The Impact of Decentralized Data Entry on the Quality of Household Survey Data in Developing Countries: Evidence from a Randomized Experiment in Vietnam

Paul Glewwe and Hai-Anh Hoang Dang

Computers were provided to randomly selected districts participating in a household survey in Vietnam to assess the impact on data quality of entering data within a day or two of completing the interview rather than several weeks later in the provincial capital. Provision of computers had no significant effect on the observed distribution of household expenditures and thus no effect on measured poverty. Provision of computers reduced the mean number of errors per household by 5–23 percent, depending on the type of error. Given the already low rate of errors in the survey, however, the goal of increasing the precision of the estimated mean of a typical variable can be achieved at a much lower cost by slightly increasing the sample size. Provision of additional computers did substantially reduce the time interviewers spent adding up and checking the data in the field, with the value of the time saved close to the cost of purchasing desktop computers. JEL Classification: C81, C93, C42, I32, O15

Public Disclosure Authorized

Household survey data are used for many policy and research purposes in developed and developing countries. Yet anyone who works with household survey data soon realizes that they can have many errors and inconsistencies. Statisticians call such errors nonsampling errors. Economists call them measurement errors.

Reducing such errors in household surveys should lead to better research and better policies. Personal computers are a useful tool for this purpose. Data from survey questionnaires can be entered into personal computers using

Paul Glewwe (corresponding author) is a professor in the Department of Applied Economics at the University of Minnesota; his email address is pglewwe@umn.edu. Hai-Anh Hoang Dang is a consultant in the Policy Research Group at the World Bank; his email address is hdang@worldbank.org. The authors would like to thank the staff of Vietnam's General Statistical Office for their cooperation in this research. The authors also benefited from comments and information provided by Graham Kalton, Frank Martin, Rob Swinkels, Phung Duc Tung, and three anonymous reviewers. Glewwe thanks INRA-LEA (Fédération Paris-Jourdan) for hospitality in the fall of 2004, when the first draft of this article was written. A longer version of this article is available at http://wber.oxfordjournals.org/.

THE WORLD BANK ECONOMIC REVIEW, VOL. 22, NO. 1, pp. 165–185 doi:10.1093/wber/lhm023
Advance Access Publication January 28, 2008

software that detects data errors (see, for example, Grosh and Muñoz 1996). Many detected errors can be resolved by returning to the households to gather more accurate information.

Using personal computers to reduce nonsampling errors in developing countries is most effective when data entry takes place near the household, soon after the interview, because survey teams often work only a week or two in one area before moving to another, at which point it difficult for them to return to the previous area to correct errors detected by data-entry programs. Providing personal computers to survey teams could avoid this problem,[1] but doing so can be very expensive for statistical offices in developing countries. Before purchasing more computers, these statistical offices need to know the impact of buying more computers on the errors in the survey data they collect.

This article analyzes a randomized experiment conducted in Vietnam in 2002. Computers for decentralized data entry were randomly provided to some districts and not others. Such experiments are rare in developing countries. The first set of results focuses on per capita expenditures, as the survey's main purpose is to use expenditure data to measure poverty and household welfare. The data show no evidence that providing computers to each district affected measured per capita expenditures.

A second set of results examines errors and inconsistencies in other variables. Decentralized data entry reduces errors by 5–23 percent, depending on the type of error. The use of desktop computers in high population density areas has the largest impact. The gains in data quality from using more personal computers are not cost-effective, however: raising the sample size would reduce measured variance at a much lower cost. Yet providing computers does reduce the time interviewers spend manually checking the data, and the value of the time saved is close to the cost of a desktop computer.

This article is organized as follows. Section I describes the design of the experiment. Section II explains how to test for the impact of decentralized data entry and defines probable errors in the data. Section III reports the results, and section IV discusses the implications for data collection in Vietnam. Section V summarizes the findings and offers recommendations for developing country statistical offices.

I. Design of the Experiment

In 2002 Vietnam's General Statistical Office (GSO) conducted the first Vietnam Household Living Standards Survey (VHLSS). Its main purpose is to monitor poverty and other indicators of household welfare in Vietnam (GSO 2006).

1. The distance could be reduced to zero by providing interviewers with laptop computers, to enable them to enter data during the interview. This is called computer-assisted personal interviewing (CAPI). The decentralized data entry examined here did not use this method.

The 2002 VHLSS collected information from about 75,000 households from all of Vietnam's 61 provinces. Of these households, 45,000 were administered a questionnaire that solicited information on income and basic household characteristics. The other 30,000 were interviewed using a questionnaire that solicited the same information plus data on household expenditures. This study focuses on the 30,000 households interviewed using the second questionnaire. The 45,000 households that completed the first questionnaire are excluded because they were interviewed in the first half of 2002, whereas the experiment was implemented in the last three months of 2002.

The sample for the 2002 VHLSS was drawn as follows. Vietnam's 61 provinces were divided into urban and rural areas, creating 122 strata. Vietnam is divided into about 10,000 communes. The GSO drew a master sample of 3,000 communes, with the number of communes per stratum determined by the square root of the number of households in those strata. Each selected commune was divided into several enumeration areas. One enumeration area was randomly selected from each of the 3,000 communes, with selection probabilities proportional to the number of households per area. From each selected enumeration area, 20 households were randomly selected, yielding a sample of 60,000 households. In early 2002 the sample was increased to 75,000 households by selecting five additional households from each enumeration area. The 30,000 households that completed the second questionnaire are a representative sample of all of Vietnam.

All VHLSS survey teams have two interviewers and one supervisor. Each district has a single team, which conducts all the interviews in its district.

This study estimates the impact on data quality of decentralized data entry using personal computers. More precisely, it examines the impact of providing data-entry computers at the district level, rather than the province level, for the 2002 VHLSS. Providing computers to each district requires 10 times as many computers.[2]

The VHLSS data are entered into computers at Province Statistical Offices in the provincial capital. District-based teams interview all sampled households in an enumeration area and check the data collected manually. This manual checking is very time consuming. Each district then sends the completed household questionnaires to the provincial capital. Data entry is done at Provincial Statistical Offices, using personal computers with a data-entry program that detects more than 100 potential errors and inconsistencies.

Under this system survey teams have few opportunities to return to households to clear up errors or inconsistencies detected by the software when data are entered at provincial capitals. In contrast, the decentralized data-entry system evaluated here allowed survey teams to return quickly to households to correct errors or inconsistencies detected by the software. Because data can be

2. In 2002 Vietnam had 607 districts and 61 provinces. Since then some provinces and districts have been split, increasing the numbers of both.

entered into computers in each district one or two days after an interview, errors and inconsistencies are detected by the software while a survey team is still working in the enumeration area. Thus teams can return to households for which the software detects errors, correct the errors, and enter the corrected data into the computer.

To estimate the impact of decentralized data entry, a classic experimental design was used in 23 provinces during the fourth quarter (October–December) of the 2002 VHLSS. In those provinces all districts included in the fourth quarter of the survey participated in the experiment, a sample of 2,895 households in 202 districts. These districts were divided into "dispersed" and "compact" districts. "Dispersed districts" were districts too dispersed geographically to allow interviewers to return to the District Statistical Office in the district capital within 24 hours of the interview for some or all of the sampled households. Of the 202 districts in the "population" of the 23 provinces, 97 were classified as dispersed. Ten were randomly selected to be treatment districts; the other 87 were controls. The teams in the 10 treatment districts received laptop computers for entering data from household questionnaires at the enumeration area within hours of each interview.

The other 105 districts in the 23 provinces were "compact districts"—districts small enough so that interviewers could return to the District Statistical Office within 24 hours of the interview. Fifteen districts were randomly selected as treatment districts; the other 90 were controls. Each treated district received a desktop computer, kept at the District Statistical Office in the district capital, which permitted data to be entered a day or two after each interview.

Thus 25 of the 202 districts in the 23 provinces were randomly provided laptop or desktop computers to speed data entry and allow interviewers to return to any households for which the software detected unusual or inconsistent data. No additional training was given to interviewers or supervisors in the districts that received computers.

II. Testing the Impact of Decentralized Data Entry

This section describes the statistical methods used to estimate the impact of providing data-entry computers to each district in Vietnam. It also explains how the data were examined to detect nonsampling (measurement) errors.

Statistical Methods Used to Estimate Impact

The impact on data quality of changing a household survey's methodology can be assessed in two ways. First, one can estimate the impact of changes on observed means, variances, and other functions of variables of interest. Second, one can investigate how changes affect the prevalence of data errors. This subsection discusses statistical methods for both approaches.

Consider a variable y. It could be household income or expenditure, the number of errors in a completed questionnaire, a dummy variable indicating a household's poverty status, or the household's "poverty gap." (The poverty gap is the poverty line minus household income; it equals zero if this difference is negative.) The objective is to test whether the observed mean of y varies across the treatment and control groups. If, for example, y measures poverty, more accurate poverty measurement from improved data collection could induce reallocations of government resources to regions or households for which current data underestimate poverty. Examples of such research are Chaudhuri and Ravallion (1994) and Gibson and others (2003).

Of course, changes in the observed mean of y as a result of new data collection methods do not necessarily imply that the new estimate is better; a change could increase bias in the estimated mean. If the observed mean has changed, the new data collection method must be carefully considered to determine whether bias has increased or decreased. The rest of this subsection explains how the observed means of y were checked to determine whether they changed after decentralized data entry was implemented in Vietnam.[3]

If y has a finite mean and variance, then by the Lindberg–Levy Central Limit Theorem its sample mean is asymptotically normally distributed; the asymptotic variance is $\text{Var}[y]/n$ (n is the sample size). If the variance of y is equal across the treatment and control groups, the difference in the sample means of those two groups is also asymptotically normally distributed, with an asymptotic variance $\text{Var}[y](1/n_t + 1/n_c)$, where n_t is the sample size of the treatment group and n_c is the control group. A potentially more precise test compares the difference in the means of the treatment and control groups of these households (which were interviewed in the fourth quarter of the survey) with the same difference for households interviewed in the first three quarters, a difference in differences test. If y has the same variance across all four groups, the double difference $(\bar{y}_{t,\,4} - \bar{y}_{c,\,4}) - (\bar{y}_{t,\,123} - \bar{y}_{c,123})$ is asymptotically normal, with a variance of $\text{Var}[y](1/n_{t,4} + 1/n_{c,4} + 1/n_{t,123} + 1/n_{c,123})$.

Regression analysis can test differences in the mean of y across treatment and control groups. Using only the fourth-quarter data, ordinary least squares regression of y on a constant and a dummy variable indicating the treatment group produces a coefficient on that variable with a t-statistic exactly equal to the test statistic for the difference in means described above. The test using all four quarters of data is replicated by regressing y on a constant, a dummy variable indicating residence (in any quarter) in a district that was treated in the fourth quarter, a dummy indicating households interviewed in the fourth quarter, and an interaction term between both variables. The coefficient on the interaction term tests whether the treatment affected the mean of y.

3. For expositional ease the discussion refers to the variable y; more precisely, it concerns observed measurements of y.

Adding control variables to these regressions may improve the precision of the estimated coefficients. Since the treatment variable is uncorrelated with any variable, adding regressors will not lead to biased estimates of treatment effects.

The first regression uses only the 2,895 households that participated in the experiment in the fourth quarter of the survey. Of these, 2,475 are in the control group and 420 are in the treatment group. They can be used to estimate

$$y = \beta_0 + \beta_1 P + \boldsymbol{\beta}_2' \boldsymbol{x} + \varepsilon \tag{1}$$

where y is the variable of interest, P is the "program" dummy variable that indicates households living in districts that received computers, $\boldsymbol{x}$ is a vector of control variables, and ε is an error that is uncorrelated with P because P was randomized. All estimated standard errors in this article account for clustering at the enumeration area level.

Ignoring the $\boldsymbol{x}$ covariates, β_1 in equation 1 equals $E[\bar{y}_{t,\,4} - \bar{y}_{c,\,4}]$, the difference in the mean of y across the treatment and control districts for the 2,895 households interviewed in the fourth quarter of 2002. That is, $E[\bar{y}_{t,\,4}] = \beta_0 + \beta_1$ and $E[\bar{y}_{c,\,4}] = \beta_0$, so $E[\bar{y}_{t,\,4} - \bar{y}_{c,\,4}] = \beta_1$.

The second regression method adds the households interviewed in the first three quarters (January–September, 2002) in the 202 districts that participated in the experiment in the fourth quarter:

$$y = \beta_0 + \beta_1 Q_4 + \beta_2 P + \beta_3 Q_4 P + \boldsymbol{\beta}_4' \boldsymbol{x} + \varepsilon \tag{2}$$

where Q_4 is a dummy variable indicating households interviewed in the fourth quarter. Ignoring $\boldsymbol{x}$, $\beta_3 = E[(\bar{y}_{t,\,4} - \bar{y}_{c,\,4}) - (\bar{y}_{t,\,123} - \bar{y}_{c,\,123})]$, the difference between the difference of the mean of y across the treatment and control districts in the fourth quarter and the analogous difference in the first three quarters. That is, ignoring covariates, equation 2 implies $E[\bar{y}_{t,\,4}] = \beta_0 + \beta_1 + \beta_2 + \beta_3$, $E[\bar{y}_{c,\,4}] = \beta_0 + \beta_1$, $E[\bar{y}_{t,\,123}] = \beta_0 + \beta_2$ and $E[\bar{y}_{c,\,123}] = \beta_0$, so $E[(\bar{y}_{t,\,4} - \bar{y}_{c,\,4}) - (\bar{y}_{t,\,123} - \bar{y}_{c,\,123})] = (\beta_2 + \beta_3) - \beta_2 = \beta_3$.

Adding a $P\boldsymbol{x}$ term to equation 1 and a $Q_4 P\boldsymbol{x}$ term to equation 2 allows one to estimate whether the program impact varies by household characteristics ($\boldsymbol{x}$).[4] Finally, separate regressions can be estimated for dispersed and compact districts, to see if the impacts vary across these two subexperiments.

4. In theory, one should also add $Q_4\boldsymbol{x}$ and $P\boldsymbol{x}$ as regressors to equation 2, but β_2 should equal zero (the program should have no impact before its implementation), so adding $P\boldsymbol{x}$ is unnecessary. $Q_4\boldsymbol{x}$ was added but later dropped, because it was seldom significant.

Direct Detection of Nonsampling (Measurement) Errors

Examining a variable of interest to see whether changing survey procedures changes its mean or another function of its distribution *indirectly* assesses whether those changes affected errors in household survey data, because that approach does not search for explicit errors. In one sense, such indirect evaluation of changing survey procedures is all that is required, because only changes in the observed distributions of those variables affect policy decisions. Yet it is also instructive to search for measurement errors (or likely errors) directly in variables of interest and even in other variables, because this information can reveal whether particular types of households tend to yield error-ridden data.

This section explains how to analyze measurement errors directly. Almost all variables from household surveys have errors, but only some errors are apparent in the data. Errors can be divided into those that can be detected by analyzing survey data and those that cannot. Examples of detectable errors are an age of 150 years or consumption of 10 kilograms of meat per person per day. Perhaps more common are inconsistencies between two or more variables, such as a five-year-old child who is married. Examples of undetectable errors are an incorrect age of an adult (if it does not contradict other variables, such as date of birth) and errors in income or consumption expenditures (assuming no contradictions with other data).

The distinction between detectable and undetectable errors is not always sharp. Some information may be doubtful but still possible, such as a 15-year-old who reports having finished secondary school or a farmer who reports unusually high fertilizer use per hectare. In practice, one can treat dubious data anomalies that have a small probability of being correct as detectable errors and classify errors that produce unusual (but still plausible) data patterns as undetectable errors. Dividing errors into these two types is subjective, but if the judgments made are explicit, the results should not be misleading.

By definition, undetectable errors cannot be directly analyzed by examining the data; they can be examined only by reinterviewing households soon after the initial interview or by checking other data sources (for example, employer's payroll data).[5]

Fortunately, detectable errors can be directly analyzed. It is useful to divide them into two subtypes: those explicitly checked by the 2002 VHLSS data-entry program and those that were not checked by that program but can be checked by examining the data. This distinction is important, because the data-entry software determines how much can be learned from analyzing each subtype of detectable errors. In particular, some data entry programs force the

5. Undetectable errors can also be analyzed indirectly, by looking at changes in the distributions of the variables of interest after changing survey methods.

data to be consistent with all explicit error checks, so the data provided by the software are "error free"—that is, the data will never have detectable errors of the type checked by the software. Most important, if survey teams cannot return to households to clear up errors detected by the software, data-entry operators or other survey staff must change the data in the office to remove all apparent errors. Such changes could leave in place real errors if the "correction" relies on guesswork, which is often the case. Thus checking only for errors that the software checks yields no information on the impact of survey design changes on nonsampling errors. In contrast, if the software does not force data-entry staff to "fix" errors that it detects and staff are not instructed to "correct" all detected errors, the effect of providing computers to each district may appear in the errors that the program explicitly checks (as well as other detectable errors, as explained below).

Regardless of whether the software forces survey staff to "correct" the errors it finds, one can still detect improvements in data quality by examining errors not explicitly checked by the software. Intuitively, if moving data entry closer to where households are interviewed allows interviewers to return to households to clear up errors detected by the program, the corrected data may reduce errors that are not explicitly checked by the software. For example, the data-entry program may find that the sum of the area of a household's plots of land does not equal the reported total amount. Explicit checking may detect an error in a plot's area, and correcting that error could lead to more accurate crop yields on that plot of land. This article examines both types of detectable errors.

The 2002 VHLSS income and expenditure household questionnaire has nine sections (box 1). The VHLSS data-entry program performed more than 100 data checks. Space constraints preclude describing any but the most common. First, several questions involved summing numerical values from specific questions; many data checks in the program verified those sums. Second, a few checks verified whether the number of people with information from a given questionnaire section matched the people listed in the household roster (Section 1). For example, Section 3 (employment) was to be completed for all household members aged 10 and older; one check is whether Section 3 included information for everyone in the household roster age 10 or older. Third, several data-entry checks involved questions about household participation in certain production activities. If the household reported participation, the data-entry checks looked for the expenditures on or income from those activities; if the household reported no involvement in an activity, the expenditures or income for that activity should be zero (or missing).[6] Fourth, for households reporting land owned or livestock or crops sold or used for other purposes, the sum of

6. The questionnaire uses skip codes, so households reporting no involvement in an activity will have missing values for that activity's income and expenditure variables. In practice, the skip codes were almost always followed, and missing data that should not be missing are rare.

the disaggregated amounts (for example, certain types of land or specific uses of livestock or crops) should equal the total amounts reported for all types or uses. Fifth, many questions had preassigned codes (such as occupation codes or industry codes); the software checked that the values recorded in the data matched one of these codes. The program does not "force" data to be consistent; data-entry personnel can enter data that violate the checks performed by the data-entry program.

BOX 1. STRUCTURE OF 2002 VHLSS INCOME AND EXPENDITURE HOUSEHOLD QUESTIONNAIRE

Section 1: List of Household Members (Roster)
Section 2: Education
Section 3: Employment
Section 4: Health
Section 5: Income and Other Inflows of Money
- Part A. Income from Salary and Wages
- Part B. Agricultural, Forestry, and Fishery Activities
  1. Agricultural, Forest, and Aquaculture Land
  2. Agricultural Production
  3. Income from Livestock
  4. Income on and Expenses for Farm Services
  5. Income from and Expenditure on Tree and Forest Crops
  6. Income from and Expenditure on Aquaculture
- Part C. Nonfarm Businesses and Processing Farm, Forest, and Fishery Products
- Part D. Other Sources of Income

Section 6: Expenditure
- Part A. Expenditure on Food and Drink
- Part B. Expenditure on Nonfood and Other Expenditures
- Part C. Other Expenses (contributions, taxes, and so forth)
- Part D. Other Expenses not Included as Expenditure (savings, insurance, and so on.)

Section 7. Fixed Assets and Durable Goods
Section 8. Housing
Section 9. Participation in Poverty Alleviation Programs

Source: GSO 2006.

The 2002 VHLSS data-entry program undoubtedly detected many errors. The protocol for correcting errors (for control group households) was the following. The supervisor at the provincial capital, where all data were entered, corrected "obvious" errors in the provincial office. For errors without obvious resolutions, the supervisor was to telephone the district office to ask the district

team to investigate, including (if necessary) by revisiting the household. Little is known about what actually happened, but examination of the data for the types of errors detected by the program reveals very few errors (about 0.03 per household). Thus the data-entry program was used to "clean" almost all errors explicitly checked by that program. The extent to which the corrections were valid, as opposed to "educated guesses" by survey teams at the provincial and district levels, is unclear; discussions with survey staff suggest that they made very few revisits to households.

Almost every section of the questionnaire has variables, or combinations of variables, that were not checked by the data-entry program but could have errors or inconsistencies. For example, Section 1 collected data on gender, relationship to the household head, date of birth, age, and marital status. Possible inconsistencies among these variables include age and date of birth; people identified as the head's spouse who report not being married; people reported as children of the head who are no more than 14 years younger (or more than 50 years older) than the head; and the head and his or her spouse reporting the same gender. Inconsistencies may also exist across sections, such as young children with high levels of education, households that report being self-employed farmers but report neither owning nor renting land, and households that report consuming a certain crop grown by the household but not reporting growing that crop in the agriculture section. (The appendix in the long version of this article, available at http://wber.oxfordjournals.org/, presents an extensive list of errors, inconsistencies, and likely errors that were not checked by the 2002 VHLSS software.)

Several control variables ($\boldsymbol{x}$) were added to the regression analysis of detectable errors. First, to control for unobserved quality in survey personnel, province dummy variables were added (district dummy variables yielded similar, but less precise, results). Second, household size was added; because many questionnaire sections collect information on each household member, larger households provide more information and thus have more opportunities for errors. Third, per capita expenditure was added; wealthier households buy more goods, increasing the amount of information collected and thus increasing opportunities for errors.

III. Empirical Results

This section uses the methodology described above to estimate the impact of decentralized data entry on the measured distribution of per capita expenditures and in the number of errors in the data. Per capita expenditure is of particular interest, because the VHLSS uses that variable to monitor poverty in Vietnam.

Impact on the Mean and Variance of per Capita Expenditures and on Poverty

Vietnam's GSO implements the VHLSS every two years. The results on poverty receive widespread attention both inside and outside Vietnam. This subsection

examines whether introducing computers in each district changed the observed distribution, and functions of the distribution, of per capita expenditures.

If measurement errors were random (uncorrelated with per capita expenditures), with a zero mean, then reductions in measurement errors from providing computers to each district would not affect the mean of observed per capita expenditures but would reduce the observed variance. Since most Vietnamese are not poor, reduced variance would reduce measured poverty (Ravallion 1994). Yet measurement errors may not be random, so decentralized data entry could affect the mean of observed expenditure, leaving the impacts on variance and poverty ambiguous.

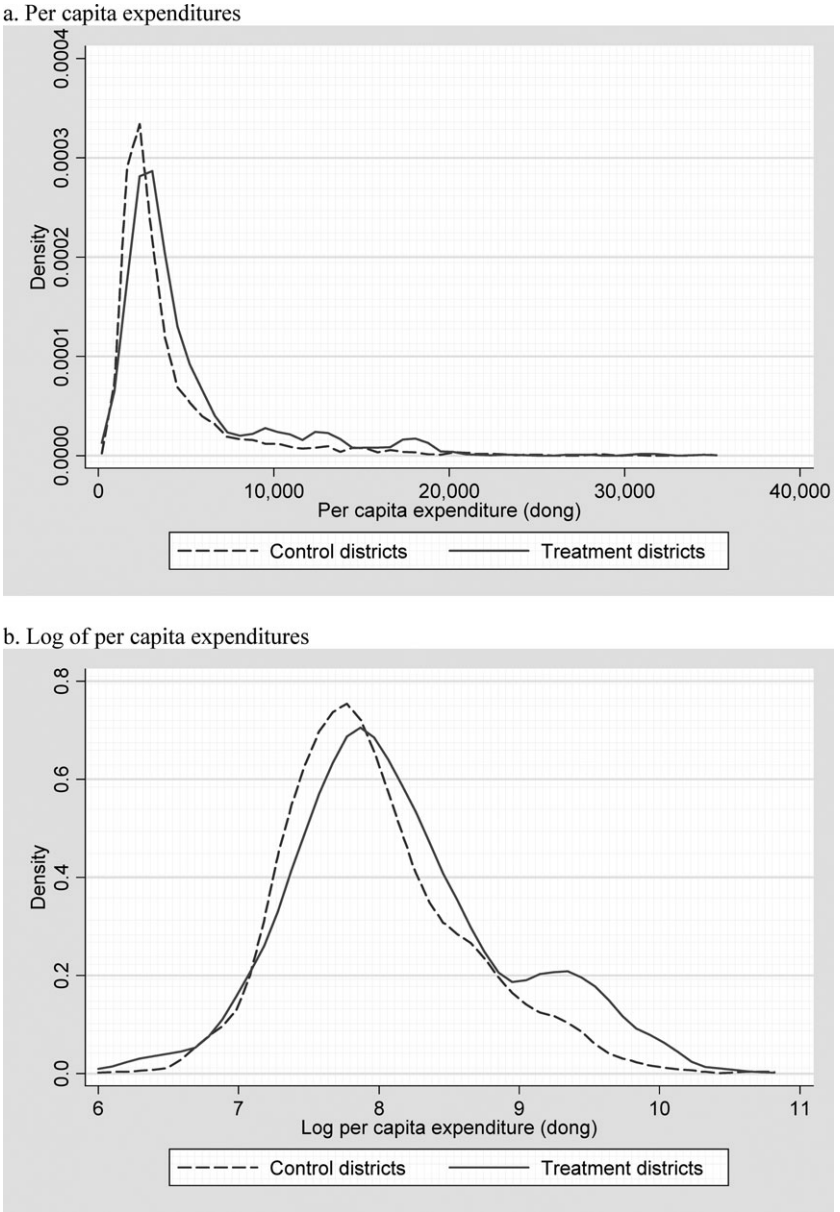
To assess the impact of decentralized data entry on measured per capita expenditures, the density function of that variable can be compared for the treatment and control households in the fourth quarter of the survey. In both panels of figure 1 the density for treatment households lies to the right of the density for control households, but only slightly so. The densities for log per capita expenditures suggest that the variance of the treatment households may be larger than the variance for control households, albeit only slightly. Indeed, Kolomogorov–Smirnov tests fail to reject, even at the 10 percent level, the null hypothesis that the distributions of per capita expenditure for the treatment and control groups are equal.

Next consider parametric tests for particular summary statistics, starting with the variance (table 1). The standard deviation of expenditures per capita is larger for the treatment group than for the control group, indicating that any "corrected" errors tended to regress to the mean (positive errors for low per capita expenditures and negative errors for high values). The Brown and Forsythe (1974) robust test of equality of variances (which is needed, because figure 1 shows that the distribution of expenditures per capita is skewed) reveals that this difference is statistically significant (p-value of 0.04).

This statistical significance is doubtful, however, for several reasons. First, applying that test to log per capita expenditures fails to reject the null hypothesis of equal variances at the 5 percent level (though it does at the 10 percent level). Second, and more important, these tests are biased toward rejecting the null hypothesis, because they ignore the fact that the data are clustered and weighted, which (if properly accounted for) raises the standard error of almost any test statistic. Third, the statistical significance of the difference in variances was also checked by bootstrapping (accounting for clustering); the null hypothesis of equal variances was not rejected, even at the 10 percent level. Thus there is little evidence that providing computers to each district affected the variance of per capita expenditures.

Turn now to the mean of per capita expenditures. By definition, random measurement errors have a zero mean and thus cannot affect the mean. If their mean is not zero, and decentralized data entry reduces measurement errors, the mean of observed expenditures should differ across treatment and control

FIGURE 1. Kernel Density Estimates of per Capita Expenditures



Note: Automatic bandwidths are used for all kernel density estimates.
Source: Authors' analysis based on data described in text.

groups. The second line of table 1 shows that the mean for the treatment group is 3,911 thousand dong, about 7.5 percent higher than the mean for the control group (3,636 thousand dong). The difference (275 thousand dong) is statistically insignificant (using tests that incorporate clustering and sample

TABLE 1. Differences in Distribution of per Capita Expenditures

Statistic	Treatment group	Control group	Difference
Standard deviation (thousands of dong)	4,034.0	3,680.5	353.5
Mean (thousands of dong)	3,910.9	3,636.4	274.6 (416.7)
Poverty rate	0.250	0.269	−0.019 (0.037)
Poverty gap	0.061	0.060	0.001 (0.012)
Squared poverty gap	0.024	0.020	0.004 (0.007)

Note: Numbers in parentheses are standard errors of the differences in means. The difference for the standard deviation is statistically insignificant, as explained in the text. Sample size is 2,895 for all rows.

Source: Authors' analysis based on data described in the text.

weights). Thus provision of computers had little effect on the mean of measurement errors (whether zero or nonzero) in per capita expenditure, leaving unchanged any bias caused by those errors.

Introducing computers at the district level could affect poverty, as measured by per capita expenditures, even if it does not affect mean expenditures. Districts that received computers have a slightly lower poverty rate, but the poverty gap was almost identical and the squared poverty gap (which is sensitive to inequality among the poor) was slightly higher. None of these differences is statistically significant, suggesting that providing computers to each district had no effect on measured poverty.

Double-difference estimates of the impact of decentralized data entry may be more precise and therefore more likely to detect any impacts (table 2). The standard errors of these double-difference estimates (equation 2, without covariates) for mean per capita expenditures and the three poverty indices reveal no increase in precision, and all differences remain insignificant.

TABLE 2. Double-Difference Estimates of Change in per Capita Expenditures

	Rounds 1–3		Round 4		
Statistic	Treatment group	Control group	Treatment group	Control group	Double-difference
Mean (thousands of dong)	3987.6	3135.5	3762.2	3704.2	−794.1 (592.3)
Poverty rate	0.253	0.282	0.259	0.260	0.027 (0.048)
Poverty gap	0.053	0.061	0.061	0.056	0.012 (0.014)
Squared poverty gap	0.016	0.019	0.023	0.018	0.008 (0.007)

Note: Numbers in parentheses are standard errors of the differences in means. Sample size is 11,040 for all rows. Figures for Round 4 are slightly different from those in table 1 because 40 observations were dropped from five districts that were surveyed in Round 4 but not in Rounds 1–3.

Source: Authors' analysis based on data described in the text.

TABLE 3. Regression Estimates of Program Impacts on Error Rates

	Mean errors per household			Regression coefficients on computer-variable	
Source of errors	Quarters 1, 2, and 3 (before experiment)	Quarter 4 control group	Quarter 4 treatment group (with district-level computers)	Level regression (quarter 4 only)	Double-difference regression (all quarters)
All individual sections	0.256	0.229	0.190	−0.183 (0.201)	−0.249 (0.196)
Household sections	0.073	0.077	0.047	−0.479 (0.310)	−0.434 (0.303)
All sections	0.329	0.306	0.236	−0.278* (0.164)	−0.307* (0.162)

*Statistically significant at the 10 percent level.

Note: Results are negative binomial regression estimates. All figures are for the 23 provinces that participated in the experiment. All regressions include a computer dummy variable, province dummy variables, household size, and expenditure per capita. Double-difference regressions add quarter dummy variables. Numbers in parentheses are standard errors, which account for clustered sampling. Sample sizes are 2,895 for the level regressions and 11,040 for double-difference estimates.

Source: Authors' analysis based on data described in the text.

To summarize, both parametric and nonparametric methods find no impact of decentralized data entry in Vietnam on the distribution of per capita expenditures or poverty indices. Point estimates of differences are small, especially in table 1. Providing data-entry computers to each district is thus unlikely to generate results that would change Vietnam's economic policies.

Impact on Detectable Errors

Level and double-difference estimates are presented for the full sample (compact and dispersed districts), with each row representing a separate negative binomial regression (table 3).[7] The fourth and fifth numbers in each row are the negative binomial coefficients that estimate the impact on errors of providing district-level computers; the fourth number is an estimate of β_1 in equation 1, and the fifth is an estimate of β_3 in equation 2. If decentralized data entry reduces errors, these coefficients should be negative. Standard errors are given in parentheses.

The first row examines all errors for individual level sections of the VHLSS questionnaire. The mean number of errors for all these sections is 0.256 for quarters 1–3 and 0.229 for the fourth-quarter control group. Both are larger than the treatment group mean (0.190), but these differences are statistically insignificant in both the level and double-difference regressions. The second row examines all errors in the household-level sections of the VHLSS questionnaire. The mean

7. All regressions were also estimated using ordered probits; the results were very similar. Additional estimates for individual sections of the VHLSS questionnaire are given in the online version of this article, available at http://wber.oxfordjournals.org/.

errors for quarters 1–3 (0.073) and the fourth-quarter control group (0.077) are noticeably higher than for the fourth-quarter treatment group (0.047), but the estimated impact falls just short of significance at the 10 percent level.

The third row sums both individual- and household-level errors. Among households interviewed in quarters 1–3 there are 0.329 errors per household; for the fourth-quarter control group there are 0.306. The figure for districts given computers in the fourth quarter is 0.236. The last two numbers suggest that decentralized data entry reduced aggregate errors by 23 percent, a statistically significant difference (at the 10 percent level) for the level and double-difference estimates. A simple *t*-test of the difference in means across the treatment and control groups in the fourth quarter is also statistically significant (at the 10 percent level).[8]

These results combine compact districts, which received desktop computers, and dispersed districts, which received laptops. The impact of providing computers could vary by type of computer or type of district. Survey teams in dispersed districts travel to the enumeration areas and stay there for several days before returning to the district capitals. The working conditions are difficult, suggesting potential for large reductions in errors from bringing data entry closer to interview sites. On the other hand, using laptops in these areas could entail additional difficulties, such as the lack of electricity for recharging batteries or the damage caused to computers by moving them frequently.

For dispersed districts, in the individual sections of the questionnaire there is no apparent impact of providing computers on total errors: the fourth-quarter control group had slightly fewer errors (0.251) than the treatment group (0.266) (table 4). The regression estimates show a statistically insignificant negative impact of providing laptops on individual-level errors. Yet the mean errors on the household-level sections in the treatment group (0.045) are less than half of the mean in the fourth-quarter control group (0.105), a statistically significant effect (at the 5 percent level). With both types of errors combined, the mean errors for treatment group households (0.311) is 13 percent lower than the error for the fourth-quarter control group (0.356), but it is statistically insignificant.

Providing desktop computers to compact districts reduces errors on individual-level sections by 41 percent, from 0.213 to 0.126 (table 5). This is statistically significant (at the 5 percent level) in the level regression but not in the double-difference regression. The drop in household-level errors is smaller (about 21 percent) and statistically insignificant. Combining both types of errors yields a 37 percent reduction in errors (from 0.266 to 0.168), which is statistically significant at the 5 percent level in the level regression and at the 10 percent level in the double-difference regression.

8. The online version of this article examines errors of consistency between what households report as consumption from own production in the food expenditures section of the VHLSS questionnaire and what they report regarding crops grown and animals raised in the farming section. There was no discernable impact on such errors of providing computers to each district, but these data were very noisy and some apparent errors may not be errors at all.

TABLE 4. Regression Estimates of Program Impacts on Error Rates: Dispersed Districts

	Mean errors per household			Regression coefficients on computer-variable	
Source of errors	Quarters 1, 2, and 3 (before experiment)	Quarter 4 control croup	Quarter 4 treatment group (with district-level computers)	Level regression (quarter 4 only)	Double-difference regression (all quarters)
All individual sections	0.235	0.251	0.266	−0.254 (0.365)	−0.237 (0.330)
Household sections	0.077	0.105	0.045	−0.928** (0.438)	−0.838** (0.415)
All sections	0.313	0.356	0.311	−0.427 (0.305)	−0.383 (0.285)

**Statistically significant at the 5 percent level.

Note: Results are negative binomial regression estimates. In addition to the computer dummy variable, all regressions include province dummy variables, household size, and per capita expenditures. The double-difference regressions add quarter dummy variables. Numbers in parentheses are standard errors, which account for clustered sampling. The sample size was 1,285 for the level regressions and 4,840 for the double-difference estimates.

Source: Authors' analysis based on data described in the text.

The different results in the dispersed and compact districts make sense. Dispersed districts are very rural, and almost all adults are self-employed farmers; most households complete the detailed household-level agricultural section. Because they are less likely to work for wages, seek medical care, or

TABLE 5. Regression Estimates of Program Impacts on Error Rates: Compact Districts

	Mean errors per household			Regression coefficients on computer-variable	
Source of errors	Quarters 1, 2, and 3 (before experiment)	Quarter 4 control group	Quarter 4 treatment group (with district-level computers)	Level regression (quarter 4 only)	Double-difference regression (all quarters)
All individual sections	0.282	0.213	0.126	−0.573** (0.291)	−0.476 (0.303)
Household sections	0.068	0.053	0.042	−0.290 (0.496)	−0.133 (−0.484)
All sections	0.350	0.266	0.168	−0.536** (0.224)	−0.439* (0.230)

*Statistically significant at the 10 percent level.

**Statistically significant at the 5 percent level.

Note: Results are negative binomial regression estimates. In addition to the computer dummy variable, all regressions include province dummy variables, household size, and per capita expenditures. The double-difference regressions add quarter dummy variables. Numbers in parentheses are standard errors, which account for clustered sampling. The sample size was 1,605 for the level regressions and 5,660 for the double-difference estimates.

Source: Authors' analysis based on data described in the text.

send their children to school, they provide less information on the individual sections. Because they provide more data on household-level sections and less data on individual-level sections, they are more likely to have household-level errors and less likely to have individual-level errors than are households in the (more-urbanized) compact districts. This tendency, seen in the first columns of tables 4 and 5, implies that improved data entry will have greater effects on household-level errors in dispersed districts and on individual-level errors in compact districts.

Does the impact of providing computers vary across households? The question can be investigated using regression analysis by adding interaction terms between the program dummy variable (P) and certain household characteristics that may increase errors. One example is household size: more household members giving individual-level information implies more opportunities for errors. Another is per capita household expenditures: better-off households purchase more food and nonfood goods, own more durables, and use more health services, increasing the data collected and thus the possibilities for errors. A third example is the education of the household head (the primary survey respondent): better-educated household heads may make fewer errors answering household questions. Surprisingly, none of these three interaction terms is statistically significant; no observable household characteristics appear to increase or reduce the incidence of errors.

IV. Implications for Data Collection in Vietnam

Providing data-entry computers to each district (instead of each province) has little effect on the measured distribution of per capita expenditures and thus little effect on poverty estimates for Vietnam. Providing computers to districts does reduce detectable errors in the VHLSS for many variables, however. When errors of consistency for reported consumption of home-produced crops and livestock are included, introducing computers reduces errors by only (a statistically insignificant) 5 percent. When such errors are excluded, providing computers to districts reduces errors by (a statistically significant) 23 percent. Errors are reduced in both dispersed districts (which received laptops) and compact districts (which received desktops), but the drop in errors was larger in compact districts (37 percent) than in dispersed districts (13 percent). The impact of computers did not vary by household characteristics.

Does decentralized data entry merit the cost of purchasing computers for all 607 districts in Vietnam? The GSO spends about $500 on desktop computers and about $1,200 on laptops; both require $100 in training costs. About one-third of Vietnam's districts are dispersed and thus require laptops. Assuming that 407 desktop computers and 200 laptops are needed, the total cost is $504,200.

It is harder to value the benefits of purchasing computers. None was found in terms of improved measurement of poverty or per capita expenditures. But it appears that providing computers reduces errors for many other variables.

To assess the value of reducing these errors, consider a random variable x and two alternatives for reducing the standard error of its estimated mean: increasing the sample to reduce the contribution of sampling error to the standard error (of the estimated mean of x) or providing computers to reduce nonsampling errors. The standard error of the mean of x is the standard deviation of x divided by the sample size: $SD(x)/\sqrt{N}$.[9] Assume that the errors have a mean of zero and are uncorrelated with the true value of x.[10] Providing computers reduces $SD(x)/\sqrt{N}$ by reducing random measurement error and thus reducing $SD(x)$; increasing the sample reduces $SD(x)/\sqrt{N}$ by increasing $\sqrt{N}$.

Consider the likely impact on $SD(x)$ of providing computers to each district. Assume that all errors, including those corrected by providing computers, have a zero mean, are uncorrelated with the true x, and have a standard deviation equal to $SD(x)$. The last assumption likely overestimates the true standard deviation of these errors. In particular, the difference between two observations randomly drawn from a normal distribution (the error from replacing the value of x for one household with that of another randomly selected household) is $0.798 \times SD(x)$ (Johnson and others 1994), so what follows is an upper bound of the impact on $SD(x)/\sqrt{N}$ of providing computers to each district. Finally, for simplicity suppose that x is usually measured without error, but for a small fraction of the observations (denoted by k) the observed x is the sum of the true value and the random error with a standard deviation of $SD(x)$. Thus the standard deviation of observed x equals $\sqrt{1+k}$ multiplied by the true standard deviation of x.[11]

Providing desktops to compact districts reduced the mean errors per household from 0.266 to 0.168 (see table 5). This result was based on checking about 40 variables, which implies a reduction in the error rate (k) from about

9. This standard error assumes a simple random sample. Yet almost all household surveys, including the VHLSS, have a multistage clustered sample design. The analysis presented here still holds, because the correct standard error of the estimated mean is $\rho SD(x)/\sqrt{N}$, where ρ is the design effect, which is unchanged when either more computers are added or the sample size is increased using the same multistage clustered design.

10. It is impossible to determine whether measurement errors are biased (have a nonzero mean), but the results presented in section II reveal no evidence that providing computers affected any bias that exists for per capita expenditures. It is also impossible to determine whether measurement errors are correlated with the true values of variables. Yet in most cases correlated errors raise the variance of observed x, so the following analysis is relevant. To see this, let x^* denote the true value of x and let u be the measurement error, so x, the observed value, is $x^* + u$. Assume a linear correlation between u and x^*: $u = \beta(x^* - \mu_{x^*}) + \varepsilon$, where μ_{x^*} is the mean of x^* (to ensure that u has a zero mean) and ε is a random error uncorrelated with x^*. Thus $\mathrm{Var}(x) = \mathrm{Var}(x^* + u) = \mathrm{Var}(x^*(1+\beta) - \beta\mu_{x^*} + \varepsilon) = (1+\beta)^2\mathrm{Var}(x^*) + \mathrm{Var}(\varepsilon)$. Clearly, $\mathrm{Var}(x) > \mathrm{Var}(x^*)$ if the correlation is positive ($\beta > 0$). For negative correlation ($\beta < 0$), $\mathrm{Var}(x) < \mathrm{Var}(x^*)$ only if $(2\beta - \beta^2)\mathrm{Var}(x^*) > \mathrm{Var}(\varepsilon)$. Intuitively, for $\mathrm{Var}(x) < \mathrm{Var}(x^*)$ the impact of the negative correlation must outweigh that of the random component (ε) of u.

11. Given these assumptions, the variance of observed x is the variance of the "true" values plus the variance of the random errors that occur for k percent of the sample. The variance of the random errors that occur for k percent of the sample variance equals the variance of the true values of x, so the variance of the observed values is $1 + k$ times the variance of the true values, and the standard deviation is $(1+k)^{1/2}$ times the true standard deviation.

0.0067 to about 0.0042 for a typical variable. Thus the benefit of providing computers to the 407 compact districts in Vietnam for reducing $SD(x)/\sqrt{N}$ is that $SD(x)$ decreases by a factor of $\sqrt{1.0067/1.0042} - 1$, a reduction of 0.1 percent. This gain is very small given its cost of \$244,200 ($407 \times \$600$). Indeed, a similar gain is obtained by increasing the sample by 0.2 percent, that is, adding just 40 households to the approximately 20,000 households in the compact districts. Increasing the sample costs about \$30 per household (personal communication from GSO), so increasing the size of the sample by 40 households would cost \$1,200—just 0.5 percent of the cost of 407 new computers. Even if the methods used in this article detected only one-fourth of the errors avoided by providing each district with a computer, so that providing computers reduces k from 0.0268 to 0.0168, provision of computers would reduce the standard deviation of observed x by only $\sqrt{1.0268/1.0168} - 1$, or 0.5 percent. This reduction in the standard error of the estimated mean of x can be realized by adding 200 households to the sample, at a cost of about \$6,000, or 2.5 percent of the cost of 407 computers.

Two other factors should be considered when assessing the benefits of purchasing computers for each district in Vietnam. First, the VHLSS is implemented every two years, so each computer can be used for two or three surveys before becoming obsolete; the appropriate comparison is thus 407 computers and 400–600 households (generously assuming that the analysis uncovered only one-fourth of the errors in the data that computers would correct). Second, the computers can be used for two other surveys of similar size. Thus the increased precision of the estimated mean of x from buying computers is equivalent to increasing the sample (summing over three surveys) by 1,200–1,800 households. The cost of increasing the sample to 1,800 is about \$54,000, still only 22 percent of the cost of 407 new desktops.

Yet another benefit of providing computers to each district is that they reduce the time VHLSS interviewers spend manually calculating sums and checking the data. According to Truong (2003), providing computers to each district saves about 3.5 hours of interviewer time for each household (for both dispersed and compact districts).

The VHLSS covers 30,000 households. Given 607 districts in Vietnam, the average district contributes about 50 households. Thus purchasing a computer saves about 175 hours of interviewer time per district. Since the VHLSS takes place every two years and the computers last for two or three surveys, the use of computers saves about 500 interviewer hours. Including other surveys the GSO could implement, using those computers could increase time saved to 1,000–1,500 interviewer hours. The GSO pays a typical interviewer in Vietnam about \$0.20 per hour, so each computer purchased can reduce costs by \$200 to \$300. This is less than the \$600 cost of a desktop computer (including training costs), but wages are rising in Vietnam and computer prices are falling; in a few years it may be cost-effective for the GSO to purchase desktop computers for compact districts in Vietnam.

V. Conclusion

Providing computers in Vietnam had little impact on measured poverty and per capita expenditures.[12] Yet regression analysis shows a statistically significant 23 percent reduction in (detectable) errors for many other variables (excluding errors of consistency between food consumption and production data). This reduction is higher for "compact districts" (41 percent), which were given desktop computers. Is this reduction worth the cost? Given the already low rate of errors in the VHLSS data, the answer appears to be "no." Simple calculations (which assume random errors) suggest that standard errors in estimated means of variables of interest can be reduced less expensively by increasing the sample size slightly.

Yet there is another benefit of providing computers to each district. One reason why the error rate of the VHLSS is low is that interviewers currently spend several hours manually checking each questionnaire. Truong's (2003) study suggests that each computer can save 1,000–1,500 hours of interviewer time, implying a savings per computer of $200–$300. A desktop computer costs $600, which is more than the amount saved, but with wages rising and computer prices falling, purchasing new desktop computers may be cost-effective in a few more years.

These results shed light on the merits of moving data-entry computers closer to where interviewers work. Even so, several caveats must be kept in mind. First, the estimated impacts of providing more computers on observed per capita expenditures and poverty are imprecise. An experiment with a larger sample might reveal policy-relevant impacts on these variables.

Second, estimates of the benefits of reducing errors in other variables assume that those errors are unbiased and uncorrelated with those variables' true values. The conclusion that reducing the standard errors of estimated means by providing more computers is much more expensive than doing so by increasing the sample depends on that assumption (although any change in assumptions about measurement errors must be dramatic to overturn this conclusion).

Third, Vietnam's situation may differ from that of other countries. The experience of the first author in many developing countries suggests that household survey data in other countries have far more errors, which implies greater benefits from introducing computers in other countries, especially those with higher interviewer wages. Moreover, the data analyzed here were already "cleaned" by the data-entry program, which may lead to underestimation of the benefits of computerized data entry (although the detectable errors examined were those that the program did not check).

12. While the per capita expenditure variable is arguably the most important variable, the VHLSS measures many other variables; the authors have not (yet) attempted to investigate other variables as intensively.

This last point suggests that more research of this type is needed. Given the cost of purchasing hundreds of computers, randomized experiments similar to this one should be conducted in any country considering decentralized data entry. For Vietnam this study contributed to the decision not to provide computers to each district. Research is also needed to determine whether providing computers reduces bias; this would require reinterviewing households or comparing other data, such as employer records, with household survey data. Finally, when important policy decisions are made using one or two variables from a household survey, such as per capita expenditures, whenever a major change in data collection methods is made a randomized trial should be conducted with a sample size sufficient to detect changes in the distribution of those variables that have substantial policy implications.

Funding

The funding to provide the computers in this experiment was from the World Bank Office in Hanoi, using funds from the U.K. Department for International Development (DFID).

References

Brown, M., and A. Forythe. 1974. "Robust Test for the Inequality of Variances." *Journal of the American Statistical Association* 69:364–67.

Chaudhuri, Shubham, and Martin Ravallion. 1994. "How Well Do Static Indicators Identify the Chronically Poor?" *Journal of Public Economics* 53(3):367–94.

Gibson, John, Jikun Huang, and Scott Rozelle. 2003. "Improving Estimates of Inequality and Poverty from Urban China's Household Income and Expenditure Survey." *Review of Income and Wealth* 49(1):53–68.

GSO (Government Statistical Office). 2006. *Vietnam Household Living Standards Survey (VHLSS), 2002 and 2004: Basic Information*. Hanoi.

Grosh, Margaret, and Juan Muoz. 1996. "A Manual for Planning and Implementing the Living Standards Measurement Study Surveys." Living Standards Measurement Study Working Paper 126, World Bank, Washington, D.C.

Johnson, Norman, Samuel Kotz, and N. Balakrishnan. 1994. *Continuous Univariate Distributions*. Vol. 1, 2nd edn. New York: John Wiley.

Ravallion, Martin. 1994. "Poverty Rankings Using Noisy Data on Living Standards." *Economics Letters* 45(4):481–85.

Truong, Thi Kiem Chuyen. 2003. "Assessment Report on Pilot In-Field Data Entry." Department of Geography, University of Social Sciences and Humanities. Ho Chi Minh City, Vietnam.