution with mean 0 and variance 169, were added to the true scores in order to simulate the pattern of observed scores that might be obtained on a test of "meaning of opposites" (after Chapman, 1914). Parallel forms of each "test" were simulated on each of the occasions of measurement in order to obtain the cited test reliabilities.

Of course, these group-level statistics will not appropriately summarize the bivariate relationship displayed in Figure 1 unless initial and final observed status are linearly related. Considerable caution is required when such statistics are used to summarize between-wave scatterplots like Figure 1 because even when the individual growth trajectories themselves are linear, between-wave relationships of status are not necessarily linear.

The slopes of the individual growth trajectories were obtained by computing the increase in observed score per "day" for each subject, where one day was assumed to have passed between the first and second occasions of measurement.

Although the assumption of normal distributions for the measurement errors is not required at this stage, the assumption has been made here because ultimately the assumption will be required by the estimation strategies to be outlined in Section V.

This suggests, of course, that the extra information required might very easily be gathered in an investigation of growth by simply collecting additional waves of data. In fact, the collection of multiwave data does enable the empirical estimation of the reliability of the growth measure in situ and therefore the separate external estimates of the reliability of initial and final observed status can be dispensed with.

Unfortunately, as the reader will note, many of the so-called time-invariant covariates in this list are, in fact, intrinsically variable—measures of teacher and parent behavior being especially volatile, but even measures of socioeconomic status and home background being somewhat subject to variation over time. In this case, the question of correlates of growth becomes considerably more complex, with statistical models being required to characterize the variation in the covariates over time before inroads can be made into the investigation of systematic individual differences in growth. This chapter concentrates exclusively on time-invariant covariates of growth.

Rogosa et al. (1982) show that, under a variety of less restrictive assumptions on the measurement errors, the reliability-weighted difference (Webster & Bereiter, 1963) and the estimated true change score (Lord, 1956) are not completely identical. However, they state that the discrepancy between the two measures is never very large and that their reliabilities are usually very similar.

Throughout this paper, we will assume that all individuals in the sample have been observed or tested simultaneously and therefore the t_i are the same for all subjects. Bock (1979) has referred to this as "time-structured data."

Strictly speaking, of course, t will only be "midway" among the occasions of measurement when the t_i are either equally spaced or symmetrically placed around their average.

Rogosa & Willett (1985) have shown it to be a parabolic function of t.

Not all systematic interindividual differences in growth need be manifest as linear relations. For example, if \theta_i were a quadratic function of \omega_i, then \beta_\omega would not adequately represent the systematic interindividual differences in growth. However, for simplicity, this paper focuses on linear relations between the parameters of the individual growth model and \omega_i, but the approach could be extended to include more general forms of the conditional expectation in Equation 22.

All the data analyses reported here were performed using SAS Version V.

This data-analytic strategy ignores any possible autocorrelation among the within-individual measurement errors. If such dependence exists, then knowledge of
the autocorrelation could be incorporated into the within-individual estimation by use of weighted least-squares and generalized least-squares regression. However, typically, the covariance structure of the within-individual measurement errors is not known and must be estimated. Such estimation is possible only when there are many time points for each subject in the sample. As Bryk and Raudenbush (1987) have noted, "for data sets with a short time series, the assumption of independent errors with constant variance is often most practical" (see also Berkey, 1982a, 1982b). Whether autocorrelation among the within-individual measurement errors is present or not, OLS estimates of the individual growth parameters remain unbiased.

"Given that the straight-line function is an adequate representation of individual growth.

"Recall that, in the traditional study of individual growth, only two waves of data are collected and the observed difference score is frequently the preferred measure of individual growth. When there are only two waves of data, the slope estimator \( \hat{\beta}_g \) is directly proportional to the observed difference score. However, with only two waves of data, the in situ estimation of measurement error variance from the pooled individual residual sums of squares is not possible, as both of the required residuals are zero. Therefore, an external estimate of the reliability of \( D_g \) is required. As might be anticipated, if the reliabilities of \( X_1 \) and \( X_2 \) are obtained empirically by split-halves estimation on each of the two occasions of measurement (a maximum-likelihood approach—Lord & Novick, 1968), then the traditional approach is identical to a four-wave growth modeling approach with a pair of "half-test" measurements grouped at the beginning of the study and another pair grouped at the end.

"Of course, the obtained increase in reliability is actually a reflection of the improved precision of individual growth parameter estimation in the first stage of the analysis. Thus, in practice, adding more waves of data collection or spreading out the existing waves has twin advantages: more precise empirical knowledge of individual growth is obtained, and the reliability of growth measurement is improved.

"Strictly speaking, to infer the existence of interindividual differences in true growth in the population (i.e., to test the null hypotheses \( H_0: \sigma^2 = 0 \) and \( H_1: \rho_{\text{error}} = 0 \)) both point and interval estimates of \( \sigma^2 \) and \( \rho_{\text{error}} \) are required. In practice, approximate standard errors and confidence intervals can still be obtained by a nonparametric resampling technique such as the jackknife (Efron, 1982; Miller, 1974, 1985).

"The pseudo-Hanushek strategy outlined in this paper is a modification of the strategy actually proposed by Hanushek (1974). Based on earlier work by Goldberger (1964), Hanushek suggests that the sample residual variance in the second-round OLS estimation \( s^2 \) (which is used as a constituent of the weights in Equation 31) should be adjusted to reflect its bias for estimating the variance of the random errors accruing to Equation 25. However, data-analytic experience and algebraic manipulation (based on Goldberger, 1964, pp. 241–243) indicate that, under conditions which frequently prevail in the analysis of educational growth data, the adjustment is relatively unimportant as it typically makes little difference to the overall outcomes of the analysis. Thus, discussion of the adjustment has been avoided here for purposes of clarity and in order to emphasize the intuitive appeal of the weights. Investigators who wish to examine panel data in which there is great variability among the s.e. (\( \hat{\sigma}_g \)) and for whom these latter standard errors have a magnitude similar in magnitude to \( s \) should be cautioned that the adjustment may no longer be trivial (see Hanushek, 1974, Equation 9). In addition, other authors have suggested alternative weighting schemes which may be marginally more appropriate (see de Leeuw & Kreft, 1986)."
REFERENCES


Willett: Measurement of Change


Rogosa, D.R., & Willett, J.B. (1983). Demonstrating the reliability of the difference
score in the measurement of change. *Journal of Educational Measurement*, 20, 335-343.


