Review of Development Economics, 8(1), 91–109, 2004

Using Quality of Interview Information to Assess Nonrandom Attrition Bias in Developing-Country Panel Data

*John A. Maluccio**

Abstract

Panel data often provide an understanding of household behavior not possible with cross-sectional information alone. However, a disturbing feature of such data is that there can be substantial, nonrandom attrition and many analysts share the concern that this inhibits the ability to make accurate inferences. The author examines attrition in the KwaZulu–Natal Income Dynamics Study 1993–1998, assesses the extent of attrition bias for a specific empirical example, and proposes and implements a selection correction methodology using quality of first round interview variables as identifying instruments. The results show that attrition does lead to statistical bias in the "behavioral" coefficients in estimation of household-level expenditure functions. Since it is typically difficult to determine the bias for a particular analysis *a priori*, and such bias is by its nature model-specific, it behooves researchers using panel data to evaluate the effects of attrition in their analyses.

1. Introduction

The analysis of panel (or longitudinal) data, where the *same* individuals or households are interviewed multiple times, contributes substantially to our understanding of a variety of economic phenomena. For example, while repeated cross-sectional surveys of different households at two points in time might reveal a constant poverty rate, they are silent as to whether this reflects chronic poverty; i.e., the same households in poverty in each period, or transitory poverty with equal proportions of households exiting and entering poverty between surveys. If appropriate policy prescriptions depend on the chronic or transitory nature of poverty, it is essential to be able to distinguish between the two. Thus panel data often permit an understanding of the dynamic behavior of individuals or households not possible with cross-sectional information, even repeated cross-sections.

A second valuable feature of panel data is that they enable one to econometrically resolve a pernicious form of omitted variable bias: that which is due to time-invariant unobserved heterogeneity. For example, rarely do surveys observe or measure a family's preferences and priorities for educating its children. It is quite likely that families that put a high priority on education perform additional work to obtain income needed to pay school fees. If we use cross-sectional data alone to determine the effect of family income on education, we risk making incorrect inferences if these omitted (time-invariant) preferences or tastes for education are correlated with included

*International Food Policy Research Institute, 2033 K Street NW, Washington, DC 20006-1002, USA. Tel: 202-862-5693; Fax: 202-467-4439; E-mail: j.maluccio@cgiar.org. I would like to thank Michael Carter, Duncan Thomas, Jere Behrman, Paul Cichello, Lawrence Haddad, Julian May, Aki Stavrou, and an anonymous referee for valuable comments. The 1998 KwaZulu–Natal Income Dynamics Study reported on here was a collaborative project between IFPRI, the University of Natal, and the University of Wisconsin–Madison. In addition, the assistance of the Southern Africa Labour and Development Research Unit is gratefully acknowledged. The data are available at www.ifpri.org.

income measures. Estimates derived from such data will tend to overstate the impact that an income transfer would have on educational decisions of families that give only an average priority to education. In contrast, with panel data, econometric methods can be used to control for these sorts of time-invariant preferences and family characteristics, allowing improved estimates of the effect of income on education.[1]

Panel data are not a panacea, however. When carried out using updated sample frames, cross-sections have a clear advantage in terms of representativeness and are therefore superior for certain types of analyses.[2] Moreover, in practice, one must balance the potentially substantial benefits against the many real difficulties encountered in survey work that lead to, in particular, errors of measurement and sample attrition. Either of these can introduce different sources of bias, inhibiting anew the capacity to make correct inferences from the data.

Panel data estimators that control for time-invariant unobserved heterogeneity, for example, fixed-effects estimators, are more susceptible to bias from measurement error than ordinary least squares. Indeed, when there are relatively large (random) measurement errors it may even be preferable to eschew these types of estimators and not control for the unobserved time-invariant omitted variables at all (Hsiao, 1986). While exact results depend on the data and the form that the measurement error takes, one can appeal to the signal-to-noise ratio for the intuition of the underlying problem. Since fixed-effects estimators rely on variation over time for identification, when there is little such variation, for a given level of noise fixed-effects estimators may actually lower the signal-to-noise ratio relative to the alternative of ordinary least squares; this can result in the inconsistency of the fixed-effects estimator being worse than that of the ordinary least squares alternative that does not exploit the panel nature of the data. Of course, in settings where there is rapid change, as in South Africa, fixed-effects estimators may increase the signal-to-noise ratio. In practice, it is not usually possible to ascertain the extent or nature of measurement error, although panel data often provide partial means to do so, for example, through repeated measurements of some variables.

Similarly, unless it is random, sample attrition (i.e., when some targeted households are not successfully interviewed in all rounds) may introduce biases into analyses based only on the nonattriting sample. Of course, the potential problem of selection bias due to nonresponse exists in cross-sectional surveys as well (although it is typically ignored), but it is typically exacerbated in panel data owing to the difficulties inherent in interviewing the same individuals or households multiple times. To put greater confidence in findings based on panel data, one must assess the magnitude of these potential problems.

This analysis focuses on the latter problem of sample attrition and the possible ensuing selectivity, with special reference to a recently collected panel survey of African and Indian households living in KwaZulu–Natal Province, South Africa, the KwaZulu–Natal Income Dynamics Study (KIDS) (May et al., 2000). The study examines attrition in the KIDS in detail in order to (1) document the procedures and outcomes of the survey as a resource, both for those using this publicly available data and for those embarking on their own survey work; (2) describe the characteristics of households that attrited in the second round of the survey and explore their correlates in a multivariate framework; and (3) propose a simple methodology to assess, and correct for, attrition bias using information reflecting the quality of the fieldwork in the first round as identifying instruments. Because the KIDS is a comprehensive survey and can be used for a variety of analyses, it will not be possible to make global statements about attrition bias. As Becketti et al. (1988) note, the hypothesis that attrition is correlated with some (possibly unobserved) variable of interest is quite broad.

Rather, the results presented here should be treated as methods to be replicated by other researchers analyzing the data these and other panel data.

2. Attrition in Panel Data

Among the earliest large-scale (e.g., 1,000+ household) panel surveys are those begun in the United States in the 1960s, including the Panel Study of Income Dynamics (PSID) and the National Longitudinal Survey of Labor Market Experience. While initially designed to study the nature and causes of poverty in the United States, these surveys have subsequently been used to examine a wide range of topics, including labor supply, earnings, family composition changes, and residential mobility (Baltagi, 1995). For the most part, large-scale panel studies began much later in developing countries,[3] facilitated by the Living Standards Measurement Survey (LSMS) projects of the World Bank in the late 1980s (Deaton, 1997; Glewwe and Jacoby, 2000).

As a result, both the level of attrition and analysis of its impact are more advanced for the US-based surveys. For example, in the PSID, over half the original 1968 sample had attrited by 1989 (Fitzgerald et al., 1998a). These high levels of attrition have spawned research that examines the reliability of conclusions drawn from these data, however. It turns out that the *level* of attrition alone need not necessarily distort inferences made using the data. On the whole, research in this area has not found large biases due to sample attrition for some commonly estimated labor market models. (A number of these studies were published in a special issue in the spring of 1998 of *The Journal of Human Resources* 33(2) entitled "Attrition in Longitudinal Surveys.") While encouraging to those working with panel data, these results do not necessarily extend to other models or settings and data where the processes underlying the attrition may differ.

There are a number of reasons why one might expect selective attrition to be more severe in developing than developed countries (Ashenfelter et al., 1986; Thomas et al., 2001). Better information and capability for tracking typically exist in developed countries; respondents are often just a phone call away. Furthermore, the high levels of mobility and long-distance migration associated with development are likely to complicate survey work in developing countries. Partly offsetting these concerns, however, are the much lower refusal rates typical of developing countries, perhaps reflecting lower opportunity costs of time and/or different cultural attitudes toward the interviewing process.

A recent survey, the Indonesian Family Life Survey (IFLS), demonstrates that with careful planning and execution it is possible to collect panel data in developing countries with similar or even lower levels of attrition than in typical developed country surveys (Thomas et al., 2001). In Indonesia, a key explanatory factor for reinterview rates above 94% after a three-year period was the decision to track movers, something not typically done in developing-country surveys. Tracking reduced attrition by more than half. Nevertheless, the attrition that remains is still nonrandom and is associated with migration, as well as with community and household characteristics. After controlling for community-level wealth, attrition in the IFLS is negatively associated with household size and positively associated with logarithmic per capita expenditures, but only for those below the 25th percentile of the per capita expenditure distribution within the community. Thus, households at the lower end of the distribution were more likely to move and not be successfully tracked.

While documenting the existence of nonrandom attrition as above is important, it is not the end of the story. What is of ultimate concern is whether, and to what extent, the attrition invalidates inferences made using the data. For example, would an analysis of the determinants of poverty using the IFLS data be influenced by the patterns of attrition described above? As a basis for exploring the possible effects in a general framework, consider a canonical, one-period, selection model as described in equations (1) and (2) (bold characters represent vectors):

$$y_i = \mathbf{x}_i'\boldsymbol{\beta}_1 + \varepsilon_i \quad (y_i \text{ observed only if } A_i^* < 0) \tag{1}$$

$$A_i^* = \mathbf{x}_i'\boldsymbol{\beta}_2 + \mathbf{z}_i'\boldsymbol{\gamma} + v_i. \tag{2}$$

Equation (1) is the model of interest. The outcome variable, y_i, is observed only for a subset of the entire sample, those for whom the latent index, A_i^*, is less than zero. Equation (2) represents a selection function depending on the same independent variables in (1) as well as additional ones. In practice, we do not typically observe A_i^* but only an indicator of whether an observation is selected or not, in this example $A_i = 0$ ($A_i^* < 0$) if selected (i.e., observed) and $A_i = 1$ ($A_i^* \geq 0$) if not (i.e., unobserved). If there is correlation between the error terms ε_i and υ_i, estimation of (1) ignoring (2) leads to inconsistent parameter estimates of $\boldsymbol{\beta}_1$; this is commonly referred to as selection bias.

If we now treat y_i as an outcome variable from the second period of a two-period panel dataset (but leave $\mathbf{x}_i$ as a vector of first-period measures, such as time-invariant factors) and recast (2) as an attrition function, we have the equivalent result for attrition in a panel survey. If there is correlation between the error terms ε_i and υ_i, estimation of (1) ignoring (2)—i.e., estimation on the nonattriting sample alone—leads to inconsistent parameter estimates of $\boldsymbol{\beta}_1$.

The stylized model presented here makes it clear that any evaluation of attrition bias is necessarily model-specific. As the outcome modeled in (1) changes from labor supply, to fertility, to child health, some or all of the right-hand-side explanatory variables may change and, in particular, ε_i necessarily changes, reintroducing the possibility that there is correlation between the error terms ε_i and υ_i in the two equations.

Building on the methodology of Becketti et al. (1988), Fitzgerald et al. (1998a) suggest a simple test for attrition bias in panel data using first-round information supplemented by knowledge of A_i, whether the household attrited at a later date. The procedure is to estimate (1) using the entire set of first-round information with $\mathbf{x}_i$ and a set of interactions between $\mathbf{x}_i$ and the attrition indicator A_i, as independent variables. The aim is to determine whether those who subsequently leave the sample differ in their initial behavioral relationships. Significant interaction terms are a warning sign for attrition bias.

If attrition bias is present, one solution, estimation of a selection-corrected model, lies with $\mathbf{z}_i$ provided it is correlated with attrition but not correlated with ε_i in the model of interest; i.e., provided it is validly excluded from (1). First, the selection (in this case attrition) function (2) is estimated including all exogenous variables $\mathbf{x}_i$ and identifying instruments $\mathbf{z}_i$, and then a selection correction factor is introduced into the second-stage estimation of (1) (Heckman, 1979).

Fitzgerald et al. (1998a,b) suggest an alternative solution for a slightly different form of attrition selection. They first distinguish between two cases: (1) selection on unobservables, the model discussed above, and (2) selection on observables, where $\mathbf{z}_i$ and ε_i are correlated but ε_i and υ_i are not. A convenient interpretation for the second formulation is that $\mathbf{z}_i$ are outcome variables measured in round 1 that might be considered endogenous in (1), perhaps even including (lagged) y_i itself. Their solution involves estimating the attrition function using (endogenous) $\mathbf{z}_i$ and then estimating model (1) by weighted least squares where the weights are constructed from the first-stage attrition estimates (Fitzgerald et al., 1998b).

Since the South African data analyzed here provide quality of 1993 interview variables that are plausibly exogenous to a variety of outcomes at the household and community levels, this analysis adopts a Heckman selection approach to correct for attrition bias. A substantial advantage to this approach compared with the weighted least-squares methodology mentioned above is that it is robust to attrition selected on both observables *and* unobservables. A drawback to the approach, however, is that although second-round dependent variables from the nonattriting sample may be modeled, without further assumptions they can be modeled using only explanatory variables from the first round, including, for example, time-invariant variables.[4]

3. 1993 Baseline Survey

The first South African national household survey, the Project for Statistics on Living Standards and Development (PSLSD), was undertaken in the last half of 1993 by a consortium of South African survey groups and universities under the leadership of the South African Labour and Development Research Unit (SALDRU) at the University of Cape Town (PSLSD, 1994).[5] This analysis uses a subset of these data comprising Africans and Indians living in KwaZulu–Natal Province and described further below. Similar to a living standards measurement survey (Grosh and Muñoz, 1996; Deaton, 1997; Grosh and Glewwe, 2000), the main instrument was a comprehensive household survey that collected a broad array of information on the socioeconomic condition of households. Among other things, it includes sections on household demographics, household environment, education, food and nonfood expenditures, remittances, employment and income, agricultural activities, health, and anthropometry (weights and heights of children). In addition to the household questionnaire, a community questionnaire was administered in each cluster or community (hereafter community) in the sample to gather information common to households in an area such as school availability, healthcare facilities, and prices for various commodities.

The sample was selected using a two-stage, self-weighting design. In the first stage, communities were chosen with probability proportional to size from census enumerator subdistricts (ESDs) or approximate equivalents where ESDs were not available. This step was also designed, via stratification, to provide representativeness at the province and homeland area levels as they were demarcated in 1993. In the second stage, a census of all inhabited physical dwellings in each chosen community was completed, the dwellings were numbered, and a list of those to be interviewed was randomly generated. In addition, a second list of "replacements" or "substitutes" was randomly generated from the remaining dwellings; when it was not possible to interview a pre-designated, first-choice random sample dwelling, the team was instructed to interview a dwelling from the replacement list.[6] Interview households were then determined from the people who lived in the selected dwellings, making it possible for more than one household, as defined in the survey, to reside in a single dwelling (PSLSD, 1994).[7]

Before turning to the 1998 second-round survey, I first document the level of nonresponse in the 1993 round for the 1998 target sample areas. While it is true that dwellings, and therefore households, on the replacement list were randomly selected, it seems unlikely that the process by which first-choice households were dropped was random. For example, more replacements in a community might be reflective of less careful fieldwork if the interview teams rushed to complete their work thereby making mistakes in the census and/or not carefully searching for selected respondents. On the other hand, there may have been households that refused to be interviewed or that migrated (temporarily or permanently) between the time of the census and the survey, although this was often only a few days and in nearly all cases less than a month. Since little or no information was collected on the first-choice households not interviewed, I follow Thomas et al. (2001) and examine the characteristics of the survey communities and their association with the frequency of replacement interviews in order to explore possible effects on the final sample interviewed.[8]

Overall, 90.0% of the pre-designated first-choice random sample of 1,354 households was interviewed in the original 1993 fieldwork, but this average first-round completion rate conceals substantial variation among communities. One-third of the communities indicated that all first-choice households were interviewed, but in the remaining areas, 1993 average completion rates vary, dipping below 70% in five rural communities in the former KwaZulu homeland area, where only Africans resided. On average, rural areas had slightly lower completion rates than urban areas (88% versus 92%, respectively); within urban areas Indian communities had slightly lower completion rates than African communities (91% versus 94%, respectively).

First-round completion rates are largely unrelated to an array of observed characteristics of the community and sample households within it, with only a few exceptions. Treating the fraction completed in each community as the outcome, the main finding from ordinary least squares regressions (robust standard errors are used to calculate the *t*-statistics in parentheses) using each of the 67 communities as an observation, is that communities with lower completion rates appear to have been either growing or experiencing relatively high mobility (Table 1).[9]

For example, specification (3) in the third column of Table 1 indicates that in communities where there was either net in- or out-migration in the 12 months prior to the 1993 survey, completion rates were 10 percentage points lower. Completion rates are also negatively (and very strongly) associated with the fraction of survey households in the community who reported in-migrating to the area any time in the past five years. These results suggest that the first-round sample may be somewhat selective toward households that were less likely to move.

In addition to considering the role of migration indicators themselves, local labor market conditions plausibly influence migration and therefore attrition. Given apartheid policies of segregation, key indicators for local labor market conditions are the area where one lived, especially rural versus urban, and whether or not one resided in the former Natal Province or former KwaZulu homeland. Owing to the general lack of spatial integration of the population, communities in the sample are entirely African or entirely Indian so race can be treated as a community-level variable in 1993 that also captures an important dimension of location. Lastly, the influence of community average daily wages is evaluated. While several of these indicators will play a role in the 1993 to 1998 attrition analysis presented in the next section, it turns out that none of them is significantly associated with 1993 completion rates.

Thus, in addition to being a possible proxy measure for survey quality as described above, the completion rate also appears to be associated with mobility within communities. For both these reasons it might be a useful predictor of attrition in later rounds, although the interpretation for why this is so would differ depending on which is the dominant influence, survey quality or migration. I return to this distinction below in the analysis of attrition between 1993 and 1998.

Table 1. Baseline Nonresponse in the 1993 KwaZulu–Natal Sample

Dependent variable: percentage of first-choice random sample households in 1993 community sample (n = 67)	(1)	(2)	(3)
Community characteristics (logs)			
(1) if in former Natal Province	2.825	13.34	–7.068
	(0.8)	(1.3)	(0.7)
(1) if urban	4.760	5.547	3.991
	(1.6)	(1.0)	(0.8)
(1) if in former Natal Province and urban	–4.496	–10.94	4.579
	(0.6)	(1.0)	(0.4)
(1) if Indian	–1.603	–0.620	5.990
	(0.3)	(0.1)	(0.7)
(1) if secondary school within community	–0.441	—	–3.107
	(0.1)		(0.8)
(1) if health clinic within community	–4.855	—	–5.700
	(1.4)		(1.5)
(1) if net in- or out-migration in past year	–8.095***	—	–9.847***
	(2.8)		(2.8)
Average characteristics of households in community			
Fraction in-migrating in past 5 years	—	–44.08	–57.40**
		(1.4)	(2.1)
Log average reported wage	—	1.890	–5.191
		(0.4)	(1.0)
Log per capita expenditures	—	3.413	–0.490
		(0.4)	(0.1)
Log household size	—	12.10	6.576
		(1.3)	(0.8)
Fraction with male household head	—	0.924	14.87
		(0.1)	(1.1)
Education of household head	—	–0.640	0.846
		(0.4)	(0.5)
Age of household head	—	–4.053	–2.463
		(1.6)	(1.1)
Age of household head squared	—	0.042*	0.031
		(1.7)	(1.4)
Fraction owning house	—	–13.86	–20.58*
		(1.1)	(1.7)
Constant	96.15***	155.4**	160.1**
	(39.4)	(2.1)	(2.4)
R^2	0.20	0.21	0.38
F-test all covariates	2.4**	1.6	2.7***
p-value	0.03	0.12	0.01

Notes: Ordinary least squares estimates. Absolute value of t-statistics in parentheses calculated using robust standard errors (StataCorp, 2001). *indicates significance at 10%, **at 5%, ***at 1%.

4. The 1998 Resurvey

South Africa has undergone dramatic political, social, and economic change since the first democratic national elections in 1994. With the aim of addressing policy research questions concerning how these changes are affecting South Africans, African and Indian households surveyed in the 1993 PSLSD in KwaZulu–Natal Province were resurveyed from March to June 1998 for the KIDS. The choice of KwaZulu–Natal was in part the result of practical considerations, including the feasibility of locating households that had been interviewed in 1993 since the original survey was not intended to be a panel. The data, as well as a sampling of the policy questions it can be used to inform, are described by May et al. (2000) in more detail.[10]

One of the administrative changes made after the 1994 elections by the South African government was the designation of new provinces and provincial boundaries. The former KwaZulu homeland area and Natal Province were combined to create KwaZulu–Natal Province. Unlike some of the other new South African provinces, however, the pre-1994 borders of Natal (which circumscribed the KwaZulu homeland area) were not altered; thus the 1993 sample, which was designed to be representative at the 1993 provincial level, remains approximately representative at the newly formed provincial level, another reason for its selection.[11]

Attrition in 1998

In theory, three factors underlie the level of attrition in a panel survey: (1) the mobility of the target population, (2) the success with which those who move are followed and reinterviewed, and (3) the number of refusals. Thus, attrition is often closely linked to migration behavior. In practice, there is also the possibility of errors in fieldwork (in both earlier and later rounds).

The 1993 (and thus 1998 target) portion of the PSLSD sample included 1,354 households (Table 2).[12] Interview teams first went to the 1993 location of a household. If it was learned that the household had moved, the team was instructed to get new loca-

Table 2. Attrition in the 1998 KwaZulu–Natal Sample

Reinterviewed?	*African rural*	*African urban*	*Indian urban*	*Row totals*
Yes, same location	667	256	149	1,072
	(80.8)	(81.5)	(69.3)	(79.2)
Yes, different location	21	20	19	60
	(2.5)	(6.4)	(8.8)	(4.4)
No, known to have moved	60	15	18	93
	(7.3)	(4.8)	(8.4)	(6.9)
No, no-trace	70	19	25	114
	(8.5)	(6.1)	(11.6)	(8.4)
Refusal	4	3	4	11
	(0.5)	(0.9)	(1.9)	(0.8)
Death	3	1	0	4
	(0.4)	(0.3)	(0.0)	(0.3)
Column totals	825	314	215	1,354
	(60.9)	(23.2)	(15.9)	(100.0)

Notes: Row and column percentages are shown in parentheses.

tion information using a form provided for this purpose. The teams sought address or other contact information from other family members, neighbors, and local facilities, such as clinics and schools. If a new address was found, and was sufficiently detailed, the household was tracked to its new location.

Of the target sample, 1,132 households (83.6%) were successfully reinterviewed, success being defined as having reinterviewed at least one adult member from the 1993 household.[13] Sixty households were tracked to new locations and successfully reinterviewed using the tracking protocol described above. Surprisingly, many surveys in developing countries do not attempt to track movers; had this strategy been followed, only 79.2% of the target households would have been reinterviewed. Put another way, the tracking procedures yielded a one-quarter reduction in the level of attrition between the 1993 and 1998 surveys.

In most surveys of this type in developing countries, refusal rates are low (Deaton, 1997; Thomas et al., 2001). This is also true for the KIDS: only 11 recontacted households refused an interview. Finally, in four households (three single-person and one two-person), all of the 1993 members had died before the 1998 resurvey.

For Africans, reinterview rates were higher in urban areas versus rural areas (87.9% versus 83.3%). Within urban areas, reinterview rates for Africans were higher still (92.5%) in the metropolitan areas, which are characterized by more permanent housing structures and street addresses (not shown). Indian households proved more difficult to reinterview (78.1%). This appears to have been related to higher mobility among Indians, as well as problems with the 1993 fieldwork; in two Indian communities, for example, the household addresses were completely mixed up, making tracking substantially more difficult.

An alternative way to categorize the households is according to whether they lived in the former KwaZulu homeland area or in the former Natal Province (not shown). Residential restrictions were stricter and property rights more limited in the former Natal Province. Also, two of the communities surveyed in rural Natal consisted of black farm workers on large, commercial white-owned farms where there appears to have been high turnover, including one farm that went bankrupt, dispersing nearly all of its residents. Finally, there was a spate of expulsions from large farms in some Natal areas, apparently a strategy by white farmers aimed at avoiding the possible consequences of anticipated land reform. Consequently, among Africans, reinterview rates were much higher in rural areas of the former KwaZulu homeland compared with rural areas of former Natal Province (86.7% versus 62.5%).

Characteristics of Attriting Households

Before turning to the multivariate analysis of attrition between the rounds, I first consider a simple comparison of some 1993 household characteristics for those who were ultimately reinterviewed versus those who were not. On average, attritors were significantly more likely to be Indian than African and have higher per capita income and expenditures, more educated household heads, and more durable assets. Of course, since these measures tend to be highly correlated—in particular, race with education, income, and assets—it is not surprising that they show similar patterns. Nevertheless, the comparisons suggest that attrition in the sample is nonrandom.

For more than one-third of the households not reinterviewed, information collected verified that the household had moved but did not provide enough detail to allow tracking to a new residence. For the remaining households, however, there was simply no trace; i.e., no one approached in the community recognized the name of any household members when presented with the 1993 household roster (see Table 2). These two groups, those who are known to have moved and those who seemingly left no trace, turn out to be different. In addition to analyzing simple attrition, then, it is informative to separately consider these two groups of attritors.

Table 3 presents two types of multivariate estimates of attrition. The first column contains the results from a binary logit attrition function where the dependent variable has two possible outcomes—reinterviewed or attrited. In the remaining columns, I consider the different types of attritors separately. To compare the characteristics of (1) those reinterviewed with (2) those not reinterviewed but known to have moved and (3) those not reinterviewed but leaving no trace, I employ multinomial logits distinguishing among the three mutually exclusive categories.

Both models include explanatory factors that parallel the baseline completion rate analysis presented in Table 1, but with two important differences. First, community average characteristics of households used above are replaced with their household-level counterparts, except for those factors most closely associated with previous migration decisions (whether the household had migrated in the past five years and reported wages). Since 60 households that had moved were tracked and reinterviewed, attrition and migration behavior are not the same thing; however, they are still partly linked. Recent household decisions regarding migration and employment are likely to be closely related to other unobservable factors in the household that directly influence migration decisions and therefore directly influence attrition. As a result, their inclusion could bias estimates of the role of all the explanatory factors. Community averages for these variables provide a means of avoiding these potential biases. The second difference is that the 1993 to 1998 attrition equations include two proxy measures of 1993 survey quality, one of which is the community average completion rate modeled in the previous section.

Table 3 presents the derivatives of the probability with respect to each regressor ($\partial P/\partial X$) for the different models, evaluated at the overall mean of the regressors for each independent variable and multiplied by 100. (Robust standard errors allowing for correlation within communities are used to calculate the asymptotic t-statistics in parentheses.) For example, in the first row, first column, –5.445, if significant, would indicate that households living in rural areas of former Natal Province (note the interaction term in Table 3 two lines below) were 5% less likely to attrit. The remaining columns present two specifications for the multinomial logit. In the first row, second column (2a), 8.325 indicates that households in former rural Natal were 8% more likely to fall into the category not reinterviewed but known to have moved (hereafter "mover") relative to households that were successfully reinterviewed, the omitted category. In contrast, –5.110 in the first row, third column (2b), if significant, would indicate that households in former rural Natal were 5% less likely to be in the category not reinterviewed and leaving no trace (hereafter "no trace") relative to successfully reinterviewed households.

Returning to the first column containing the results from the binomial attrition function, few community characteristics appear to be strong predictors of attrition over the period. Households in communities with a health clinic were 9% less likely to attrit, probably reflecting the fact that clinics provided an important source of information for the interview teams when tracking. The fraction of households reporting that they had migrated to the area in the past five years is positively associated with attrition, suggesting that more mobility in an area is associated with more attrition. Labor market indicators (location and average wages), however, have no significant influence

Table 3. Binomial and Multinomial Logit Attrition Regressions for the 1998 KwaZulu–Natal Sample

Omitted category: reinterviewed in 1998 (n = 1,354)	(1) *Attritors*	(2a) *Mover*	(2b) *No-trace*	(3a) *Mover*	(3b) *No-trace*
Community characteristics (logs)					
(1) if in former Natal Province	–5.445	8.325*	–5.110	1.179	–3.964
	(0.5)	(1.9)	(0.8)	(0.2)	(0.5)
(1) if urban	–1.431	1.514	–4.480	2.093	–3.299
	(0.2)	(0.6)	(0.9)	(1.0)	(0.7)
(1) if former Natal Province and urban	2.758	–9.864*	10.42	–4.809	6.100
	(0.2)	(1.9)	(1.1)	(1.0)	(0.6)
(1) if Indian	9.952	8.886*	–0.924	10.06*	0.230
	(1.0)	(1.8)	(0.1)	(1.9)	(0.0)
(1) if secondary school within community	3.167	3.953	–2.111	5.001**	–0.719
	(0.6)	(1.6)	(0.6)	(2.2)	(0.2)
(1) if health clinic within community	–8.699*	–1.190	–5.762*	–1.324	–6.693**
	(1.9)	(0.6)	(1.9)	(0.6)	(2.1)
(1) if net in- or out-migration in past year	–6.849	—	—	–3.163	–3.457
	(1.3)			(1.4)	(0.9)
Mean characteristics of households in community					
Fraction in-migrating in past 5 years	6.877*	—	—	4.960**	1.644
	(1.9)			(2.3)	(0.8)
Log average reported wage	–2.366	—	—	–5.454**	2.785
	(0.5)			(2.5)	(0.7)
Log average per capita expenditures	–8.772	–7.311**	–0.198	–6.125*	–2.181
	(1.5)	(2.5)	(0.1)	(1.9)	(0.6)
Household characteristics					
Log per capita expenditures	–0.826	2.345**	–2.876**	2.276**	–2.673*
	(0.4)	(2.2)	(1.9)	(2.2)	(1.8)
Log household size	–9.011***	–2.301*	–5.890***	–2.073	–5.835***
	(3.9)	(1.9)	(3.9)	(1.7)	(3.9)
Log if male household head	2.693	2.518	–0.460	2.870*	–0.093
	(1.2)	(1.4)	(0.3)	(1.7)	(0.1)
Education of household head	0.247	–0.037	0.343**	0.001	0.224
	(0.9)	(0.2)	(2.1)	(0.0)	(1.5)
Age of household head	0.137	–0.055	0.102	0.035	0.179
	(0.4)	(0.2)	(0.3)	(0.1)	(0.5)
Age of household head^2 × 10,000	–0.129	0.086	–0.157	0.027	–0.234
	(0.4)	(0.4)	(0.5)	(0.1)	(0.7)
Log if own house	–2.372	0.314	–2.262	0.727	–2.556
	(0.6)	(0.2)	(0.9)	(0.4)	(1.1)
Quality of interview variables					
Average 1993 completion rate	–26.91*	–1.985	–17.56**	–1.482	–21.84**
	(1.8)	(0.2)	(2.1)	(0.2)	(2.4)
(1) if questionnaire verified	9.436	–1.792	8.239**	–1.383	8.187**
	(1.5)	(0.8)	(2.4)	(0.6)	(2.3)
Constant	74.63*	13.10	29.32	22.21	33.60
	(1.9)	(0.7)	(1.2)	(1.1)	(1.3)
χ^2 test quality variables	3.4	0.9	7.1**	0.9	8.7***
p-value	0.18	0.63	0.03	0.65	0.01
Pseudo R^2	0.09	0.09		0.11	
χ^2 test all covariates	57.2***	139.4***		299.0***	
p-value	0.01	0.01		0.01	
χ^2 test movers and no-trace the same		39.0***		43.8***	
p-value		0.01		0.01	

Notes: Multinomial logit estimates, derivatives ($\partial P/\partial X$), at the overall mean of the regressors for each independent variable are shown ×100 except where indicated. Absolute value of asymptotic t-statistics in parentheses calculated using robust standard errors, allowing for within cluster correlation (StataCorp, 2001). *indicates significance at 10%, ** at 5%, and *** at 1%. Estimation treats 11 refusals and 4 deaths as attritors in (1) and as no-trace in (2) and (3).

either individually or as a group. Finally, the KIDS survey was more likely to reinterview larger households, a result I interpret below.

Identifying and reinterviewing households in a panel survey relies heavily on the accuracy of the original fieldwork. Measures of quality for the original interview, then, may help predict the success of reinterview (Zabel, 1998). Two such measures of quality of the 1993 interview are considered next. The first is an indicator of whether the questionnaire was verified (signed) by the team supervisor. Verification was indicated as having been done for all Indian households but only 23% of African ones. The maintained hypothesis is that properly verified questionnaires were unlikely to have been bogus interviews and more likely to have been accurately and fully completed (correct names, address, etc.), making recontact more likely.

The second measure recasts the 1993 average completion rate for first-choice, random sample households modeled in the previous section as an independent variable and proxy measure for survey quality in communities. Despite the possibility that it may also in part reflect mobility (addressed below), I include this measure since there is no variation in the verification indicator for the Indian subsample but there does appear to have been variation in the quality of fieldwork for that group as well.

The verification indicator is positive but insignificant in the binomial attrition function, but the 1993 community average completion rate is negatively associated with attrition and has large effects in the hypothesized direction—higher completion rates decrease the likelihood of attrition.

While this result is consistent with their interpretation as quality variables, the earlier findings regarding the completion rate and its association with migration activity cannot be disregarded. Furthermore, since the verification indicator does not vary for the Indian subsample—for whom there were higher rates of migration—it is possible that the quality indicators are measures of *both* migration activity and quality of the earlier fieldwork.[14] I provide two additional types of evidence in order to examine further the quality of interview interpretation for these measures. The first is to separately distinguish the attritors known to have moved and those who left no trace; the quality variables affect only the latter group, indicating they are not associated with migration only. Second, I include other available migration and labor market information; these additional factors are significant but do not alter the effect of the quality variables.

The results presented in columns (2a) and (2b) of Table 3 first explore the role of the quality variables in specifications that do not include all the migration and labor market factors. Reflecting the weaker property rights described earlier, households residing in former rural Natal were 8% more likely to have moved relative to being reinterviewed; this effect is entirely offset if the area was in former urban Natal, however. Indians, as described above and seen in Table 2, were more likely to have moved. Households in communities with clinics were less likely to be in the no-trace group, supporting the interpretation made above that clinics provided an important source of information for the interview teams when tracking. Households living in wealthier communities, as measured by the logarithm of community average per capita expenditures, appear less likely to have moved, thereby facilitating reinterview. Community-level wealth is not associated with the probability of being in the no-trace category, however.

Two household-level characteristics are also strongly associated with attrition. Conditional on the logarithm of community average per-capita expenditures, individual households with more resources measured in the same fashion were more likely to be movers relative to those reinterviewed, but less likely to be in the no-trace group. In comparison to the Indonesian case, where households at the lower end of the distribution were more likely to move and not be successfully tracked, in KwaZulu–Natal it was wealthier households that were more likely to move and not be successfully tracked while poorer households were more likely not to leave a trace. Wealthier households seemingly had higher profiles in the community, making it easier for the survey team to learn their whereabouts whereas the converse holds for poorer households.

Unsurprisingly, the KIDS survey was more likely to reinterview larger households, especially relative to the no-trace group, though additional members increased the likelihood less and less. Various linear splines for household size were considered and verify that the relationship is nonlinear with diminishing probability for additional members (not shown). This suggests that larger households were less likely to move, consistent with associated higher moving costs, and, among those that did move, the survey teams were more likely to find some trace for larger households who presumably had more contacts within the community.

Both quality indicators are significant for the no-trace group only; higher 1993 completion rates substantially decrease the probability of being in the no-trace group while the verification indicator—conditional on all the other covariates—has a significantly positive effect on attrition. Since neither measure is related to movers, the quality of interview interpretation appears to dominate the migration interpretation.

The final specification presented in columns (3a) and (3b) of Table 3 includes the migration and community average-wage indicators as further controls to explore their effect on the quality variables and ensure the latter are not merely picking up dimensions of migration and its effects on attrition. Previous in-migration activity in the community over the past five years is positively associated with movers relative to being reinterviewed but has no association with the no-trace group. The logarithm of community average reported wages is also associated with movers with higher wages leading to fewer households in the mover group, but unrelated to the no-trace group. In contrast, however, the quality measures retain their significant effects on being in the no-trace group but have no effect individually, or jointly, on movers.[15]

In summary, the evidence presented here indicates the following. A large percentage (83.6%) of the original sample was successfully reinterviewed after nearly five years, and the decision to follow those who had moved contributed a substantial portion to the overall success rate. However, attrition in the KIDS survey is nonrandom and varies with, among other things, household size and community and household-level resources. Furthermore, despite following movers, attrition remains closely linked to migration. The characteristics of the households that were not reinterviewed but known to have moved differ from those who left no trace, suggesting that the processes underlying their attrition were also different. Indicators of quality of the interview in 1993 were identified that significantly influence the probability of being in the no-trace group, but not in the mover group, and therefore can be used to correct for attrition bias due to sample selection on unobservables.

5. Attrition Selection-Corrected Expenditure Functions

While observable differences between attitors and nonattritors (as well as within the former group) indicate attrition is nonrandom, this does not necessarily imply that estimated relationships based only on the nonattriting sample suffer from attrition bias, but only that they might. Indeed, attrition bias could still be a problem even if there were no *observable* differences between the two groups; it depends on the existence of correlation between the error terms ε_i and υ_i in equations (1) and (2) shown above. For example, if attrition is selective on observable right-hand-side covariates, and the model is well specified, it may be possible to condition on those variables allowing consistent estimation of (1) while ignoring (2). This is not an option, however, if there is selection on unobservables. In that case, a possible solution is a standard selection-correction methodology (Heckman, 1979; Maddala, 1986). This is the strategy employed below using the verification indicator and the average 1993 completion rate as identifying instruments.

One of the chief aims of the KIDS was to assess the changes in income and expenditure of households five years after South Africa's first national elections. As such, the data are being used in efforts to understand the underlying mechanisms for households that were able to escape poverty (Carter and May, 2001). A simple framework to begin analyzing this is the household-level expenditure function. Below I present estimates of such a function to assess whether and how attrition influences estimates based only on the nonattriting sample. To some extent, this is a loaded example since we learned above that attrition was associated with per capita expenditures; nonetheless, it is illustrates the methodology to test for, and correct, attrition bias.

Following Fitzgerald et al. (1998a), I first test for attrition bias estimating an expenditure function using all the 1993 first-round data and a complete set of interactions with an attrition dummy variable for those households that attrited in 1998. (Estimates are based on ordinary least squares with robust standard errors allowing for correlation within communities used to calculate the *t*-statistics in parentheses.) The results presented in column 1 of Table 4 are consistent with standard findings in the literature, and over 50% of the variation in logarithmic per capita expenditures is explained. Per capita expenditures are higher for urban households, Indian households, smaller households, households with male heads, and households with more educated heads. Households that later attrite are different in terms of the 1993 relationship between the logarithm of per capita expenditures and this standard set of conditioning variables; the *F*-test on all the attrition interaction terms (excluding the attrition indicator dummy) is 2.6, which is significant at the 2% level (Table 4, column 1). Attrition bias appears to be affecting the estimates.

Next, I consider two formulations of a "permanent" expenditure function estimating 1998 logarithmic per capita expenditures using initial (1993) values of explanatory variables. This formulation is necessary in order to have the complete set of observations for estimation of the selection corrected estimates that follow. Column 2 in Table 4 presents this specification of the expenditure function estimated using ordinary least squares. As before, the results are consistent with those typical in the literature so I do not describe them.

The second specification for the 1998 expenditure function is a maximum-likelihood, selection-corrected version using the 1993 average completion rate and the verification indicator as identifying variables for the first-stage probit predicting selection into the sample; i.e., nonattrition (note this formulation is in contrast to Table 3 where attrition was being predicted). Column 3 in Table 4 shows the results from the selection (nonattrition) probit. The quality-of-interview variables significantly predict attrition above and beyond the other conditioning variables in the expenditure function with a joint χ^2 test statistic of 5.3 (significant at the 7% level).

Column 4 in Table 4 presents the maximum-likelihood results from a Heckman selection-corrected model of 1998 expenditures (Heckman, 1979; Maddala, 1986). The selection term is positive and significant, indicating there was positive selection into the sample of nonattritors; that is, nonattritors had on average higher per capita expen-

Table 4. Reduced-form Household Expenditure Functions and Selection Correction

Dependent variable: Log per capita household expenditures	(1) *1993 Log per capita expenditure*	(2) *1998 Log per capita expenditure*	(3) *Nonattrition probit*	(4) *1998 Log per capita expenditure*
1993 community characteristics (logs)				
(1) if former Natal Province	–0.587***	–0.227***	–0.303*	–0.305***
	(3.9)	(2.7)	(1.8)	(3.0)
(1) if urban	0.260**	0.368***	0.282**	0.407***
	(2.3)	(4.4)	(2.4)	(4.5)
(1) if Indian	0.998***	1.081***	0.137	1.102***
	(7.1)	(10.0)	(0.8)	(9.2)
1993 household characteristics				
Household size	–0.095***	–0.062***	0.080***	–0.053***
	(10.1)	(9.8)	(5.1)	(7.3)
(1) if male household head	0.113**	0.067*	–0.070	0.060
	(2.5)	(1.8)	(0.7)	(1.5)
Education of household head	0.054***	0.072***	0.002	0.072***
	(6.5)	(9.0)	(0.1)	(8.4)
Age of household head	0.010	0.005	0.003	0.006
	(1.2)	(0.6)	(0.2)	(0.8)
Age of household head squared ×1000	–0.047	0.038	–0.036	0.030
	(0.6)	(0.5)	(0.0)	(0.4)
Attrition indicator	0.440	—	—	
	(0.8)			
Attrition indicator interactions	Yes	No	No	No
Inverse Mills ratio (lambda)	—	—	—	0.563***
				(3.1)
Average 1993 completion rate	—	—	0.444	—
			(1.2)	
(1) if questionnaire verified	—	—	–0.298**	—
			(2.0)	
Constant	5.237***	5.095***	0.128	4.905***
	(18.2)	(23.3)	(0.2)	(20.1)
F-test (attrition interactions)	2.6**	—	—	—
p-value	0.02			
Hausman test (column 2 versus column 4) *p*-value				37.1***
				0.01
R^2 (Pseudo R^2 in column 3)	0.58	0.55	0.07	—
F-test (χ^2 in columns 3 and 4)	55.1***	124.1***	83.3***	833.3***
all covariates *p*-value *F* or χ^2 test	0.01	0.01	0.01	0.01
N	1,354	1,132	1,354	1,132

Notes: Absolute value of (asymptotic in columns 3 and 4) *t*-statistics in parentheses calculated using robust standard errors allowing for within cluster correlation (StataCorp, 2001). *indicates significance at 10%, **at 5%, and ***at 1%.

ditures conditional on the included characteristics. A Hausman test rejects the hypothesis that the slope coefficients in the uncorrected (column 2) and corrected (column 4) expenditure functions are equal. For this example, then, I conclude that the "behavioral" coefficients in the expenditure functions are indeed biased by attrition in the sample, although this bias is not confined to any single coefficient but rather spread among several of them. For example, the uncorrected version underestimated the deleterious effect of residing in areas of the former Natal Province by about 25% and underestimated the positive effect of living in an urban area by about 10%. The substantive importance of differences such as these would of course depend on the underlying purpose of the analysis.

Finally, it should be noted that these findings do not imply the decision to follow movers in the KIDS survey had no effect on attrition bias; in fact it reduced the bias, further supporting the value of this practice. Results (not shown) from redoing the analysis treating all those who had moved but were reinterviewed as *not* having been reinterviewed (i.e., as if movers had not been followed) lead to greater attrition bias in the estimated relationships.

6. Conclusions

Panel data often provide an understanding of the dynamic behavior of individual households not possible with cross-sectional or time-series information alone. However, a disturbing feature of such surveys in both developed and developing countries is that there is often substantial, nonrandom attrition. In developing countries, attrition is closely linked to migration, which is the result of household-level decisions and is likely to be selective on both observable and unobservable household characteristics. For these reasons, many analysts share the concern that attrition inhibits our ability to make accurate inferences using panel data. This paper has examined attrition in the KwaZulu–Natal Income Dynamics Study (1993–1998) and assessed the extent of attrition bias in the context of a specific empirical example.

Multivariate regressions are used to describe the characteristics of households reinterviewed and households attriting in 1998, distinguishing among two types of attriting households, those that are known to have moved and those that left no trace. In addition to finding a number of observable differences among the different groups, I find quality of first-round interview variables predict attrition well, particularly for the no-trace group.

While observable differences between attritors and nonattritors (as well as within the former group) indicate that attrition is nonrandom, this does not necessarily imply that estimated relationships based on the nonattriting sample suffer from attrition bias. To more directly explore attrition bias, which is by its nature model-specific, I estimate household-level expenditure functions correcting for attrition bias using standard Heckman selection procedures and the quality of 1993 interview variables as identifying instruments. The results show that, at least for this simple case, attrition does lead to statistical bias in the "behavioral" coefficients. This bias, however, is not confined to a single coefficient.

These findings are important both for researchers preparing panel surveys and for those using them. Those involved in survey work would do well to emphasize the quality of enumerator training, for example, and endeavor to collect information on the quality of the interviews and reasons for attrition, and include them in public releases of the data. Analysts can then use the quality data in selection-correction models as illustrated here.

In related research, Alderman et al. (2001) use the Fitzgerald et al. (1998a) techniques to explore attrition bias in three developing-country panel datasets, including the KIDS data examined in this paper. For a majority of the outcome variables they consider across the different datasets, estimated models show little sign of attrition bias. In particular, for the KIDS sample, estimates of a variety of child anthropometric outcomes indicate attrition bias in only a few of them.

The Alderman et al. (2001) results, in conjunction with those presented in this paper, reinforce the notion that attrition bias for models estimated on panel data is indeed model-specific. Large levels of attrition do not always lead to attrition bias; however, sometimes they do. Since it is typically difficult to determine the bias for a particular analysis *a priori*, it behooves researchers using panel data not to avoid using panel data when there is attrition, but to evaluate the effect of such bias on the analysis at hand.

References

Alderman, Harold, Jere Behrman, Hans-Peter Kohler, John A. Maluccio, and Susan Cotts Watkins, "Attrition in Longitudinal Household Survey Data: Some Tests for Three Developing Country Samples," *Demographic Research* 5(4) (2001):77–124.

Ashenfelter, Orley, Angus Deaton, and Gary Solon, "Collecting Panel Data in Developing Countries: Does It Make Sense?" Living Standards Measurement Study working paper 23, Washington, DC: World Bank (1986).

Baltagi, Badi, *Econometric Analysis of Panel Data*, New York: John Wiley (1995).

Becketti, Sean, William Gould, Lee Lillard, and Finis Welch, "The Panel Study of Income Dynamics After Fourteen Years: an Evaluation," *Journal of Labor Economics* 6 (1988):472–92.

Carter, Michael and Julian May, "One Kind of Freedom: Poverty Dynamics in Post-Apartheid South Africa," *World Development* 29(12) (2001):1987–2006.

Deaton, Angus, *The Analysis of Household Surveys*, Baltimore, MD: Johns Hopkins University Press (1997).

Fitzgerald, John, Peter Gottschalk, and Robert Moffit, "An Analysis of Sample Attrition in Panel Data," *Journal of Human Resources* 33 (1998a):251–99.

———, "The Impact of Attrition in the PSID on Intergenerational Analysis," *Journal of Human Resources* 33 (1998b):300–44.

Glewwe, Paul and Hanan Jacoby, "Recommendations for Collecting Panel Data," in Margaret Grosh and Paul Glewwe (eds.), *Designing Household Survey Questionnaires for Developing Countries: Lessons from 15 years of the Living Standards Measurement Study*, Washington, DC: World Bank (2000).

Grosh, Margaret and Paul Glewwe (eds.), *Designing Household Survey Questionnaires for Developing Countries: Lessons from 15 years of the Living Standards Measurement Study*, Washington, DC: World Bank (2000).

Grosh, Margaret and Juan Muñoz, "A Manual for Planning and Implementing the Living Standards Measurement Study Survey," Living Standards Measurement Study working paper 126. Washington, DC: World Bank (1996).

Heckman, James, "Sample Selection Bias as a Specification Error," *Econometrica* 47 (1979):153–61.

Hsiao, Cheng, *Analysis of Panel Data*, New York: Cambridge University Press (1986).

Maddala, G. S., *Limited Dependent and Qualitative Variables in Econometrics*, New York: Cambridge University Press (1986).

May, Julian, Michael Carter, Lawrence Haddad, and John Maluccio, "KwaZulu–Natal Income Dynamics Study (KIDS) 1993–1998: a Longitudinal Household Database for South African Policy Analysis," *Development Southern Africa* 17 (2000):567–81.

PSLSD (Project for Statistics on Living Standards and Development), "South Africans Rich and Poor: Baseline Household Statistics," South African Labour and Development Research Unit, University of Cape Town, Cape Town (1994).

StataCorp, *Stata Statistical Software: Release 7.0*, College Station, TX: Stata Corporation (2001).

Thomas, Duncan, Elizabeth Frankenberg, and James P. Smith, "Lost But Not Forgotten: Attrition in the Indonesian Family Life Survey," *Journal of Human Resources*, 36(3) (2001):556–92.

Zabel, Jeffrey, "An Analysis of Attrition in the PSID and the Survey of Income and Program Participation," *Journal of Human Resources* 33 (1998):479–506.

Notes

1. Hsiao (1986) notes two other econometric advantages of panel data: (1) increased number of observations or degrees of freedom, and (2) increased variability of regressors since they are measured over both space and time. Ashenfelter et al. (1986) demonstrate that panel data are superior for estimating differences in means over time (but not necessarily the means themselves) in the presence of serially correlated measurement error.

2. The current period representativeness of a panel sample deteriorates over time, and this may occur more quickly in rapidly changing societies. Thus, many analyses appropriate for a representative cross-sectional survey, such as estimation of population means, are not appropriate for individual rounds of a panel survey after the first one.

3. One exception is the Indian National Council of Applied Economic Research Additional Rural Incomes Survey initiated in 1968 on a national sample of over 4,000 households followed for three years and then reinterviewed again in the early 1980s (Glewwe and Jacoby, 2000). Various institutions, including the Institute for Crop Research in the Semi-Arid Tropics and the International Food Policy Research Institute, conducted a number of smaller scale surveys beginning in the 1970s.

4. A third possible solution to the selection problem is to use fixed-effects estimators as described in section 1. To the extent that the correlation between the error components in equations (1) and (2) is fixed over time for households this would be a valid methodology. It does not, however, allow direct evaluation and testing for attrition bias as the methods carried out here do.

5. The PSLSD has also been referred to as the SALDRU survey, the South African Integrated Household Survey, and the South African LSMS.

6. In addition to replacements at the household level, a small number of first-choice random sample clusters were also replaced because they were considered too dangerous to carry out fieldwork. These were replaced with nearby communities displaying otherwise similar characteristics (PSLSD, 1994).

7. An important component of the survey design, as with any household survey, was the definition of a household. To account for the complexity of the South African situation with its history of residential restrictions and migrant labor, a two-tiered definition for household members, resident or nonresident, was formulated based on time spent in residence. *Resident household members* were defined as those who (i) lived under this roof for more than 15 of the last 30 days; (ii) when they are together they share food from a common source (i.e., they cook and eat together); and (iii) contribute to or share in, a common resource pool (PSLSD, 1994, p. iv). The household was also defined to include *nonresident members*—those that satisfied conditions (ii) and (iii) and had lived under the same roof at least 15 days out of the past year. Only limited information was collected from nonresident household members.

8. The ensuing analysis treats 32 households for which the replacement indicator is missing as having been first-choice random sample draws. The results are qualitatively unchanged if these are treated as replacements instead.

9. The lack of association for other variables is not an artifact of the large number of controls. Only the presence of a secondary school and the migration indicators are significantly correlated with completion rates in unconditional bivariate comparisons. A variety of nonlinear specifications for expenditures were also considered.

10. During follow-up field research in May 2001 it was discovered that all 39 household interviews in clusters 217 and 218 had been fabricated in both 1993 and 1998; these households are

dropped in this analysis, leading to minor discrepancies between the figures concerning attrition reported here and those reported in previously published work, in particular, May et al. (2000). In light of this, a careful assessment of all suspicious clusters was carried out in 2002 and it was determined that 4 additional clusters were also dubious. The most recent release of the data (January 2003) excludes these as well—as a result those currently working with the data begin with a sample of 1,247 in 1993 (about 100 fewer than reported in this article). All the analyses reported in this article were repeated on the revised sample and the qualitative results for all regressions are the same and, if anything, the effect of the quality of interview variables is larger and more significant. Since the new release occurred after the editing process began, however, it was not possible to update the numbers presented in this article.

11. The extent to which it is currently representative is of course somewhat weakened by the passage of time and the loss of the two clusters mentioned in note 10.

12. These include 1,139 African and 215 Indian households but exclude 112 white and 53 colored households not targeted for reinterview in 1998.

13. Since household structure is not static over time, an alternative and conceptually cleaner approach to examining attrition is at the individual level. This was accounted for in the design of the 1998 resurvey with key decision-makers (essentially household heads) being specifically targeted for reinterview as described in May et al. (2000). This paper adopts the commonly used household level approach since researchers often treat the household as the unit of analysis.

14. The verification indicator is not a significant predictor when averaged to the cluster level and included as an explanatory variable for 1993 completion rates as in Table 1.

15. Another measure considered that had no discernable effect on the results was the community-level unemployment rate (based on the average for household respondents). By design, the explanatory variables in the attrition functions include only information available in 1993. However, I have also considered specifications using information on economic shocks in the communities between 1993 and 1998, including labor market shocks. Several variations of these are significant in predicting “movers” but none predict the “no-trace” group. The quality of interview variables is virtually the same when economic shocks are included.