Problems of Longitudinal Studies with Children: Practical, Conceptual, and Methodological Issues

Wolfgang Schneider

1. The Need for Longitudinal Studies in the Field of Developmental Psychology

While there seems to be a broad consensus that developmental psychology should focus on changes occurring over time within the organism, the majority of research conducted in the field of developmental psychology has been based on a methodology inappropriate for the study of change (cf. McCall, 1977; Wohlwill, 1973, 1980). That is, most developmental studies cannot be considered truly developmental because they used cross-sectional designs. Consequently, they focused on developmental *differences* among various age groups and ignored developmental *changes* within individuals over age which can only be assessed via longitudinal approaches.

These criticisms represent serious challenges to the purpose of developmental psychology and the way its hypotheses are traditionally investigated and interpreted (cf. Appelbaum and McCall, 1983). Although these criticisms have been around for a while, their impact on current research methodology has been negligible. For example, a recent review on studies conducted in the field of memory development revealed that more than 99 % of these studies have been cross-sectional in nature (cf. Schneider and Weinert, in press).

Interestingly, this does not mean that researchers are still unaware of the problem: Calls for longitudinal studies are frequent in the developmental literature. Given the discrepancy between theory and practice, however, one conclusion could be that there are also various problems with longitudinal studies serious enough to keep off many developmental researchers. The critical analysis of potential problems and possible coping strategies will be a major goal of this chapter. When discussing problems of longitudinal studies, I will not restrict myself to the more general issues typical of most longitudinal investigations, but also refer to problems inherent in longitudinal studies with children.

Discussing problems of longitudinal studies is a complicated matter. First of all, it is difficult because we do not have a precise definition of what constitutes a longitudinal study. Actually, the term "longitudinal" does not describe a simple

method but a broad variety of methods. As Baltes and Nesselroade (1979) pointed out, the spectrum ranges from single-case studies in time-series arrangements to broad-band panel designs including thousands of subjects. Moreover, available longitudinal studies range from repeated single-variable assessment completed within a couple of months to life-span multivariate investigations. The only common denominator of longitudinal research is variation of time and repeated observation of a given entity.

Given the broad variety of research designs subsumed under the label "longitudinal", it follows that the problems discussed in the remainder of this chapter may be relevant for many – so I assume – but not for all longitudinal studies conducted with young children. In my view, there has been considerable confusion about what has to be considered a "true", general, and uncurable problem of longitudinal investigations. It is one goal of the present chapter to illustrate the relativity of many problems and their dependence on research aims. More specifically, it is assumed that problems vary as a function of the respective rationale for longitudinal research. According to Baltes and Nesselroade (1979), there are three different rationales that relate to *description* of development: (1) direct identification of intraindividual change; (2) direct identification of interindividual differences in intraindividual change; and (3) the identification of interrelationships among classes of behavior during development. Two further rationales concern the *explanation* of development: (4) the analysis of causes of intraindividual change; and (5) the analysis of causes of interindividual differences in intraindividual change.

In the remainder of this chapter, general problems as well as problems specific to longitudinal research based on the above rationales will be discussed in more detail. Three classes of general problems will be considered: *Practical problems* concerning cost factors, the long-term recruitment of staff, data storage, and funding; *conceptual problems* referring to the fact that there seems to be no broad consensus among longitudinal researchers about how the concept of change should be defined (note that the solution of this conceptual problem is crucial for the realization of all five rationales for longitudinal research). Finally, general *methodological problems* will be addressed concerning the assessment of change and stability over time. The solution of these problems is equally important for all five research goals mentioned above. As will be shown below, related methodological problems, for example, the choice of adequate statistical tools, depend on the specific goal of longitudinal analysis. While those problems seem relevant to most longitudinal studies, their importance

varies as a function of the type of longitudinal study under consideration.
In addition to these more general problems, longitudinal studies with young children have to cope with more specific problems that are primarily related to the data generation process. Examples from the Munich *L*ongitudinal Study on the *G*enesis of *I*ndividual *C*ompetences (LOGIC; cf. Weinert and Schneider, 1986, 1987) will be used to illustrate problems of verbal assessments (e.g., interviews) with preschool and kindergarten children and their implications for stability of test scores.

2. General Problems of Longitudinal Studies

2.1 Practical Difficulties

Harway, Mednick, and Mednick (1984) summarize the most obvious practical problems of (long-term) longitudinal studies. Among those, the costs associated with conducting longitudinal studies over an extended period of time and difficulties with funding such costly projects are usually considered the major obstacles. According to Harway et al. (1984), this objection to longitudinal research has been typically overrated. In their view, the initial data collection phase is most costly in that staff has to be trained, tasks have to be developed and pretested, the samples have to be recruited, and the research design confined. Costs for subsequent follow-up assessment should be comparably low.
Given the variety of longitudinal designs mentioned above, however, it is difficult to evaluate the importance of the cost problem. Harway et al. (1984) judgment seems adequate for long-term longitudinal studies conducted with a single cohort and including only a few follow-ups or a rather restricted set of test instruments. The situation seems completely different, however, for a more complex longitudinal design. There is little doubt that the cost problem remains a serious practical difficulty for longitudinal studies operating with several cohorts, including a broad variety of test instruments and several measurement points. According to our own experience, costs are even likely to rise in later assessments if data collections in subsequent waves aim at the entire cohort – and do not limit themselves to subsamples, as Harway et al. suggest. The rise in costs experienced in the latter case is mainly due to mobility problems. Researchers interested in keeping the attrition rate low have to spend additional (travel) money in order to keep mobile subjects in the sample.

Given the fact that long-term longitudinal studies are costly enterprises, obtaining and maintaining funding is not an easy task. Usually, a necessary (but not sufficient) condition for obtaining funding is an elaborated research design combined with a sophisticated developmental theory. It is most important to convince reviewers that the planned study has to be longitudinal and cannot be replaced by a series of related cross-sectional studies. As there is an obvious lack of information on intraindividual changes in many developmental areas, the task seems difficult but solvable in principle.

With regard to staffing, the likelihood that only a few staff members will stay with a longitudinal study for the duration of the project is not a real problem. Although shifts in personnel are often costly, periodic changes in personnel are not necessarily unhealthy. The occasional addition of "fresh blood" does not only minimize the risk of data bias due to frequent interactions among experimenters and subjects in a long-term longitudinal study, but also increases the possibility that already existing data will be analyzed in ways not considered by permanent staff members (cf. Harway et al., 1984).

Two other practical difficulties frequently mentioned by opponents of longitudinal research refer to the publication record of longitudinal investigators and the timeliness of long-term longitudinal enterprises. In view of the "publish-or-perish" principle guiding scientific careers, a longitudinal researcher may be in a bad position because it usually takes several years before the first results are available for publication. Even worse, the problem of timeliness of publication arises in the case of long-term longitudinal studies in which it takes a long time before the harvesting of data is possible. In those cases, publishing results could be a difficult enterprise because the topic under investigation may be "out of fashion"; a problem frequently encountered in various disciplines of social sciences. However, there are solutions to both problems. More specifically, it is of crucial importance to design longitudinal studies in a way that: (1) there remains sufficient time for analyzing the data and writing up reports between two adjacent measurement points, and (2) the design is flexible enough to allow for the possibility of change. Accordingly, an extreme delay in publication can be avoided if sufficient time is available for the different phases of a longitudinal study (i.e., data collection, data analysis, report writing). With regard to flexibility, precautions should be taken to ensure that the study is not too narrow in scope. That is, a broad-band investigation including several measures from different domains copes with the problem of timeliness in that the data

might be reanalyzed at a later date and interpreted in the light of theoretical and technological advances (cf. Block and Block, 1980, 1984; Harway et al., 1984). All in all, there is no doubt that there are several practical difficulties with longitudinal studies. However, as there also seem to be practicable solutions to most of the problems discussed in this section, those problems should not be overestimated by researchers interested in conducting longitudinal research.

2.2 Conceptual Problems

All five rationales of longitudinal research listed above referred to the assessment of change. At first glance, there seems to be no problem with conceptualizing and studying developmental change. A closer look at the literature, however, reveals that conceptualizing human development is a complicated issue. As emphasized by Baumrind (1987), instability and discontinuity in human development can only be seen against a background of stability. The question of what stability can mean in the context of changing individuals (Wohlwill, 1980) has been answered differently by different researchers.

Given the space restrictions, I do not want to reiterate the discussion of conceptual discrepancies but summarize the existing consensus (cf. for a more detailed discussion Asendorpf, in press a; Kagan, 1980; Overton and Reese, 1981; Rutter, 1987; Wohlwill, 1973, 1980).

There is a general consensus that two types of longitudinal inquiry can be distinguished. One aspect of developmental inquiry concerns what Wohlwill (1973) called the developmental function, that is, the average value of a dependent variable plotted over age. Typical examples would be growth curves for physical height or weight based on a sample of individuals. The second realm of developmental inquiry concerns individual differences. More specifically, the question is whether individual subjects maintain approximately the same relative rank ordering within their group at one age as they do at another (cf. Appelbaum and McCall, 1983; McCall, 1977). McCall (1977) used the term *continuty/discontinuity* to refer to the developmental function, whereas *stability/instability* refers to the individual differences approach. Other researchers have further differentiated among various meanings of the stability concept: They make a distinction between the stability of a variable and the stability of an individual (Wohlwill, 1980), and further refer to ipsative stability as the

persistence of a pattern of variables for an individual subject over time (cf. Asendorpf, in press a, 1987; Kagan, 1980; Rutter, 1987).

The important distinction to be emphasized here is between the continuity or discontinuity of a growth function for an attribute and the degree of stability or instability of individual differences in an attribute. Note that continuity/discontinuity in developmental function is *conceptually* independent from stability/instability in individual differences: The relationship between both concepts is an *empirical* question (cf. Appelbaum and McCall, 1983).

Longitudinal researchers have frequently overlooked the fact that developmental functions and individual differences represent two separate aspects of the same problem. Appelbaum and McCall (1983) provide examples for such confusions. One general problem of longitudinal research is the tendency to emphasize information on the stability of individual differences at the expense of data on developmental change. That is, most longitudinal researchers focused on predicting later differences in a given variable without considering the fact that this could not tell them anything about the developmental functions of that variable. McCall (1977) correctly stated that many longitudinal studies cannot be considered truly developmental because they ignored developmental change in individual differences over age.

In the present context, it is important to note that most longitudinal studies either focus on individual differences or the developmental function. Studies simultaneously combining these two aspects are rare. With regard to methodological problems, the conclusion is that only a few problems to be discussed below are representative of the two types of longitudinal studies. For example, the extensive literature on the problems of assessing individual changes over time only refers to those empirical studies dealing with data on the developmental function. The problem with measuring change has not been a relevant topic for the many longitudinal studies dealing with individual differences. The reader should keep in mind that many problems of longitudinal research to be discussed below are relative rather than general.

2.3 Methodological Problems of Longitudinal Research

Given the conceptual problems discussed above, it is not surprising that longitudinal research in the social sciences has been dominated by several fundamental misunderstandings and damaging myths (cf. Rogosa, 1988). Many problems

are related to the use of inadequate designs or inappropriate statistical tools. Appelbaum and McCall (1983), for example, emphasize the fact that applied statistics has made considerable progress within the last few years, and that researchers engaged in longitudinal studies should become acquainted with more recent statistical developments useful for the study of change. Thus, narrowing the knowledge gap between statisticians and researchers is considered a precondition for improving longitudinal research. In the following, I will first refer to several well-established "myths" of longitudinal research and then go on to more serious problems related to the assessment of change.

2.3.1 Problems with Measuring Change

The debate about the measurement of change has a long tradition in the psychometric literature. Since the classical article by Cronbach and Furby (1970), longitudinal researchers have been warned repeatedly of the hazards of change scores. These warnings seem to have created the belief that change scores are unreliable, misleading, and unfair, and therefore should be avoided at all costs (cf. Maxwell and Howard, 1981). However, several methodologists have provided evidence that it is high time to debunk this myth.

For example, Maxwell and Howard (1981) showed that the analysis of change scores is valid in randomized pretest-posttest designs. Social scientists familiar with the hazards of change scores may prefer a repeated measures ANOVA over an ANOVA on posttest-pretest change scores, assuming that any problem with change scores are avoided with the repeated measures ANOVA. They are obviously unaware of the fact that an ANOVA on posttest-pretest change scores yields exactly the same F value as obtained from the interaction test of the repeated measures design (cf. Maxwell and Howard, 1981; Nunnally, 1982). Accordingly, there is no problem with using change scores for *group* analyses.

However, what about the use of difference scores for the assessment of *individual* change? Again, there is strong evidence that problems have been overestimated. Proponents of difference scores emphasize the fact that they constitute the very heart of longitudinal investigations and represent unbiased estimates of true change (Nunally, 1982; Rogosa, 1988; Rogosa, Brandt, and Zimowski, 1982). Indeed, it can easily be shown that several objections (e.g., unreliability, unfairness, bias through regression toward the mean) do not generally hold.

For example, Rogosa et al. (1982) demonstrated that the reliability of difference scores is not generally low. The difference score will have low reliability as long as individual growth rates vary little across subjects. In this case, reliability indicates the accuracy with which subjects can be ranked on the growth rate function on the basis of their difference scores, whether the estimates of the growth rate function are precise or not. The important message is that low reliability does not necessarily imply lack of precision, that is, does not preclude meaningful assessment of individual change. Moreover, the reliability of the difference score is respectable when considerable individual differences in change are present (for illustrations see Rogosa and Willett, 1983).

Similarly, it has been shown that the importance of regression-toward-the-mean effects for the study of change has been overestimated in the social sciences literature (cf. Nesselroade, Stigler, and Baltes, 1982; Rogosa, 1988). While there is a lack of explicit descriptions of the phenomenon, the traditional meaning is that, on average, you are going to be closer to the mean at Time 2 than you were at Time 1. The crucial message provided by Nesselroade et al. (1982) is that regression toward the mean is not an ubiquitous phenomenon that has unalterable effects. On the contrary, Nesselroade et al. demonstrated that the often-held belief that measurement error necessarily produces a regression effect that makes it impossible or difficult to measure change properly is not correct, at least not with multiwave data.

It should be noted in this connection, that the popular assumption that residual change scores should be chosen instead of difference scores because they adjust for effects of individual differences in initital status is problematic. As Rogosa (1988; Rogosa et al., 1982) points out, there are logical problems with the residual change approach as well as statistical and psychometric shortcomings. Logically, the question "How much would individual p have changed on attribute x if all individuals had started out 'equal'" stimulates the subsequent question "Equal on what?". Does it mean equal on true initial status, observed initial status, initial status in combination with other background variables, or what? The correct answer is unknown. Obviously, addressing the question "How much did individual p change on attribute x?" is comparably simple (cf. Rogosa, in press). See Rogosa et al. (1982) as well as Rogosa and Willett (1985a) for a detailed treatment of statistical and psychometric shortcomings of residual change scores.

All in all, methodological papers written in defense of the difference score have accumulated over the past few years. The important message for the longitudinal researcher is that more effort should be invested into developing models of individual growth (cf. Bock, 1976; Bryk and Raudenbush, 1987; Rogosa et al., 1982) and constructing proper longitudinal designs. In general, two measurement points provide an inadequate basis for studying change. Data collected on multiple occasions are better suited to control for regression effects and estimate individual growth curves (cf. Nesselroade et al., 1980; Rogosa et al., 1982). To my knowledge, Bryk and Raudenbush's (1987) two-stage model of growth represents one of the most promising approaches to the study of change. It allows for studying the structure of individual growth, examining the reliability for measuring status and change, investigating correlates of status and change, and testing hypotheses concerning the effects of background variables. Given the impressive demonstrations presented by Bryk and Raudenbush, longitudinal researchers should be encouraged to adopt such a hierarchical linear modeling approach for their studies which seems broadly applicable to the study of change.

So far, our discussion has been restricted to change scores based on classical test theory. Note that probabilistic measurement models provide an alternative possibility to quantify individual change. Various versions of the linear logistic test model (Rasch model) exist that allow for unbiased estimation of item difficulty parameters and person ability parameters (cf. Fischer, 1976; Fischer and Formann, 1982). In these Rasch models, changes in either abilities or item difficulties over time can be simply assessed by analyzing the respective difference scores (see Schneider and Treiber, 1984, for an empirical example).

To summarize, there seem to be several possibilities of correctly assessing developmental change or the developmental function, mainly due to recent methodological advancements. Potential pitfalls have been definitively overrated by many researchers.

It should be noted, however, that this conclusion only holds if change scores are measured on a common scale. More specifically, the precondition for the analysis of change scores is that the same measurement instruments were used over time. It is obvious, then, that longitudinal studies with children run into serious problems whenever tests or questionnaires are designed for a restricted age range. While the same instrument can be used at all ages with certain measurements like height or weight, this is not true for the majority of meas-

ures designed to assess cognitive or personality development in children. Although Goldstein (1979) discusses the possibility of constructing a common scale for different instruments by using various transformation procedures, this only allows for the assessment of *relative* change, that is, for a comparison of subgroups of a given population. Consequently, while the use of different measurement instruments over time does not cause problems for the longitudinal analysis of individual differences, it definitively restricts the analysis of the developmental function.

But even when the same instrument is used on each occasion, the interpretations of it may differ (cf. Magnusson, 1981). As Baumrind (1987) emphasized, a variable or construct may appear to be the same at various ages when in fact it is not. An example referring to motor development may illustrate the case: There is no doubt that crawling has a different meaning for a 9-month-old child than it has for a 4-year-old child. Obviously, the neglect of *qualitative* change or discontinuity in the organization of individual behavior can lead to erroneous conclusions (see also Rutter, 1987). Block, Gjerde, and Block (1986) found considerable transformations in the psychological meaning of indicators of categorization breadth from age 4 to age 11. While the use of relatively broad categories in early childhood reflected an inability to organize experience effectively, the use of relatively broad categories in preadolescence reflected a rather creative ability (see Baumrind, 1987, for a similar empirical example). Baumrind (1987) and Block et al. (1986) recommend the use of multiple and diverse measures of behavioral constructs in order to examine whether continuity across time periods is given or not. Methodologically, this means that a cluster of variables defining a construct must load on second-order factors in the same way across time to have the same meaning and validity.

2.3.2 Problems of Longitudinal Studies Analyzing Individual Differences

As already mentioned above, longitudinal research focusing on individual differences is concerned with different definitions and types of stability. In most cases, correlation coefficients (e.g., Time 1 – Time 2 correlations) are used as measures of the consistency of individual differences. Further methods include repeated measures ANOVA, cross-lagged panel correlation analysis, path analysis regression, and structural equation models using latent variables.

Several myths or misunderstandings relate to the interpretation of the correlation coefficient. While there are several limitations with correlation coefficients and problems concerning their interpretation (cf. Rutter, 1987; Valsiner, 1986, one fundamental misunderstanding is that the correlation matrix for longitudinal data tells you whether or not you are measuring the same thing over time (Rogosa, 1988). Theoretically, it is possible that the rank ordering of individuals within a given group remains constant over time (indicated by a large correlation coefficient), although the theoretical construct under study changes its meaning for the subjects. Of course, the opposite could be also true. Consequently, more elaborated validation procedures (e.g., assessment of related reference variables) seem necessary to ensure that correlation coefficients are correctly interpreted.

A further misunderstanding concerning methods of the individual differences approach is that structural regression models tell us much about change, or that cross-lagged panel correlation procedures inform about reciprocal causal effects (cf. Rogosa, 1980, 1985, 1988). In the first case, the myth to be debunked is that structural parameters can be indicative of individual growth rates in observed or latent variables. As illustrated by Rogosa (1988), structural regression coefficients can actually badly mislead about exogeneous influences on growth. The message is that the analysis of correlations or covariance structures should not be undertaken to reach conclusions about individual growth. Rather, they should be used to investigate stability or consistency issues and be concerned with the prediction of events.

The myth concerning the potential of cross-lagged panel correlations seems even more popular. Roughly speaking, the research question is whether variable x causes y or vice versa. Thus studies of reciprocal effects investigate problems of causal predominance or causal ordering. In the two wave – two variable case typically used to illustrate the problems with the procedure, r_{x1y2} and r_{y1x2} represent the sample cross-lagged correlations of specific interest. The attribution of causal predominance is based on the difference between the two cross-lagged correlations. Usually, causal predominance is assumed when the null hypothesis of equal cross-lagged correlations is rejected. Rogosa (1980, 1985) has convincingly demonstrated serious methodological flaws of the procedure. Accordingly, cross-lagged panel correlations cannot be recommended in order to detect patterns of causal influences and should best be forgotten. Fortunately, several alternatives are available. In particular, structural equation modeling (SEM) procedures have been developed recently that can be used to systematic-

ally develop and test theories (cf. Bentler, 1980, 1985; Jöreskog and Sörbom, 1984; Lohmöller, 1984). SEM procedures using a latent variable approach like LISREL or EQS not only seem better suited for the analysis of reciprocal causal effects but are also appropriate for estimating and testing more complex causal models including intervening variables. While structural equation models can principally be applied to cross-sectional data, they seem promising when used with longitudinal data. In short, their major advantages – as compared to traditional regression analysis – are that: (1) a verbal theory has to be translated into a mathematical model that can be estimated; (2) structural/causal relationships are estimated at the level of latent variables or theoretical constructs and not on the basis of fallible observed variables; (3) the distinction between a measurement model describing the relationships among observed variables and latent factors and a structural model describing interrelations among theoretical constructs also allows for a separate estimation of measurement errors in the observables and specification errors in the structural part of the model: large specification errors usually indicate that the causal model is not completely specified, that is, important predictor variables are obviously missing; and (4) Several so-called goodness-of-fit tests exist that detect the degree of fit between the causal model and the data set to which it is applied. Causal models are said to be "confirmed" when the goodness-of-fit parameter indicates better-than-chance fit between the model and the data. (See the Special Section on Structural Equation Modeling in the first issue of *Child Development*, 1987, for a more detailed description of SEM procedures and for numerous applications drawn from different areas of developmental psychology.)

While SEM procedures generally operate on correlation or covariance matrices, mean structures can also be considered. For example, McArdle and Epstein (1987) illustrate the possibilities of a longitudinal model that includes correlations, variances, and means and is described as a latent growth curve model. The inclusion of mean structures makes this longitudinal model more similar to repeated-measures ANOVA and MANOVA traditions. As a consequence, this type of model may also be used to assess the developmental function, that is, group changes in the amount of a latent variable over time.

Although there is broad agreement that SEM procedures represent powerful general tools for the analysis of longitudinal data, they should not be conceived

of as panaceas. Several potential problems with SEM procedures have been addressed in the literature (cf. Connell, 1987; Martin, 1987; Rogosa, 1988). First of all, the use of a multiple indicator approach based on an elegant statistical model cannot compensate for poor-quality data, careless operationalization of major constructs, and inappropriate designs (cf. Martin, 1987; Rudinger and Wood, 1987). It is the researcher who has to make sure that theoretical assumptions are justified, and that the longitudinal sampling of occasions and variables allows for the discrimination between interesting alternatives. Further, some SEM procedures (e.g., LISREL) require multivariate normality of data, a criterion rarely met in the case of longitudinal data (cf. Bentler, 1986, 1987). In addition, the effectiveness of most SEM procedures depends on the accessibility of large data sets (more than 100 subjects as a rule of thumb). Unfortunately, SEM procedures applicable to nonnormally distributed data (e.g., EQS) require even larger sample sizes (cf. Tanaka, 1987). As Tanaka puts it, the "cost" of making fewer distributional assumptions about data is the necessity of a large sample size. It should be noted, however, that this restriction does not apply to distribution-free, exploratory SEM procedures also known as soft-modeling procedures. For example, causal models with latent variables based on partial-least-squares (PLS) estimation procedures can be used as a starting point whenever theoretical knowledge is scarce and/or only small samples are available (cf. Lohmöller, 1984; Schneider, 1986).

One last problem to be mentioned concerns the adequate assessment of reciprocal causal effects discussed earlier in the paper. Traditionally, first-order autoregressive or simplex models have been argued to be optimal models for studying stability and change in developmental applications (cf. Rogosa, 1988). In these models, variables are represented as causes of themselves over two or more points in time. As it is well known that autoregressive models define changes over time to be independent of prior changes, they do not seem to be an optimal method for assessing change in many developmental applications. Note that the growth curve models introduced by McArdle and Epstein (1987) seem preferable in that they define changes over time to be dependent on the prior changes. Further, as Hertzog and Nesselroade (1987) illustrated, simplex models may be a particularly poor way of representing change and reciprocal causal relationships between state (nontrait) phenomena. Finally, there is also evidence that it is often too easy to fit autoregressive models (including cross-lagged regression models) to longitudinal data. For example, Rogosa and

Willet (1985b) demonstrated that a simplex model marvelously fit a covariance matrix from growth curves that were maximally "unsimplex".

All in all, the problems presented so far have illustrated that careful theoretical analyses are a prerequisite for an adequate model building process. Of course, the quality of data available for analyses further complicates the issue. Recent developments in SEM procedures, however, seem to minimize this problem. That is, long-linear path analysis models for discrete data or causal models with categorical/nonnormal dependent variables are now available for longitudinal researchers who cannot rely on continuous variables (cf. Goldstein, 1979; Muthén, 1984, 1987). Altogether, the number of statistical techniques available for the analysis of discrete longitudinal data has considerably increased during the last decade (cf. for reviews Henning and Rudinger, 1985; Markus, 1979).

To summarize, there seem to be relatively few problems with analyzing longitudinal data focusing on individual differences and stability/instability issues. Due to recent methodological advancements, generally applicable and elegant statistical tools are available that can be used for elaborate model building. This is particularly true for the assessment of what Hertzog and Nesselroade (1987) call *mean* stability or *covariance* stability over time. Relatively little attention has been paid to *intraindividual* stability, that is, change within the given sampling unit (but see Asendorpf, in press a, b for the construction of a coefficient assessing individual stability over time).

3. Specific Problems of Longitudinal Studies With Young Children: The Case of the Munich Longitudinal Study on the Genesis of Individual Competencies (LOGIC)

In the last section of this chapter, I will focus on problems typical of longitudinal studies with preschool and kindergarten children. Although these problems are well-known to most researchers working with young children, they are rarely mentioned in scientific reports. In particular, problems related to the test situation, to practice and experimenter effects will be discussed in more detail. Moreover, problems related to the instability of test scores and the implications for model building and prediction purposes will be addressed.

As already noted, empirical examples demonstrating some of the problems typical of longitudinal studies with young children will be taken from our

Munich longitudinal study on the genesis of individual competencies (LOGIC). In this study, a sample of about 220 four-year-old subjects was first tested immediately after the children had entered kindergarten. This was done to make sure that the subjects' experience with social groups was limited at the very beginning of the longitudinal study. As one of the major goals of LOGIC was the study of the effects of social group experiences on cognitive development, it seemed important to start at this particular point in time. Since 1984, children have been annually tested on a broad variety of variables, including measures of intelligence, memory and metamemory, social cognition, social competence, moral development, and achievement motivation. In order to identify important prerequisites and determinants of school (reading) achievement, a number of experimental tasks tapping different aspects of phonological processing (i.e., children's use of phonological information in processing oral or written language) were additionally included at the third measurement point, that is, at the end of the kindergarten period. The study is designed to be active until the end of elementary school (4th grade). The major methodological problems experienced during the kindergarten phase of the study are summarized below.

3.1 Problems Related to the Test Situation

It is well-known from cross-sectional studies that testing or interviewing young children is a complicated matter. Given the large number of experimental tasks, psychometric tests, and interview procedures included in LOGIC, it turned out to be an extremely difficult task to keep our subjects motivated and interested in the study. One problem repeatedly encountered during the first measurement point was that children felt insecure and uncomfortable in the test situation. It usually took a long warming-up phase before children were ready to participate in the various tasks and answer the numerous questions. In several cases, however, even extended interactions with the child before testing did not have the expected effects: A small number of 4-year-old children refused to answer any questions in several test situations, particularly in the more difficult and demanding interviews (e.g., the metamemory interviews). They also did not reproduce any items in several experimental tasks (e.g., recall of texts or word lists). How does one proceed with such "untestable" children? They surely cannot be treated like missing cases because they did not actually miss the test session. On the other hand, it seems obvious that their poor performance in the

test situation represents a seriously biased estimate of their true competence. In other words, the problem of measurement error seems particularly serious in the case of "untestable" children.

A related problem concerns the experimenter or observer effect. Even if young children are willing to participate in an interview or test session, it may be that they only want to interact with certain experimenters. Methodologically, one "disadvantage" of an extended warming-up procedure is that young children get used to their adult partners. As a consequence, they do not want to be tested by other experimenters in subsequent sessions or measurement points. This phenomenon has been frequently observed in our study. The problem is that it is difficult if not impossible to control for experimenter or observer effects. Again, this means that there is systematic bias and additional noise in the data. Ways of coping with this problem will be described below.

3.2 Problems with Instability of Test Scores

An analysis of the LOGIC memory data (Schneider and Weinert, in press) revealed considerable test instability over time. That is, retest correlations for measures of word list recall, text recall, and memory span assessed at ages 4 and 6 were rather low with coefficients ranging between .22 (word span) and .36 (word list recall). This finding indicates that the subjects did not maintain their relative standing within their group, and that individual differences were not preserved between age 4 and age 6. These results are not in accord with those reported by Kunzinger (1985), who reported high across-age group stabilities for his sample of elementary school children between ages 7 and 9. Additional analyses concerning *individual* stability, that is, the amount of across-age variable shown in an individual's relative standing within the referent group, did not change the overall pattern of results. So-called "lability scores" (Kunzinger, 1985; Wohlwill, 1973), representing the across-age standard deviation of an individual's *z* scores, were computed for the various recall measures, and yielded considerably higher values (i.e., high lability) than those reported by Kunzinger (1985). This trend was not restricted to our memory data. When lability scores were computed to assess the across-age stability in text recall, verbal intelligence, and motor skills, we found them to be almost three times as high as those obtained by Kunzinger (1985). Interestingly, lability scores were

comparable across the three tasks considered. It seems, then, that high levels of instability are not only typical of memory performance at that particular age, but can be generalized across different domains (Tab. 1).

Tab. 1: Individual Across-Age Lability of Test Score for Selected Variables of the LOGIC Study, split for Sex (N = 208)

Variable	Boys	Girls	
General metamemory	.71	.55	s
Sorting in a sort-recall task	.79	.71	ns
Recall in a sort-recall task	.59	.64	ns
Text recall	.63	.66	ns
Memory span	.64	.61	ns
Verbal intelligence	.46	.50	ns
Nonverbal intelligence	.69	.74	ns
Motor skills	.52	.50	ns
Social competence	.87	.72	ns

Note: "s" indicates that sex differences were significant at the p = .05 level;
"ns" indicates that no sex differences were found.

Given the high instability of test scores over time, an interesting question is whether this is due to high unreliability of measurement instruments or rather to the high fluctuation of the phenomenon under study. Note that low stability over time does not necessarily mean low reliability or low internal consistency of the measure in question: Hertzog and Nesselroade (1987) used SEM procedures to demonstrate that so-called state measures like anxiety or fatigue showing low stability over time can nevertheless be reliably and validly assessed.

Interestingly, in the case of young children, trait measures such as indicators of psychometric intelligence seem to "behave" like the state measures in the case of Hertzog and Nesselroade's (1987) study with older adults: Whereas their stability over time is rather low, their internal consistency is not.

The situation seems more complicated, however, when internal consistency of measures cannot be assessed, that is, whenever the measures of interest consist of only one of a few items. Unfortunately, this is true for many experimental tasks (e.g., measures assessing different aspects of memory). In those cases, short-term, test-related correlations should be obtained to make sure that the variables of interest are indeed measured reliably (cf. also Asendorpf, in press a).

Problems with low stability of test scores obtained from young children are not restricted to longitudinal assessment. We found similarly low intraindividual consistency across related tasks in cross-sectional settings. For example, only weak to moderate intercorrelations were obtained when memory tasks tapping similar skills were compared for one point in time.

Young children's low intraindividual consistency across similar measures represents a serious problem when the researcher's goal is to predict future performance. In the LOGIC study, for example, the Bielefeld Screening Test developed by Marx and Skowronek (this volume) was used in the last period of kindergarten to identify children at risk, particularly with regard to reading and spelling. The screening procedure consisted of nine different tests assessing various aspects of phonological processing (e.g., rhyming, syllable segmentation, visual word matching, or sound blending). Although intertask correlations were moderate to high, there were only a few children scoring consistently low on most measures. Tab. 2 illustrates the degree of inconsistency observed for the screening test measures. We checked the number of times each child belonged to the bottom 10% of the distribution in the nine subtests of the screening test. As can be seen from Tab. 2, there were only 8 out of 208 subjects who belonged to the bottom 10% in more than five out of nine subtests. Given the low intraindividual consistency across measures, it seems that the early prediction of school achievement is a difficult task.

Tab. 2: Number of Times a Subject Belonged to the Bottom 10% in the Various Subtests of the Bielefeld Screening Test (N = 208)

N of times	N of subjects	Percent
0	99	47.6
1	60	28.8
2	21	10.1
3	11	5.3
4	9	4.3
5	4	1.9
6	2	1.0
7	2	1.0

3.3 Possible Coping Strategies

How can we cope with all the problems related to longitudinal studies with young children? It seems that the risk of working with measures swamped by error is particularly high in studies using preschoolers and kindergarteners as subjects. To enhance the reliability and validity of test scores, several measures can be taken (cf. Block and Block, 1980; Block et al., 1986). First, the use of multiple kinds of data seems advantageous. If test data, observational data, self-report, and questionnaire data are available that all refer to the same theoretical concept, the chances of getting closer to the "true" score increase. Similarly, multiple measurement within each kind of data seems suited to reduce error variance. Block et al. (1986) refer to the psychometric truism that the proportion of concept-related variance in a given measure can be improved by basing that measure on an average or composite of a number of concept-related items, each of which may contain only a small proportion of concept-related variance. Moreover, the use of multiple measures also allows for more sophisticated model building and theory testing via SEM procedures.

In addition, it seems important to draw independent cross-sectional samples to check generalizability of findings and to assess possible practice effects. Furthermore, this measure could be useful for identifying cohort effects. How-

ever, while controlling for cohort effects seems a very important problem of life-span longitudinal studies (see Baltes, Cornelius, and Nesselroade, 1979, for a detailed treatment of this point), they do not seem equally relevant in longitudinal studies with young children conducted within a comparably restricted time interval. Nonetheless, the recruitment of independent samples in studies with young children may well be informative in that the existence of age-group-related influence patterns can be examined.

4. Concluding Remarks

All in all, the discussion of selected practical, conceptual, and methodological problems of longitudinal studies with children revealed that the importance of different kinds of difficulties has been overrated in the literature. This is particularly true of methodological problems: As has been shown above, several myths concerning the problems of change scores or the analysis of cross-lagged panel data still exist and need to be debunked. Recent developments in applied statistics have made it possible to effectively deal with change scores and also provide several possibilities for the analysis of reciprocal causal effects. Thus the message is that most methodological problems of longitudinal studies can be successfully handled, regardless of whether the focus is on the developmental function or individual differences.

On the other hand, the overview of the literature also revealed that longitudinal studies have to be planned carefully in order to be successful in the long run. For example, many practical problems reported in the literature seem related to poor planning efforts; for example, insufficient time lags between data assessment and data analysis. Other studies suffered from the problem that their long-term goals were never precisely defined. That is, it remained unclear in those cases whether the focus was on developmental changes or on individual differences. As mentioned above, this decision should be made early in the planning process because it definitively affects the choice of measurement instruments. For example, the selection of conceptually related but different measures does not cause major problems in studies dealing with individual differences, but seems disadvantageous in studies focusing on the developmental function. In fact, one of the few serious problems discussed in this chapter

concerns the question of how to build up a common scale for different measures tapping the same underlying construct.

Taken together, however, it seems that numerous coping strategies are available that can deal effectively with potential problems of longitudinal work. Given the unique importance of longitudinal data for our proper understanding of child development, we can only hope that the powerful tool of longitudinal analysis will be more frequently used in future developmental studies than it is today.

References

Appelbaum, M.I., and McCall, R.B. (1983): Design and analysis in developmental psychology. In P.H. Mussen (Ed.), *Handbook of child psychology* (Vol. 1, pp. 415–476). New York: Wiley

Asendorpf, J. (in press a): Individual, differential, and aggregate stability of social competence. In B.H. Schneider, G. Attili, J. Nadel, and R. Weissberg (Eds.): *Social competence in developmental perspective*. Dordrecht: Kluwer

Asendorpf, J. (in press b): Coefficients of individual and differential stability. *Methodika*

Baltes, P.B., Cornelius, S.W., and Nesselroade, J.R. (1979): Cohort effects in developmental psychology. In J.R. Nesselroade, and P.B. Baltes (Eds.), *Longitudinal research in the study of behavior and development*. New York: Academic Press

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

Baumrind, D. (1987): *The permanence of change and the impermanence of stability*. Paper presented at the biennial meetings of the Society for Research in Child Development, Baltimore

Bentler, P.M. (1980): Multivariate analysis with latent variables. Causal modeling. *Annual Review of Psychology, 31*, 419–456

Bentler, P.M. (1985): *Theory and implementation of EQS: A structural equations program*. Los Angeles: BMDP Statistical Software Corp.

Bentler, P.M. (1986): EQS – Ein Ansatz zur Analyse von Strukturgleichungsmodellen für normal- bzw. nichtnormal verteilte quantitative Variablen. In C. Möbus and W. Schneider (Eds.), *Strukturmodelle für Längsschnittdaten und Zeitreihen*. Bern: Huber-Verlag

Bentler, P.M. (1987): Drug use and personality in adolescence and young adulthood. Structural models with nonnormal variables. *Child Development, 58*, 65–79

Block, J., Gjerde, P.F., and Block, J.H. (1986): Continuity and transformation in the psychological meaning of categorization breadth. *Developmental Psychology, 22*, 832–840

Block, J.H., and Block, J. (1980): The role of ego–control and ego–resiliency in the organization of behavior. In W.A. Collins (Ed.): *Minnesota Symposia on Child Psychology* (Vol. 13). Hillsdale, NJ: Erlbaum

Block, J.H., and Block, J. (1984): A longitudinal study of personality and cognitive development. In S.A. Mednick, M. Harway, and K.M. Finello (Eds.): *Handbook of longitudinal research, Vol. 1: Birth and childhood cohorts* (pp. 329–352). New York: Praeger