

Modeling Longitudinal and Multilevel Data

*Practical Issues, Applied Approaches
and Specific Examples*

Edited by

Todd D. Little
Yale University

Kai U. Schnabel
Jürgen Baumert

Max Planck Institute for Human Development



2000

LAWRENCE ERLBAUM ASSOCIATES, PUBLISHERS
Mahwah, New Jersey

London

Copyright © 2000, by Lawrence Erlbaum Associates, Inc.

All rights reserved. No part of the book may be reproduced in any form, by photostat, microform, retrieval system, or any other means, without prior written permission of the publisher.

Lawrence Erlbaum Associates, Inc., Publishers
10 Industrial Avenue
Mahwah, NJ 07430

Cover design by Kathryn Houghtaling Lacey

Library of Congress Cataloging-in-Publication Data

Modeling longitudinal and multilevel data : practical issues, applied approaches, and specific examples / edited by Todd D. Little, Kai-Uwe Schnabel, Jürgen Baumert.

p. cm.

ISBN 0-8058-3054-5 (cloth : alk. paper)

1. Social sciences—Statistical methods 2. Longitudinal method. I. Little, Todd D. II. Schnabel, Kai-Uwe. III. Baumert

QA76.76.E95S32 2000

006.3'31—dc21

97-5613

CIP

Books published by Lawrence Erlbaum Associates are printed on acid-free paper, and their bindings are chosen for strength and durability.

The final camera copy for this book was prepared by the author and therefore the publisher takes no responsibility for consistency or correctness of typographical style.

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

CHAPTER TEN

Selectivity and Generalizability in Longitudinal Research: On the Effects of Continuers and Dropouts

Todd D. Little
Yale University

Ulman Lindenberger
Heiner Maier
Max Planck Institute for Human Development, Berlin

Perhaps the quintessential goal in science is to ensure the accuracy of a study's results. When this goal is met, a study's conclusions can be generalized to a larger space of potential measurements and ultimately to an intended spectrum of generalization instead of being restricted to just the observed events and the sample in question. This goal is encapsulated in concepts such as *measurement representativeness*, which refers to the degree to which observations can stand for other nonmeasured events (McArdle, 1994), and *validity*, which refers to the degree of veridicality of conclusions that are drawn from empirical findings.

A common but often neglected threat to generalizability in longitudinal research is sample selectivity, or nonrandom participation. In this context, selectivity refers to the potential for systematic differences between continuers and dropouts. Although selective participation is only one possible factor endangering a study's validity, in our view, it is also one of the least considered threats to the representativeness and generalizability of longitudinal investigations. Few empirical studies have examined selectivity effects fully or attempted to quantify their influences on variables measured at later time points (cf. Lindenberger et al., 1999; McArdle & Hamagami, 1991; McArdle, Hamagami, Elias, & Robbins, 1991). In this chapter, our primary goal is to emphasize the importance of examining selectivity and to review general procedures to analyze their effects (see also Lindenberger et al., 1999).

Overview of the Problem

Sample attrition occurs when not all persons who have been asked to participate in a longitudinal study partake in all assessments. Attrition can lead to selectivity (or bias) if the persons who continue to participate (i.e., *continuers*) differ from those who do not (e.g., *dropouts*) on any characteristics that are relevant to a study (Kessler, Little, & Groves, 1995; Little & Rubin, 1987). Selectivity stemming from sample attrition is a unique threat to validity that is quite distinct from other sources of possible bias such as those related to sampling procedures (i.e., selective sampling vs. selective dropout; Baltes, Reese, & Nesselroade, 1988) or measurement and analytic procedures (T. D. Little, Lindenberg, & Nesselroade, in press). For example, bias related to nonrepresentative sampling of persons is well understood. Here, as is commonly acknowledged, random sampling procedures (i.e., whereby each person in a population has an equal chance of being selected) provide sound ways of minimizing systematic sampling-related biases (Kruskal & Mosteller, 1979a, 1979b, 1979c). Generally speaking, random sampling is advantageous because systematic sources of bias are much less likely to occur than with nonrandom sampling, thereby guarding against selectivity, or invalidity, due to sampling.

In typical longitudinal studies, however, researchers have little control over who drops out and who does not, and this process may or may not occur randomly. If (non)participation is random (i.e., unsystematic) then, all other things being equal, the integrity of a study would be maintained. If (non)participation is nonrandom, then, independent of the degree of generalizability associated with other aspects of a study, the effects of sample attrition would lead to some degree of bias. Under such conditions, conclusions based on selective samples are no longer fully accurate reflections of the original parent sample or the intended population. When sample attrition has led to some degree of sample selectivity, all conclusions would need to be qualified accordingly (i.e., acknowledging and/or accounting for the estimated degree of observed selectivity).

A related problem in longitudinal designs is the effect that continuers have on the outcomes of a study. Those who continue may engender characteristics that also bias the study's results. For example, continuers have more exposure to various aspects of the testing situation. Repeated exposure can lead to a number of unwanted influences, from practice and memory effects to reactivity and boredom. Although these influences are not strictly related to selectivity influences, they are associated with the selective status of being a continuing participant. Therefore, not only do the characteristics of dropouts affect the outcomes of a study, the characteristics of the continuers do as well.

In what follows, we describe procedures for assessing the degree of selectivity and thereby the degree of representativeness and generalizability in longitudinal studies. Some of these procedures are exemplified using a fictitious example based on data from the Action Control and Child Development Project (T. D. Little, Oettingen, & Baltes, 1995), and more detailed examples can be found on the resource web page

of this volume (http://www.mpib-berlin.mpg.de/research_resources/index.html). However, before we present our overview of the types of selectivity questions and review methodologies that can be applied to examine them, we first define concepts related to selectivity.

On Examining Participation Selectivity

In any study, values on variables are assigned to persons or groups of persons. Depending on the topic of study and the type of variable, these measurements are typically summarized as frequencies (prevalence rates), means, variances, and correlations. These mathematical summaries of the distributions of scores are used to characterize a sample and to address substantively meaningful questions. For example, (a) what proportion of adolescents smoke or drink alcohol (prevalence), (b) how large is the social network of rejected children versus popular children (mean), (c) how large are individual differences in intellectual abilities among 70- to 80-year-olds (variance), and (d) how closely linked are agency beliefs and school performance (correlation)?

Without further information, the validity of such statements is limited to those participants who were actually measured on the relevant variables. In other words, a study's conclusions are initially restricted to only those individuals who actually complete the measurement process. Because not all persons selected for a longitudinal study share the same participation profile and because each participation profile may have selective influences that are related to it, the question arises whether examining the full original sample (i.e., the parent sample) would have yielded different results—are the results based on continuers also true for the parent sample and, by implication, the population to which one wishes to generalize?

For instance, adolescents from disadvantaged households or with motivational difficulties may be less likely to continue in a longitudinal study of agency and school achievement. Conversely, adolescents who are quite conscientious or have high achievement motivation may be more likely to continue in a study. In scenarios such as these, statements about the links between agency beliefs and performance may, for example, misrepresent the true relations in the original sample and the population. Generally speaking then, validity is limited to the extent that observed (as well as non-observed) characteristics predicting participation or nonparticipation are correlated with variables of interest. Biased results would emerge, of necessity, if variables predicting those who drop out versus those who continue are systematically related to the variables under scrutiny (Graham & Hofer, chap. 11, this volume; Little & Rubin, 1987).

Quite too commonly, selectivity is addressed only in terms of mean-level differences between dropouts and continuers. Such an approach, as we emphasize in our discussion, is deficient because a greater wealth of information must also be examined in order to determine the full effects of selectivity. For example, even though questions about frequency or mean-level differences are often posed, information

about the variances and covariances is often neglected and, in addition, estimates of the effects of such differences on later relations among variables are nearly non-existent.

In our view, at least two types of selectivity-related questions can be examined (these questions are described in more detail next). The first question looks at systematic differences between participants and nonparticipants on variables that are measured in common to the subgroups. The second type of question looks at the influence that any systematic differences may have had on the core outcomes of a longitudinal study. Although both questions address similar aspects of selectivity, examining each type of question provides a full and informative picture of the degree of selectivity and allows any biasing effects in a longitudinal sample to be considered.

Selectivity Analyses: Goals and a Paradox

Selectivity analyses of longitudinal data represent a methodological precaution to reduce the likelihood of false conclusions and misleading generalizations (Lindenberger et al., 1999). Assessing the degree of selectivity helps to address basic validity questions, such as, are educational levels overestimated because fewer people of lower education agree to participate in a study than do those with higher education (or underestimated for some similar reason), or is the variance of intelligence underestimated because both good performers and low performers are more likely to drop out than individuals of average intelligence (or overestimated for some similar reason)?

Unfortunately, in trying to answer such questions, selectivity analyses are confronted with a fundamental paradox: In order to optimally document the degree and nature of selectivity, precisely the information that is missing must be known. The "Catch-22" then, is that to truly know the characteristics of the nonparticipants, they would have had to have been participants. On the other hand, even with only limited information on dropouts, selectivity analyses can still yield useful information regarding the representativeness and generalizability of the obtained results.

A minimum precondition of selectivity analyses, therefore, is that at least some basic pieces of information are gathered on all persons, including early dropouts (Dalenius, 1988; von Eye, 1989; Herzog & Rodgers, 1988; Oh & Scheuren, 1983; Panel on Incomplete Data, 1983). At the most sparse level, external sources of information such as basic census data can be used. At the other extreme, longitudinal studies that utilize large batteries at each wave have a wealth of information available regarding the possible selectivity of dropouts versus continuers.

In relation to this paradox, however, we must emphasize that selectivity analyses are limited in that they can only show the extent of observed selectivity and not the extent of selectivity that is possible in principle. Selectivity analyses try to relate available data to each other and to make optimal use of the inherent information, but obviously they cannot deal with issues for which additional information is necessary. This necessary deficit is especially relevant for nonparticipants for whom only a few variables can be used to document any observable selectivity (e.g., early dropouts).

Whether more information would have identified greater or more selectivity effects, however, cannot be answered conclusively.

Selectivity Subgroupings

We have described selectivity subgroupings, thus far, in the general terms of participants and nonparticipants. As mentioned in our general introduction, one common form of attrition is to end participation at some point after a study has begun (i.e., dropouts). Although the concept of a dropout is well understood, one type of dropout is often overlooked. Namely, those who dropout at the beginning of a study, before the first measurement is taken. Even with a truly representative sample of possible participants, dropouts at this very early stage of a study, for various reasons (unwilling, unable, unreachable), reflect possible sources of selective attrition. Early dropouts may lead to selectivity because, for example, they may be more likely to be married, come from single-parent families, be of higher socioeconomic status, and so on. Clearly, such differential characteristics can lead to fundamental bias in the results of an investigation (Baltes et al., 1988). In fact, even if a study is not longitudinal, the effect of early dropouts can lead to selectivity in the sample. In this regard, many of the selectivity analyses that we describe are also relevant to nonlongitudinal research designs.

In addition to the basic groupings (continuers and dropouts), other groupings can be introduced. For example, *returners* would be individuals who participated in earlier assessments, dropped out, but then returned for later assessments. In many cases, alternative methodologies for treating missing data can be utilized to condition the data set such that the information on the returners is reflected in the data (see Graham & Hofer, chap. 11, this volume; Wothke, chap. 12, this volume). In terms of our general discussion of selectivity, such groups can also be examined for the degree to which they influence the patterns of results inherent in the data. That is, the procedures we review are generalizable to other types of selective influence.

Finally, a related feature of some longitudinal studies is to include *new participants* at later time points. As with returners, new participants can also be used to examine for selectivity effects. For example, the new participants can be compared to both the continuers and the dropouts to determine the robustness of possible selectivity effects, and effects of early dropouts in both the original sample and the newly added sample can be compared for similarities and differences, and so on. Although such design features can complicate the picture, the generality of the methods that we describe is sufficient to accommodate these various forms of comparisons in order to examine the degree of representativeness and generalizability of a study's results.

In any longitudinal investigation, the general classifications between dropouts and continuers will be relevant for examining the degree of bias that is inherent in the sample under scrutiny. However, for studies stretching over a longer period of time, we augment these dichotomies with the graded concept of *participation levels* or participation depth (Lindenberger et al., 1999). For selectivity analyses, this distinction

has the important advantage that, at each wave (i.e., from one participation level to the next), the continuers can be compared with early, midway, and late dropouts on all previously measured variables. In this way, one can identify (a) variables that initially distinguish the two groups (early dropouts vs. continuers) and (b) variables that progressively differ or progressively converge in characterizing the nature of the continuing subsample in relation to the different levels of participation.

Another advantage of a levels approach to selectivity analyses is that estimates for constructs that are assessed at later time points can be calculated, given certain assumptions, such that they take into account the observed selectivity at previous time points. In other words, the later estimates can be adjusted given the degree of selectivity implicated by the early subsample of dropouts. These adjusted estimates can then be compared to the observed (i.e., unadjusted) estimates to index the degree of selectivity that is present in the data and the degree to which statements or conclusions would then need to be cautioned.

THE TWO TYPES OF SELECTIVITY QUESTIONS

To clarify the relations between the types of selectivity questions and methods that can be used to examine them, we employ a formalized representation of a longitudinal data structure (Lindenberg et al., 1999). With V_{it} , we refer to the variables (i.e., variable vectors) related to different participation levels or times of measurement (e.g., V_{t1} = Time 1 variables, V_{t2} = Time 2 variables, and so on). For example, for persons who participate in all times of measurement, we have observations (i.e., value assignments on variables) for the groups of variables measured at Time 1, V_{t1} , Time 2, V_{t2} , and so on, up to the final time of measurement, V_{tn} . For persons who drop out after the first measurement occasion, for example, only observations on the variables V_{t1} are available.

For each time of measurement, we define y_{it} as an indicator of participation versus nonparticipation. For example, $y_{t1} = 1$ if a person participated at level 1 (i.e., Time 1). If a person was not assessed at level i , y_{it} is set to zero. We refer to y_{it} as the participation indicator for a given time of measurement. Each participation indicator can then be used to determine a participation profile. A participation profile, p , is defined as the sum of each y_{it} where y_{it} is multiplied by 10 for each participation level higher than y_{it} . For example, if a study has three times of measurement, $p = (y_{t1} \times 10 \times 10) + (y_{t2} \times 10) + y_{t3}$. A participation profile provides a unique index of each observed pattern of participation in the longitudinal study. With three times of measurement, for example, seven unique profiles are possible (e.g., 111, 110, 011, 101, 100, 010, and 001).

The statistical methods that can be used to assess selectivity questions are based on the assumption that V_{it} and p reflect samples from a "superpopulation" that is characterized by a probability distribution (Cassel, Särndal, & Wretman, 1977). The concept of a superpopulation is useful because it allows aspects of the subsamples to be

expressed as parameters and allows use of standard statistical procedures to estimate them. In other words, given this assumption of a superpopulation, various statistical methods can be applied to address the two different types of questions that can be posed to clarify sample selectivity.

Selectivity Question Type I: Multivariate Differences Between Participation Subgroups

The first type of selectivity question focuses on the multivariate differences between the participation subgroups. From this viewpoint, the classifications of dropouts, returners, continuers, and so on would be used as an independent variable to compare differences in the multivariate relations among the measured variables and constructs that are common to the subgroups. This type of selectivity question focuses on whether a given participation profile has a different set of relations among the variables than another profile. For example, do continuers differ from dropouts in terms of the means, variances, or covariances on the previously observed variables?

In other words, for each observed participation profile (p):

$$E(V_{it} | p_j, y_{it} = 1) = E(V_{it} | p_k, y_{it} = 1), \text{ where } j \neq k, \quad (1)$$

and

$$\text{cov}(V_{it} | p_j, y_{it} = 1) = \text{cov}(V_{it} | p_k, y_{it} = 1), \text{ where } j \neq k. \quad (2)$$

Here, if continuers differ from dropouts, some aspects of the means, variances, and/or covariances among the variables at Time i (V_{it}) for the continuers are not the same as for the dropouts, for example. Under such conditions, some degree of selectivity is present that would compromise the integrity of any generalizations to the original population from which the sample was originally drawn.

In our view, the optimal method of analysis in this case would be a multiple-group comparison of the mean and covariance structures (MACS; T. D. Little, 1997; McArdle & Hamagami, 1991) among the variables and constructs, with a grouping defined for each participation profile, p , and where $y_{it} = 1$. In comparison to other procedures such as ANOVA, MANOVA, or a Box-M test, a MACS approach allows one to examine all of the primary moments (i.e., means, variances, and covariances) and to conduct significance tests for single matrix elements as well as groups of elements. This flexibility allows one to pinpoint those elements that reflect selectivity effects and those that do not. In other words, MACS analyses are ideally suited to examine selectivity effects on the means, variances, and covariances or correlations of continuous and normally distributed variables. We note, however, that the techniques are not suited, for example, to determine logistic probabilities for frequency distributions (see Lindenberger et al., 1999) and may be problematic when the sample sizes of specific participation profiles become too small.

Selectivity Question Type II: Projected Effects of Sample Attrition on Later Measured Constructs

The second type of selectivity question focuses on the projected effects of differential participation profiles when two or more different levels of participation are examined. For example, had the dropouts been included in the completing sample, would the relations among the constructs be different? With this type of question, information from previous measurements that include dropouts, returners, and/or new participants is used to provide projected estimates of the relations among the variables. As mentioned, these projections, or adjusted estimates, can be compared with the observed relations to index the degree of selectivity effects. Stated more generally, what are the effects of the different participation profiles on the basic moments (i.e., means, variances, and covariances) of the variables at later points in time?

In other words, across each observed participation profile (p):

$$E(V_{tn} | p_j, y_{ti} = 1) = E(V_{tn} | V_{ti}, p_k, y_{ti} \neq 0), \text{ where } j \neq k, \quad (3)$$

and

$$\text{cov}(V_{tn} | p_j, y_{ti} = 1) = \text{cov}(V_{tn} | V_{ti}, p_k, y_{ti} \neq 0), \text{ where } j \neq k. \quad (4)$$

For this type of question, at least three approaches can be taken. A first approach is to use the selectivity formula developed by Pearson (1903), Aitkin (1934), and Lawley (1943). This general formula provides estimates of the means, variances, and covariances that take into account (adjust for) observed selectivity (Meredith, 1964, 1993; Muthén, Kaplan, & Hollis, 1987; Smith, Holt, & Smith, 1989). From this vantage point, variables assessed or estimated for previous points in time are referred to as *selection variables*, and these variables are distinguished from the *dependent variables* (e.g., variables on which only continuers have scores). Means, variances, and covariances of the continuing sample are estimated on the basis of (a) the linear relationships between the selection variables and the dependent variables and (b) the differences in means between the lower level samples and the continuing sample.

The Pearson-Aitkin-Lawley method uses information on the relationships between the means, variances, and covariances assuming that the regressions of the dependent variables on the selection variables are linear and that the conditional variances are constant (homoscedasticity; although Meredith, 1993, suggests that this assumption is unnecessary). Under these assumptions (which cannot be tested empirically), the method allows a direct estimation of selectivity effects on the persons who complete a study. Within the framework of the linear model, this projection makes optimal use of nearly all available information (Meredith, 1964).

Aitkin (1934) and Lawley (1943) also showed that their formula can be applied repeatedly. For selectivity analyses, the additivity of the successive use of the basic formulae means that the variables available at Time 1 can be used to estimate, or project, the original sample's effects on the Time 2 variables. Then, the variables at Time 1 and 2 (i.e., the observed values at Time 1 and the estimated values at Time 2) can

serve as selection variables and the variables available at Time 3 become the dependent variables, and so on.

When interpreting the results obtained with this method, the following caveat, related to the basic paradox of selectivity analysis mentioned earlier, needs to be kept in mind: The more closely variables at lower participation levels (i.e., the selection variables) are associated with variables on the following level (i.e., the independent variables), the more meaningful the calculated estimates become (Lindenberger et al., 1999). Dependent variables that do not have variables that predict them to any degree at *previous measurement occasions* possess very little information that can be used to correct them and, by necessity, the dependent variables will maintain the values that were actually observed in the subsample. Another potential disadvantage of using the Pearson-Aitkin-Lawley formula is that standard errors of the reference values are not automatic outcomes of the procedure. Alternative approaches to estimating the standard errors, such as by bootstrapping the samples or systematically imputing the parameter estimates, would need to be applied.

A second approach to answer this second type of selectivity question is to use full information maximum likelihood estimation procedures to model the relations among all pieces of information that are present across the various participation profiles (see McArdle & Bell, chap. 5, this volume; Wothke, chap. 12, this volume). One advantage of this technique is that the maximum likelihood estimator provides information that is quite useful to examine the significance of selectivity; namely, standard errors of estimates. Another advantage is that the selectivity influences of all subgroupings are *simultaneously estimated*. With the Pearson-Aitkin-Lawley approach, one needs to employ the formula a number of times for each relevant comparison.

A third approach to answer this second type of question is to impute the missing information using the full information that is available in the data set. Although imputation techniques are powerful, guidelines on the limits of how much of a data set can be imputed reliably have yet to be established (see Graham & Hofer, chap. 11, this volume).

Summary

The two selectivity questions refer to (a) the differences between participation subgroups on the means, variances, and covariances of observed variables and constructs that the subgroups share in common; and (b) the relationships between variables at previous time points and variables assessed at later time points, which are used to project, or adjust, the estimates of the multivariate relations in the completing subsample.

If differences between the various subgroups (e.g., dropouts, continuers) on the analyzed variables become apparent using the various procedures, it indicates that characteristics of the participants and nonparticipants have influenced the obtained results and some selectivity is present. The extent of selectivity can be represented by the reference values that emerge from the analyses. Clearly, such information is useful

to determine both the degree of selectivity and, by implication, the degree of generalizability. However, we must emphasize, again, that the reverse does not necessarily hold. Finding no differences does not indicate that sample selectivity has not occurred. For instance, some relevant variables that predict sample attrition may not have been assessed in the first place and therefore could not be analyzed. This problem is particularly important for analyses of sample loss at the earliest stages of participation (e.g., early dropouts vs. continuers). At this level, only rudimentary demographic variables might be all that is available. In addition, the procedures that we outline here do not capture all forms of possible sample selectivity. They are limited to frequency distributions, means, variances, and covariances and the linear relations among them. However, given that these basic statistics are the most common and robust summaries of distributions, they are still quite relevant for most longitudinal investigations.

BRIEF EMPIRICAL EXAMPLE

Details of the Samples and the Data

To demonstrate some of these techniques, we created two subsamples from a group of 425 boys and girls (approximately equally distributed across grades 2–6) with complete data at two times of measurement (see T. D. Little, Oettingen, & Baltes, 1995; T. D. Little, Oettingen, Stetsenko, & Baltes, 1995; Oettingen, Little, Lindenberg, & Baltes, 1994 for details of the sample). To mimic a selective process, we randomly (but with a systematic bias) assigned participants to be dropouts or continuers. The systematic bias was introduced by using the intellectual skill scores (i.e., the Raven progressive matrices) of the participants. Specifically, we selected more participants with low Raven scores to be dropouts (approximately a 2:1 ratio). The dropout group ($n = 122$) thereby represents a subgroup that is selective with regard to Raven intelligence. In order to simplify our example, we selected three constructs to examine for selectivity: school performance, academic achievement, and personal agency.

For the school performance measure, we used the teacher-assigned grades in math and language. For academic achievement, we used the scores from the math and language subscale of the Begabungstestsysteem (BTS), a German-language achievement test. For the personal agency measure, we used the agency beliefs for effort, ability, and luck from the Control, Agency, and Means-ends Interview (CAMI; T. D. Little, Oettingen, & Baltes, 1995). Agency beliefs, which are similar to self-efficacy beliefs (Bandura, 1997), reflect a child's personal perception of whether he or she can utilize such means as effort, ability, and luck to obtain good school grades.

Selectivity Question Type I

To exemplify this question we conducted a two-group MACS analysis comparing the continuers and dropouts on the constructs represented at the first time of measurement. We used the MACS framework detailed by T. D. Little (1997).

Assessing Measurement Equivalence

When fit as a combined multiple-group model with no cross-group equality constraints, the basic model showed acceptable fit, $\chi^2_{(22)} = 28.73$, $NNFI = .992$, $IFI = .9996$, $RMSEA = .0402$, indicating that the general structure is tenable. To test for measurement equivalence, we first evaluated the loadings and then, because this model was tenable, subsequently added constraints to the intercepts. Specifically, when invariance of the loadings and intercepts was enforced, the overall model fit was again quite acceptable, $\chi^2_{(28)} = 37.28$, $NNFI = .991$, $IFI = .994$, $RMSEA = .0406$.

Taken as a whole, all the fit indices showed quite minimal differences in the sequence of steps between the freely estimated model and the measurement-equivalent model. Therefore, on the basis of the minimal differences in fit between Model 1 and Model 2 (i.e., a modeling rationale; see T. D. Little, 1997), these results indicate that the constructs have equivalent measurement properties across the two groups (continuers vs. dropouts). If measurement equivalence were not tenable, the nature of the selectivity effects would be quite pronounced because it would have affected the underlying factorial composition among the indicators, yielding noncomparable constructs. As a result, selectivity would be qualitative in nature and the quantitative degree of selectivity could not be estimated. In other words, such a situation would be quite problematic and reflect a serious threat to the validity of any conclusions. On the other hand, with comparably measured constructs, the nature of selectivity can be quantified and evaluated in terms of the degree of possible bias.

Testing for Selectivity Differences

Because construct comparability was tenable, we tested the three sets of basic moments for differences across the participation subsamples; specifically, we examined (a) equality of the latent means, (b) equality of the latent standard deviations, and (c) equality of the latent correlations. Table 10.1 contains the nested-model comparisons used to test for possible differences on these parameters.

As seen in Table 10.1, two of the tests of equivalence on the latent parameters were significant, indicating some degree of selectivity across the different participation profiles. For the mean-level tests, all three constructs showed significant differences whereby the dropouts had lower school performance, lower academic achievement, and lower personal agency. However, the selective process did not influence the variances of the three constructs (see tests of the standard deviations in Table 10.1). Finally, the correlational structure among the constructs also showed some evidence of

TABLE 10.1
Multivariate Comparison of Continuers Versus Dropouts

<i>Tested parameter</i>	<i>Difference test</i>			
	$\chi^2_{(31)}$	$\Delta\chi^2$	Δdf	<i>p</i>
<i>Means</i>	43.8	8.0	3	<.05
<i>Standard deviations</i>	36.5	0.7	3	>.80
<i>Correlations</i>	42.4	6.6	3	<.10

Note. Comparison $\chi^2_{(28)} = 35.8$; Δ = a difference between the comparison model and the tested models; χ^2 = the maximum likelihood chi-squared statistic; *df* = degrees of freedom; *p* = the probability level.

selectivity, although the effects were not as pronounced as the mean-level effects. Follow-up analyses revealed that the correlational manifold among the three constructs was higher in the continuing sample than in the dropout sample. Specifically, for continuers versus dropouts, respectively, the correlation between personal agency and school performance was .71 versus .57; between personal agency and academic achievement, .34 versus .03; and finally, between school performance and academic achievement, the correlation was .54 versus .38.

Given these differences, one would be compelled to conclude that the continuing sample was no longer representative of the original sample and that various sources of selective bias have influenced various aspects of and relations among the measured variables. One problem, however, is that analyses such as these still do not indicate the degree of bias that the selectivity effects have on the parameter estimates at later measurement occasions. To address questions such as this, one needs to utilize an analysis that explicitly examines the influence of selectivity on the later time points. We now turn to one such approach.

Selectivity Question Type II

Table 10.2 shows the results of our analyses addressing the second type of selectivity question. Here we used the full information maximum likelihood approach wherein the complete longitudinal model is fit to all 425 participants, but the 122 dropouts have no values on the variables at the second measurement occasion. In the first column of Table 10.2, as a point of comparison, we present the results of the analyses as performed only on the subsample of participants who were assessed at each measurement occasion. These values would be the information from which one would typically draw conclusions.

In comparison to these values, the second column in Table 10.2 presents the analysis that include the subsample of dropouts. Here, the information on the dropouts at Time 1 is explicitly represented in the analyses such that the parameter estimates at Time 2 include the influences of the dropout subsample.

TABLE 10.2
Selectivity Bias in Estimates of Time 2 Parameters

<i>Focal parameter at occasion 2</i>	<i>Continuers</i>	<i>Continuers with dropouts</i>	<i>Population</i>
Variance AGENCY	.942	.923	.984
Variance PERFORMANCE	1.037	1.041	1.038
Variance ACHIEVEMENT	.859	.862	.897
Corr (AGENCY, PERFORMANCE)	.732	.726	.713
Corr (AGENCY, ACHIEVEMENT)	.512	.488	.444
Corr (ACHIEVEMENT, ACHIEVEMENT)	.637	.618	.613
Mean AGENCY	.024	.035	-.019
Mean PERFORMANCE	-.039	-.033	-.039
Mean ACHIEVEMENT	.543	.577	.581

Note. Corr = correlation; AGENCY = agency beliefs for effort, ability, and luck; PERFORMANCE = teacher-assigned school marks; ACHIEVEMENT = scores on the BTS (see text). *Continuers* are the estimates based on the sample of 303 continuers without any selectivity corrections. *Continuers with dropouts* are the estimates based on the derived sample of 303 continuers with 122 dropouts estimated simultaneously using the Full-Information Maximum Likelihood approach. *Population* are the true estimates based on the full population of 425 participants.

Finally, in the third column we present the true population values for these participants (recall that the dropout classification was artificial and that the data set we used started as a complete data set for all participants). Here we see that the values are also quite different from those for the continuers but they are generally more similar to the values in the second column, which reflect the adjustments based on the information about the dropouts at the first measurement occasion only.

The difference between the estimates in column two and column three reflect the fact that the selection variable (Raven intelligence) is not perfectly correlated with the variables of interest. Therefore, the adjustments that result from the full information approach are only part adjustments and still contain some evidence of bias. However, the nature of the adjustments is clearly more accurate than the estimates based on the continuers alone. For example, consistent with the higher correlational manifold among the constructs at the first occasion, the constructs show a higher positive correlational manifold at the second occasion for the continuers than for the adjusted estimates which, in turn, were closer to the estimates found in the population.

We wish to emphasize that the significance of these selectivity effects in our example is small. Our example was meant to be a conservative one in that our selection variable was only moderately correlated with the measures we examined and, because of the graded selection process, the strength of the association between being a dropout and variables we examined was quite low ($r = 0.2$). Therefore, the degree of selectivity on the constructs at the second occasion that we identified was relatively small. Future Monte-Carlo work can be done to systematically vary the association between

the selection variable and the primary constructs under scrutiny to determine thresholds and cutoffs for the nature and degree of resulting bias.

CONCLUSION

Our primary goal has been to re-emphasize the importance of examining selectivity in longitudinal research. We highlighted two fundamental types of questions that can be addressed when examining the impact of selectivity. Clearly further work is needed to examine which methods are most appropriate to assess the degree of bias that selectivity introduces. In addition, further work is needed to explore the boundaries of generalizability given various degrees of bias. However, it is also clear that the analytic machinery that is available to detect, estimate, and correct selectivity bias has made tremendous advances. Our parting admonition is for researchers to utilize these techniques, when possible, to fully explore the degree of selectivity bias that may be evident in their longitudinal data sets. Relying on less sophisticated traditional approaches (seemingly in the hopes that their simplicity will yield null results) runs tremendous risks. Because selectivity is an inevitability, researchers must be mindful that it is a matter of degree. A small degree of bias will likely not invalidate broader conclusions. A large degree of bias would necessarily temper the breadth of one's conclusions. Either way, acknowledging the existence of selectivity bias and exploring as fully as possible its influences are needed to render one's generalizations as veridical as the data, the design, and the analytic techniques allow.

ACKNOWLEDGMENT

This work was supported in part by the Berlin-Brandenburg Academy of Sciences, the Max Planck Society, the Max Planck Institute for Human Development in Berlin, the Max Planck Institute for Demography in Rostock, and Yale University. We are grateful to the many members of the Berlin Aging Study, the Institute, and to various visiting scholars for discussions on the issues presented in this work. We are particularly thankful to Paul Baltes and John Nesselrode for their invaluable advice and support.