Comparing a Direct with an Indirect Approach to Collecting Household Level Data: Who tells the truth?

Karen Macours

University of California at Berkeley

Version: October 29, 2002

Paper prepared for presentation at the NEUDC – Williamstown, October 25-27, 2002.

Contact information:
Karen Macours
Department of Agricultural and Resource Economics
Giannini Hall, 207, #3310
University of California
Berkeley, CA 94720
Tel: (510) 643 0846 Fax: (510) 643 8911
Email: macours@are.berkeley.edu

Comparing a Direct with an Indirect Approach to Collecting Household Level Data: Who tells the truth?

1. Introduction

This paper aims to validate an indirect survey methodology that was designed to collect census data in a cost-effective way. The methodology is particularly cost and time efficient when information on a limited set of characteristics needs to be collected for a large number of observations. The indirect approach relies on the fact that a lot of private information is public at the level of the community. Hence a selected key informant from the community can be used to answer questions about individual community members on matters that are locally public. The use of key informants not only allows obtaining data on a large number of households in a relatively short time, but also prevents problems of non-response or non-inclusion of certain types of households which is important when one wants to characterize the complete universe.

The use of key informants to obtain information about other households in a community is most common in Rapid Rural Appraisal (RRA) exercises. While for RRA key informants are usually asked to provide rankings of households, in a recent paper Takasaki et al. (2000) explore the possibility of collecting data on asset ownership from key informants. The approach discussed in this paper is related to theirs. However indirect information was obtained on a much wider variety of household and plot characteristics.

The use of key informants to obtain household level information also offers an alternative for data collection on topics for which refusal to answer or misreporting by household heads might be a concern. Typically, the head (or other members) of a household interviewed for a survey, might be reluctant to provide information on certain topics, or might purposely misrepresent their own situation, expecting potential future benefits or attempting to avoid negative consequences by doing so. Such strategic answering might be related to certain household characteristics that could be of interest in the analysis and hence might cause important biases. The possibility of strategic answering, as well as the potential reluctance to cooperate with the survey, is especially high for surveys on sensitive topics, such as land related surveys in most of Latin America, with its long history of latent and sometimes violent land conflicts.

The use of key informants also reduces the risk of misinterpretation of the questions and facilitates consistency of information across household as sufficient time is spent with the informant and the same questions are repeated each time. In addition, consistency in the reporting of a transaction for the two households in the transaction can be checked and enforced during the survey process, which is crucial when one is concerned with analyzing matching patterns.

The disadvantages of the indirect approach come from the strong reliance on a relatively small number of key informants and from the limitations imposed by which information is local public knowledge in a particular context. Detailed data on income or expenditures for instance can

probably not be collected.[1] The quality of the data obtained with the indirect survey approach obviously depends on the level of knowledge and the errors made by the key informant regarding the household characteristics of interest. If certain characteristics of the informants or of the communities lead to systematic under- or overreporting, this could lead to important biases.

This paper compares data collected using a standard household survey and data collected on the same households using the indirect survey methodology, and analyses the differences and their potential sources. In the next sections, the related literature is reviewed and the data are discussed in section 3. In section 4, we compare the direct and indirect measures of a set of variables, and discuss the characteristics of the differences. Section 5 analyzes the determinants of these differences and section 6 concludes.

2. Literature Review

2. 1. Reliability and validity of Rapid Rural Appraisal data

In Rapid Rural Appraisal exercises, informants are typically asked to rate the different households according to their standard of living or food security status. The motivations to use key informants in this context - as in ours - are (1) the opportunity of exploiting local knowledge and reducing the risk of omitting key characteristics (Adams et al., 1997); and (2) the fact that respondents in household surveys are often reticent to disclose specific details regarding their socioeconomic status or may be predisposed to providing "desired" answers in order to please the interviewer or to satisfy perceived self-interest (Guijit, 1992).

A number of recent studies have investigated the reliability and validity of the group ratings from RRA exercises, by comparing data obtained using different survey methodologies about the same households. Adams, et al. (1997) show that wealth group ratings obtained by key informants in rural Bangladesh reflect statistical differences in health, demographic and socio-economic variables collected through direct household surveys. Bergeron, et al. (1998) compare food security ratings obtained from different groups of key informants on the same households in rural Honduras. They find low levels of agreement and derive a number of important points and hypothesis for the use of key informants: different informants might have different info about the households they are rating, different informants might have different understanding of the criteria used, groups may be dominated by one or more individual, poor informant selection introduces error in the ratings, differences at the margin are hard to rate and the training of the enumerators is crucial.

Takasaki et al. (2000) compare wealth rankings and asset ownership obtained from a group of village informants with data obtained by surveying the households themselves in the Peruvian Amazon. They show accuracy rates of over 80% for most productive assets, and 62% for consumer durables, and find some evidence of attribution error by the informants. They also find that the accuracy of the actual wealth ranking obtained from the key informants is rather high,

[1] It is important to note however that the use of qualitative indicators, such as living standard rankings, might eliminate the need for quantitative information in certain contexts.

and indicate that it might be a good indicator for the stratification of a further sample for household surveys.

In summary, the RRA literature provides some evidence on the reliability of wealth and asset rankings and ratings. However, it also points to the importance of the choice of variables and informants.

2.2. Comparison of different measures of education and labor market outcomes

Ashenfelter and Krueger (1994) compare the answers of identical twins regarding their own education, their twin sibling's education and their parents' education. They find that the correlation between the reports of the twins on their education, are between 0.87 and 0.92. The correlation of the reports on the education of the parents is between 0.74 and 0.85. They interpret these correlations as a reliability ratio and conclude that the economic return to schooling must be 10% downwardly biased (given a reliability index of 0.90).

A series of articles in the labor economics literature compares data from different sources on the same variables, usually comparing data from payroll tax records to information as reported by individuals in income surveys. Bound and Krueger (1991) and Bound et al. (1994) assume that the payroll data do not have measurement error, and hence use the difference between the two measures to analyze the characteristics of the distribution of the error from the survey responses. In a related paper, Rodgers et al. (1993) report rather low correlations between the reported and the recorded data, ranging from 0.24 for dollars/hour in a usual time period to 0.79 for annual pay of earnings.

Barron, et al. (1997) compare responses of employers and employees to identical questions regarding different types of training. Also in their dataset, the correlations between the two measures are low (between 0.17 and 0.47). They allow for measurement error on both sides, but assume that it is random and analyze the difference. They show that these differences seem not to correlated with other employee characteristics, and conclude that the measurement errors are unlikely to cause bias, other than attenuation bias, in wage regressions.

Hence, validation studies in the labor literature often show low levels of correlation between 2 measures of the same variable collected from different respondents. An important conclusion of these studies is that random measurement error alone is often enough to explain the differences.

3. Data

The data analyzed in this paper were collected in the context of a study aimed at analyzing matching between landlords and tenants in the land rental markets in the Dominican Republic (Macours, 2002). The indirect survey approach was used to obtain data on each household and each plot within 30 communities.

In a first step, basic information about all households and all plots in a community was obtained in order to (1) define the complete land rental market, (2) match landlords with their respective tenants and obtain information about the partners on both sides of the transaction, and (3) obtain a sampling frame for more detailed household and plot level questions. In order to ensure the

inclusion of all households in the community, the informant was asked to draw a map of the community.[2] In a second step, a stratified sample of households was drawn in order to oversample the landlords and tenants in the population. This was necessary to guarantee sufficient observations on the variables of interest, as rental in some communities is a rare event. All landlords and tenants were selected, complemented by a random sample of all other households in the community. In addition to household-level information, data on all the plots these households owned (either owner-cultivated or rented out) and rented was obtained from the key informant. Data on community characteristics were also collected.

In addition, data were collected on 8 households in each community, through direct interviews with the household heads. A stratified random sample of households in each community was drawn, to represent different standards of living, and different positions in the land rental market (landless, landlord, autarkic landowner, landlord). These 8 households per community are the basis for the analysis in this paper.

Data were collected in the regions of Constanza, San Francisco de Macoris and Jacagua (80 households in each region). Constanza is located in a fertile valley in the mountainous area and is characterized by a very intensive irrigated horticultural production, which depends to a large extent on hired labor. San Francisco de Macoris is located in the flatlands and agricultural production is mainly rice, complemented by plantains and pastures. Land in this region was redistributed during the land reform and land conflicts are more common here than in the other regions. Jacagua is located in a mountainous area and agricultural production is extensive tree production combined with subsistence crops. In the three regions, agricultural income is an important component of households' budgets. In Constanza and San Francisco de Macoris houses are concentrated in a nucleus along the road, while in Jacagua distances between houses tend to be larger. Recent immigration into all communities is relatively small (less than 20 % of all households in the last 10 years).

The informants were selected by the enumerator explicitly for their knowledge on the issues covered in the survey. Often they are leading figures in the community (about 60% are actual leaders of one of the community organizations), and have lived for a long time in the community. The number of households that one informant reports on varies between 35 and 148. In case of large communities, the community was divided in different sections and information on households in the different sections was obtained from different informants. The division was made according to the indications of the informants themselves as to which households they had accurate information about.

[2] The use of a map prevents reliance on existing administrative lists that might be out-dated or incomplete, and might be biased against the poorer households (see for evidence e.g. Christiaensen, et al. (2001) who compare administrative lists with lists obtained from mapping for rural communities in Mali).

4. Comparison of the Direct and Indirect Measures and Analysis of the Difference

In order to compare the different key variables collected with the two alternative methods, table 1 shows the percentage of observations for which the direct and indirect measures match.[3] Overall the level of agreement seems fairly high, especially concerning the status in the land market, physical and human capital. There is less consistency between the two measures regarding the number of household members, membership in community organization and in particular the total amount of land owned. However, the correlation coefficient for land owned in the second column is rather high. The lower correlation coefficients for the amount of land rented out and rented in, and for cattle ownership is mainly due to a few outliers. Overall the level of agreement and correlation is relatively high when compared to the validation studies discussed in section 2.

Scatterplots (figures 1) of the answers by the two respondents provide further insight into the relationships between the direct and the indirect measures. The straight line indicates the 45-degree line. Figure 1a shows the scatterplot for land ownership, land rented, land rented out and self-cultivated land. The figure for land rented out shows a first indication of consistent differences between the direct and the indirect survey as several household heads report no land being rented out, when the informant does. A similar, but less obvious pattern can be seen for land owned. In the Dominican Republic, as in most Latin American countries that went through a phase of expropriative land reform, households might be reluctant to give information about renting out land, as it can be seen as a signal for having "too much" land. In addition, for land acquired through the land reform, renting out land is, although widely practiced, not officially allowed without specific permission.

Table 2 reports a number of characteristics of the distributions of the differences between the direct and indirect measures for all variables. First of all, the mean difference for most variables is relatively small, compared to the value of the variable itself.[4] It is however large and statistically significant from zero for the dichotomous variable indicating renting out (landlord), indicating more reporting of renting out by the informants than by the household heads. On the other hand, the difference is significantly negative for the variables indicating the number of household members and the education of the household head, indicating lower reporting of these variables in the indirect approach. The mean difference is also relatively large, although not significant for cattle and machinery ownership, and for the amount of land rented out.[5]

Similar results are found when considering the median of the difference, conditional on that difference not being equal to zero. The differences for the dichotomous variable for renting out, the amount of land rented out, the variables indicating the number of household members and the education of the household head are all found to have a median that is statistically different than zero. In addition, also the median difference for the age of the household head and for

[3] A match is defined as a difference of less than 10% for continuous variables. For the other variables, only when the indirect and the direct measures are exactly the same, are they considered to match.

[4] The means of the indirect and the direct measures are reported in table 1.

[5] In fact, when two large outliers on land ownership are excluded, the difference for total amount of land owned and for the amount rented out is very significantly positive.

machinery ownership is statistically different from zero, and for livestock ownership and total amount of land owned it is different from 0 at the 10 % level.

In the next column we report the results of a Wilcoxon rank test, testing whether the distribution of the difference is symmetric around 0, which we can reject for the same set of variables. We also reject the normality of the distribution for all continuous variables. This is similar to the results found in Bound and Krueger (1991), and can be explained by the tails being thicker than with a normal distribution (partly because of some extreme outliers), or alternatively, by the large spike near zero.

Considering all the tests in table 1 and 2 together, we distinguish three groups of variables. A first group for which the match between the direct and indirect is relatively high (more than 75%), and for which we could not reject that the mean and the median difference between the direct and indirect was significantly different from zero and could not reject that the distribution of the differences is symmetric around zero. This category contains the variables indicating ownership and tenant status, the number of household members living abroad, having a female household head, being a leader of a community organization or participating in a collective initiative, and the household head being a farmer, having an off-farm occupation, and at least one household member having an off-farm occupation. We conclude that the measurement error for these variables is relatively small in both direct and indirect, and non-systematic.

A second group for which none of the tests in table 2 could reject the nul hypothesis, but for which the match between the direct and indirect measures was lower than 75%. This group includes the variables measuring membership in an organization, the amount of land owned, rented, or owner cultivated, title ownership and the share of land with title, and the number of household member with an off-farm occupation. For this group, measurement error in the direct, indirect, or in both, must be relatively large.

The third group of variables include these for which at least one of the tests in table 2 rejected the nul hypothesis, and includes the variables measuring landlord status, livestock and machinery ownership, age and the education of the household head, number of household member and the amount of land rented out. The rejections indicate bias that might be coming either from the direct or the indirect survey.

5. Determinants of the differences between direct and indirect data

5.1. Level of disagreement

We next turn to analyzing the determinants of the level of disagreement, for those variables for which the tests in table 1 and 2 indicate that there are important differences between the direct and the indirect approach, i.e. for the variables of the second and third group distinguished in section 4.

We analyze whether community or informant characteristics can explain the differences in the level of disagreement, by regressing the absolute value of the differences on a number of key community and informant characteristics. Less private information might be publicly known in

communities with certain characteristics, such as communities that are less united, with more conflicts, more households, a higher share of new households, more people working in off-farm employment and communities that are closer to urban areas. In addition, certain characteristics of the informant might be correlated with his knowledge about the different households, such as his education, age and leadership in the community.

The results of these regressions in table 3 show that informant characteristics affect the level of the disagreement for most of the variables in group 2, but not for the variables in group 3, except for livestock. In particular, differences between the direct and indirect approach for the amount of land owned and owner cultivated are smaller when the informant is a leader, and differences for land rented in and livestock are smaller when informants are older, but not too old. Small differences for the number of household members with off-farm employment and for the education of the household head are found when the informants have finished primary education. These results suggest that the selection of better-educated and somewhat older informants that are leaders of the community, helps to obtain more accurate information.

Considering the community characteristics, we find that the occurrence of recent conflicts in the community reduces the difference between the reports on land owned and land rented out. While counterintuitive because conflicts might lead to less availability of public knowledge, a possible explanation could be that facts about land ownership and renting are more important in communities with recent conflicts, and therefore better known. The amount of land rented out seems to be less well-known in communities with more recent migrants, while communities with more off-farm employment tend to have larger differences for livestock. The puzzling positive effect of communities being united on the difference in the number of household members might be because it might be harder to distinguish who belongs to exactly which household when mutual transfers are high. However, given that few community variables are significant at all, the few significant signs should be interpreted with caution. Overall, the results suggest that community characteristics do not have a strong effect on the levels of disagreement.[6]

5.2. Direction of disagreement

To further analyze possible reasons for the disagreement between respondents, we know turn to analyzing the difference itself. To explore potential biases resulting from the indirect approach, we want to test whether the difference between direct and indirect measures are related to attribution error by the key informant. The informant might attribute certain characteristics to a household, based on his knowledge of other characteristics of that household. We would expect that attribution error will occur for less visible assets or characteristics, and the informant will attribute these to the more visible ones. For instance, if the informant does not know (exactly) how much cattle a certain household owns, he is likely to guess a certain number according to the general wealth status of that household.

Misreporting by the household head is another potential reason for the difference between the direct and indirect measures. Strategic answering is one reason why household heads with certain characteristics might under- or overreport answers to sensitive questions. We hypothesize that

[6] For the interpretation of the results, it is important to keep in mind that lack of more significant results might be due to the fact that the communities are all relatively small communities with little immigration, which augments the likelihood of private information being locally sufficiently public.

strategic answering error will occur for variables reflecting wealth and asset ownership, by household heads with lower levels of education and/or less contact with the outside world, as they are likely to be more suspicious about the aim of a survey asking them about their asset ownership. They might think that the answers will either lead to potential future benefits, which will induce them to underreport asset and wealth related variables, or, especially in case of questions about land, think it is for tax - or even worse, expropriation - purposes.

If we would know the true value of the different variables, we would be able to test these hypotheses regarding attribution error and strategic answering in a straightforward way. However, since we do not have the true values, we need to make some additional assumptions.

Let

$$x_{hi}^{d} = x_{hi}^{*} + u_{hi}^{d} \quad (1)$$

$$x_{hi}^{in} = x_{hi}^{*} + u_{hi}^{in} \quad (2)$$

with x_{hi}^{d} a household characteristic of household i, as measured in the direct household survey; x_{hi}^{in} the same household characteristic of household i, but as measured with the indirect approach; x_{hi}^{*} the true value of that characteristic, u_{hi}^{d} the error term associated with the direct measure; and u_{hi}^{in} the error term associated with the indirect measure. If we rewrite (1) and (2) to allow for the different types of error in the direct and indirect measures we get

$$x_{hi}^{d} = x_{hi}^{*} + \gamma^{d} X_{hi}^{*} + \varepsilon_{hi}^{d} \quad (3)$$

$$x_{hi}^{in} = x_{hi}^{*} + \gamma^{in} X_{hi}^{in} + \varepsilon_{hi}^{in} \quad (4)$$

with X_{hi}^{*} a vector of household characteristics that might reduce strategic answering, X_{hi}^{in} a vector of household characteristics to which other characteristics might be attributed, as measured in the indirect approach, γ^{d} reflecting strategic answering error γ^{in} reflecting attribution error and ε_{hi}^{d} and ε_{hi}^{in} random measurement error.
Taking the difference between (4) and (3) we get

$$\Delta x_{hi} = x_{hi}^{in} - x_{hi}^{d} \text{ or} \quad (5)$$

$$\Delta x_{hi} = \gamma^{in} X_{hi}^{in} - \gamma^{d} X_{hi}^{*} + \varepsilon_{hi}^{in} - \varepsilon_{hi}^{d} \quad (6)$$

Estimating equation (6) would require observing the true values of the household characteristics that are hypothesized to lead to strategic answering. It seems reasonable to assume that the error of the household head's reports on their education, membership and/or leadership in community organizations, and their occupation, is random. If so, replacing the true values with the direct measures of these variables, will not induce bias in the results, other than attenuation bias. Hence while we should be careful in interpreting the results, the finding of coefficients significantly different from zero for the direct measures does suggest evidence of strategic answering. Given

that the attenuation bias might also cause bias in the estimations of the other variables, we test for the robustness of the results in specifications with different sets of right hand side variables.[7]

Hence the equations we estimate can be written as:

$$\Delta x_{hi} = \gamma^{in} X_{hi}^{in} - \gamma^{d} X_{hi}^{d} + \eta_{hi} \quad (7)$$

$$\text{with } \eta_{hi} = \varepsilon_{hi}^{in} - \varepsilon_{hi}^{d} - \gamma^{d} U_{hi}^{d} \quad (8)$$

X_{hi}^{in} containing the indirect measures of education of the household head (dichotomous variable indicating primary education completed), membership and leadership in a community organization, non-farm occupation by at least one household member, living standard, and a dummy for female household head;
X_{hi}^{d} containing the direct measures of the education of household head (dichotomous variable indicating primary education completed), membership and leadership in a community organization and non-farm occupation by at least one household member;
U_{hi}^{d} containing the difference between the direct and the true vale of education of household head, membership and leadership in organization, and non-farm occupation;
Given that our findings in table 3 suggest that the variance of the error might vary according to informant or community characteristics, we allow for heteroscedasticity in the estimation and use the Huber-White estimator of variance.

The dependent variables for which this equation is estimated is the difference between the direct and the indirect measures of different types of variables of group 2 and 3 (as defined in section 4). Note that we can distinguish between errors related to under (or over) reporting by the household head, and attribution error by the informant, because for the later, the characteristics of one variable as reported by the informant, is correlated with the characteristics of other variables as reported by the informant, and not with those variables as reported by the household head.

5.3. Results

Equation (7) was estimated using ordered probit regressions for differences in dichotomous variables and OLS for the others. Table 4 reports only the significant signs of the whole set of regressions. The gray areas show which variables were not included in the different regressions. To control for regional heterogeneity, regional fixed effects were added. Note that the signs of the household characteristics as measured by the direct survey, reflect $-\gamma^{d}$. A negative sign should hence be interpreted as decreasing the likelihood of underreporting.

For each dependent variable, the first line reports the results with the set of independent variables discussed in the previous section. The second and third line report the results for the regressions with only the indirect and direct measures of the independent variables respectively. The comparison of these results with the regression in the first line, allows one to determine whether

[7] Deaton (1997) shows that the attenuation bias depends on the variance-covariance matrix of the measurement errors in the different independent variables.

results in the first regression are caused by multicollinearity between the direct and indirect measures. The fourth line reports the result of a regression that was added as a further robustness check. It includes the direct variables, and in addition the measures of living standard and female household variables. The living standard variable is a categorical variable indicating how the informant ranked the household in living standard. Because of the ranking nature of this variable, no similar measure was obtained from the household directly. No distinction could be made between the direct and indirect measures for the dummy for female household head, as they are the same for all but three households. Hence, although these variables are indirect measures, it is not possible to determine whether in fact they capture reporting errors by the informant or by the household head, because we cannot control for equivalent direct measures.

The regressions show evidence of errors from reporting by the household head on a number of variables. The results suggest that heads of households with at least one member with an off-farm occupation are less likely to underreport being a landlord (i.e. renting out land), livestock holdings and ownership of a title of the land. The share of land with title is also less likely to be underreported by leaders in community organizations. Membership in a community organization is correlated with less overreporting of age, and less underreporting of the number of female household members. These results are consistent with the hypothesis that household heads that are more exposed to outside contacts, are less likely to underreport, and could indicate strategic answering behavior. The results show less evidence for attribution errors by the informants, although it seems to play a role for a few variables. Informants tend to overreport education levels of leaders, and underreport membership in organizations for households with off-farm occupations.

Household with higher living standards have a larger difference between the indirect and direct measures for livestock and a smaller difference for the amount of land rented out. Our regressions do not allow identifying whether this comes from misreporting in the direct or indirect survey. However, one possible interpretation for the difference in land rented out is that poorer households might be more likely to underreport. This would be consistent with the finding that households with off-farm occupations are less likely to underreport renting out land. Finally, the difference for the number of household members seems to be correlated with the direct and the indirect measure of off-farm occupation hence no firm conclusion can be drawn from this. For other variables, the findings are not robust across the different specifications.

In summary, for the variables in group 3, for which the distribution of the difference suggested bias, we find that for the landlord, age, and the amount of land rented out, the bias results from misreporting by the household members in the direct survey. On the other hand, our results suggest that the source of the bias for education is attribution error by the informant. Bias for livestock and for the number of household members seems to originate in both the direct and indirect measures.

6. Conclusion

In this paper, we compare data collected through a direct household survey, with data collected on the same households with an indirect survey methodology that relies on local public knowledge and key informants. We show that the overall level of agreement between the two types of respondents is relatively large, but that there are important differences among the variables. The results suggest that selection of better-educated, somewhat older informants that are leaders in the community as key informants can help to reduce random measurement error. We find some evidence of attribution error by the informants, particularly on less visible household characteristics, such as education. This finding points to the need of careful selection of variables in each context.

The results in the paper furthermore provide evidence of systematic under- and overreporting by household heads that are less exposed to outside contacts. This result points to important potential biases in direct household surveys. This paper introduces an alternative method to collecting household level information when such biases might be off concern. More importantly, the proposed methodology that allows collecting information on a large number of observations in a relatively short time is shown to provide reliable data on information that is local public knowledge.

References

Adams, Alayne M., Timothy G. Evans, Rafi Mohammed and Jennifer Farnsworth, 1997, "Socioeconomic Stratification by Wealth Ranking: Is it Valid?", *World Development*, 25(7): 1165-1172.

Ashenfelter, Orley and Alan Krueger, 1994, "Estimates of the Economic Return to Schooling from a New Sample of Twins", *The American Economic Review*, 84(5): 1157-1173.

Barron, John M., Mark C. Berger, and Dan A. Black, 1997, "How Well Do We Measure Training", *Journal of Labor Economics*, 15(3): 507-528.

Bergeron, Gilles, Saul Sutkover Morris and Juan Manuel Medina Banegas, 1998, "How Reliable are Group Informant Ratings? A Test of Food Security Ratings in Honduras", *World Development*, 26(10): 1893-1902.

Bound, John and Alan B. Krueger, 1991, "The Extent of Measurement Error in Longitudinal Earnings Data: Do Two Wrongs make a Right?", *Journal of Labor Economics*, 9(1): 1-24.

Bound, John, Charles Brown, Greg J. Duncan and Willard L. Rodgers, 1994, "Evidence of the Validity of Cross-Sectional and Longitudinal Labor Market Data", *Journal of Labor Economics*, 12(3): 345-368.

Christiaensen Luc, John Hoddinott and Gilles Bergeron, 2001, "Comparing Village Characteristics Derived from Rapid Appraisals and Household Surveys: A Tale from Northern Mali", *Journal of Development Studies*, 37(3):1-20.

Deaton, Angus, 1997, *The Analysis of Household Surveys. A Microeconometric Approach to Development Policy*, Johns Hopkins University Press: Baltimore and London.

Gijit, I., 1992, "The Elusive Poor: A Wealth of Ways to Find Them", *RRA Notes*, (15): 7-13.

Macours, Karen, 2002, "Insecurity of Property Rights and Matching in the Land Rental Market", University of California at Berkeley, mimeo.

Rodgers, Wilard, L., Charles Brown and Greg J. Duncan, 1993, "Errors in Survey Reports of Earnings, Hours Worked and Hourly Wages", *Journal of the American Statistical Society*, 88(424): 1208-1218.

Takasaki, Yoshito, Bradford L. Barham and Oliver T. Coomes, 2000, "Rapid Rural Appraisal in Humid Tropical Forests: An Asset Possession-based Approach and Validation Methods for Wealth Assessment among Forest Peasant Households", *World Development*, 28(11): 1961-77.

Table 1: Comparison between measures recorded by direct and indirect approach

	# obs	mean indirect	mean direct	% match*	Correlation
Status in land market					
Owner	240	0.57	0.53	80	
Landlord	240	0.24	0.14	84	
Tenant	240	0.30	0.30	83	
Physical capital					
Number hh members living abroad	238	0.23	0.24	92	73
Livestock	238	1.69	0.93	80	35
Machinery	240	0.10	0.15	88	
Human capital					
Age hh head	240	0.12	0.13	58	89
Female hh head	240	51.09	51.17	98	
Primary education hh	240	0.57	0.66	78	
Number of household members					
All	240	2.37	2.67	63	58
Male	240	1.34	1.54	70	58
Female	240	1.03	1.13	76	59
Social capital					
Member of organization	238	0.42	0.44	62	
Leader	239	0.16	0.18	81	
Participant collective initiative	239	0.44	0.41	77	
Amount of land					
Owned	238	34.46	37.12	48	65
Rented out	238	9.95	6.76	78	49
Rented in	239	14.47	14.65	67	74
Owner cultivated	239	24.40	30.23	55	69
Title					
Title ownership	125	0.62	0.62	64	
Share titled	125	0.60	0.59	59	59
Occupation					
Farmer	240	0.60	0.65	79	
Off farm occupation head	240	0.25	0.22	85	
hh member with off farm occupation	240	0.43	0.48	78	
# hh members with off-farm occupation	240	0.38	0.42	73	51

* The match reflect the share of all observations for which the direct and indirect differ less than 10%
hh stands for household

Table 2: Characterizing the difference between the indirect and direct measures[a]

	Mean	Weighted St. Dev*	P-value Test mean=0	P-value Test median =0**	P-value Test distr. symmetric around 0 **
Status in land market					
Owner	0.0375	0.8065	0.4092	0.2430	0.1893
Landlord	0.1042	2.0580	0.0036	0.0001	0.0001
Tenant	-0.0083	1.3970	0.8421	0.8776	0.7576
Physical capital					
Number hh members living abroad	-0.0126	2.8263	0.7682	0.4807	0.3391
Livestock	0.7605	5.5250	0.1063	0.0789	0.0501
Machinery	-0.0500	2.6218	0.1024	0.0357	0.0233
Human capital					
Female hh head	-0.0083	1.0683	0.7794	0.6250	0.3173
Age hh head	-0.0833	0.1564	0.8719	0.0181	0.1911
Primary education hh	-0.0917	0.7563	0.0389	0.0038	0.0028
Number of household members					
All	-0.3042	0.4349	0.0000	0.0002	0.0001
Male	-0.2000	0.5555	0.0001	0.0003	0.0002
Female	-0.1042	0.5625	0.0085	0.0163	0.0111
Social capital					
Member of organization	-0.0168	1.4435	0.7115	0.7520	0.6733
Leader	-0.0251	2.5637	0.4541	0.4614	0.3763
Participant collective initiative	-0.0251	1.1487	0.5524	0.5044	0.4227
Amount of land					
Owned	-2.6618	3.6321	0.7524	0.0848	0.0645
Rented out	3.1870	3.9584	0.1386	0.0062	0.0040
Rented in	-0.1736	2.5219	0.9418	0.3742	0.3005
Owner cultivated	-5.8243	4.3367	0.4479	0.7163	0.7403
Title					
Title ownership	0.0000	0.7433	1.0000	1.0000	1.0000
share titled	0.0091	0.7175	0.8108	1.0000	0.8748
Occupation					
Farmer	-0.0417	0.7288	0.3458	0.2026	0.1573
Off farm occupation head	0.0292	1.6211	0.4514	0.3105	0.2367
hh member with off farm occupation	-0.0458	1.0367	0.3131	0.1690	0.1308
# hh members with off-farm occupation	-0.0375	1.5652	0.3569	0.5386	0.4447

[a]: hh stands for household

* Weighted standard deviation equals $\frac{St.Dev}{(\bar{x}^{in} + \bar{x}^{d})/2}$

** Conditional on the difference not being equal to zero (Wilcoxon rank test)

Table 3: Significant results of tobit and probit regressions on the absolute value of the difference between the indirect and direct measures[a]

	Informant characteristics				**Community characteristics**					
	primary education	leader	age	age2	united	conflict	total # hh	% new hh	# off-farm employ.	distance to town
Group 2 variables										
Member of organization										
Amount of land owned		-***				-**				
Amount of land rented in			-**	+**				+**		
Amount of land owner cultivated		-***				-*				
Title ownership										
Share of land with title		-*								
# hh members with off-farm occupation	-**									
Group 3 variables										
Landlord										
Livestock			-**	+**					+**	
Machinery						+*				
Age hh head										
Primary education hh head	-*									
# all household members					+**					
# male household members										
# female household members										+*
Amount of land rented out						-**				

[a] Regional fixed effects not reported; hh stands for household.

Table 4: Significant signs from OLS and ordered probit regressions on the difference between direct and indirect measures (robust standard errors)[a]

		Errors from reporting by household head DIRECT independent household head variables				**Attribution error** INDIRECT independent household head variables				**Undefined error**	
Dep. Variable Difference in:		Primary education	Member org	Leader	Off-farm occup	Primary education	Member org	Leader	Off-farm occup	Living standard	Female hh head
Landlord (rents out)	(1)				-**						
	(2)										
	(3)				-**						
	(4)				-**						
Livestock	(1)	-*			-**		-**			+**	
	(2)						-**			+**	
	(3)				-**						
	(4)	-*			-**					+**	
Machinery	(1)					-**				+*	
	(2)									+**	
	(3)										
	(4)										
Age hh head	(1)		+***				-*				
	(2)										
	(3)		+**								
	(4)		+**								
Primary education hh head	(1)				-*			+*	+**		
	(2)							+**			
	(3)										
	(4)										
# all household member	(1)				-**				+**		
	(2)								+*		
	(3)				-*						
	(4)				-*						
# male household member	(1)				-**				+**	-*	
	(2)								+*	-**	
	(3)				-**						
	(4)				-**						
# female household member	(1)		-**								
	(2)										
	(3)		-**	+*							
	(4)		-**	+*							

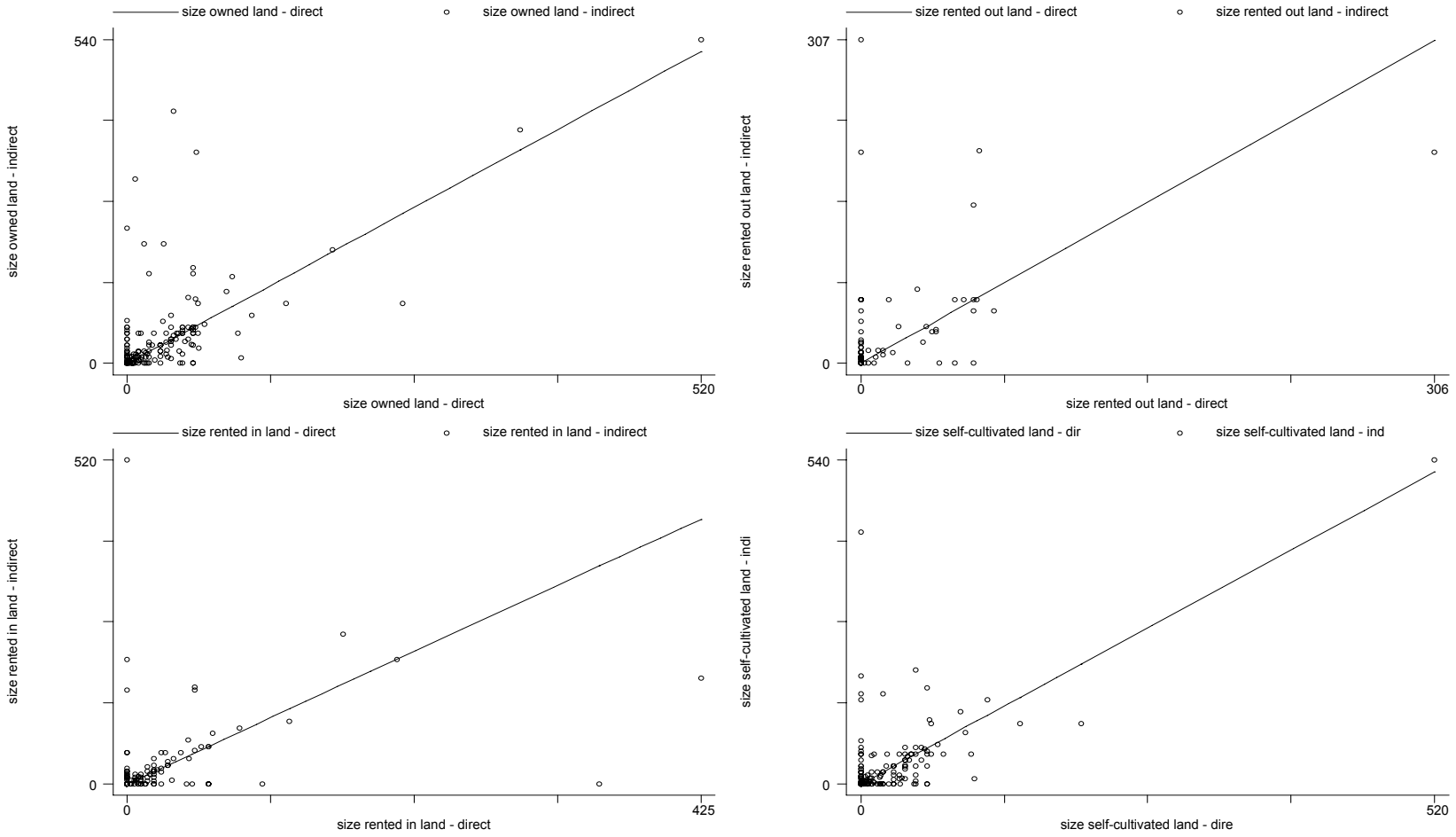
[a]: Shaded areas indicate the variables that are not included

Table 4 continued

		Errors from reporting by household head				**Attribution error:**				**Undefined error:**	
		Direct independent household head variables				Indirect independent household head variables					
Dep. Variable		educ.	member org	leader	nonfarm occup	educ	member org	leader	nonfarm occup	Living standard	Female hh
Member of organization	(1)								-*		
	(2)								-*		
	(3)										
	(4)										
Amount of land owned	(1)										
	(2)										
	(3)				-*						
	(4)				-*						
Amount of land rented out	(1)									-*	
	(2)									-**	
	(3)										
	(4)									-*	
Amount of land rented in	(1)										
	(2)										
	(3)										
	(4)										
Amount of land owner-cultivated	(1)										
	(2)										
	(3)				-*						
	(4)				-*						
Share of land with title	(1)			-*	-**						
	(2)								-*	+*	
	(3)			-**	-***						
	(4)			-**	-***						
Title ownership	(1)				-**						
	(2)									+*	
	(3)			-*	-**						
	(4)			-*	-**						
# hh members with off-farm occup.	(1)										
	(2)										
	(3)										
	(4)										

[a]: Shaded areas indicate the variables that are not included

Figure 1a: Scatterplots of land size (tareas)*



* Excluding 2 largest outliers for presentation purposes

Figure 1b: Scatterplots of labor endowments (with circles of observations indicating number of observations)

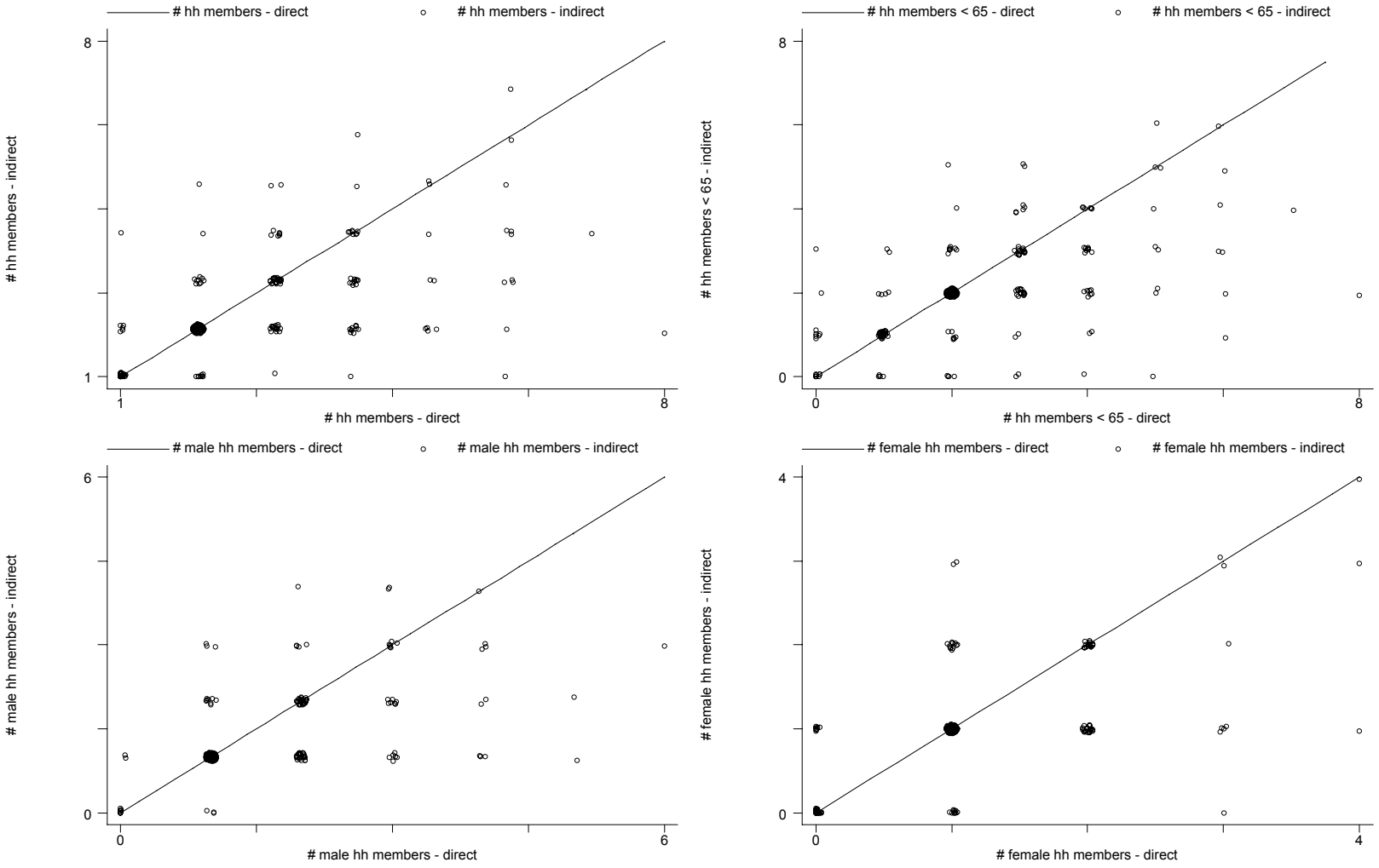


Figure 1c: Scatterplots of asset endowment (with circles of observations indicating number of observations)