Policy Research Working Paper 9410

Gender Bias in Agricultural Child Labor

Evidence from Survey Design Experiments

Jose Galdo
Ana C. Dammert
Degnet Abebaw

WORLD BANK GROUP
Development Economics
Knowledge and Strategy Team
September 2020

Policy Research Working Paper 9410

Abstract

Agricultural labor accounts for the largest share of child labor worldwide. Yet, measurement of farm labor statistics is challenging due to its inherent seasonality, variable and irregular work schedules, and the varying saliences of individuals' work activities. The problem is further complicated by the presence of widespread gender stratification of work and social lives. This study reports the findings of three randomized survey design interventions over the agricultural coffee calendar in rural Ethiopia to address whether response by proxy rather than self-report has effects on the measurement of child labor statistics within and across seasons. While the estimates do not report differences for boys across all seasons, the analysis shows sizable self/proxy discrepancies in child labor statistics for girls. Overall, the results highlight concerns on the use of survey proxy respondents in agricultural labor, particularly for girls. The main findings have important implications for policymakers about data collection in rural areas in developing countries.

This paper is a product of the Knowledge and Strategy Team, Development Economics. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at http://www.worldbank.org/prwp. The authors may be contacted at galdo@carleton.ca, ana_dammert@carleton.ca, and degnet06@yahoo.com.

Gender Bias in Agricultural Child Labor: Evidence from Survey Design Experiments

Jose Galdo **Ana C. Dammert*** **Degnet Abebaw**

JEL-codes: C8, J22, 012, Q12

Keywords: Survey design, farm labor, gender, labor statistics, child labor

* Jose Galdo is an Associate Professor at Carleton University, Ottawa; (e-mail: jose_galdo@carleton.ca). Ana C. Dammert (corresponding author) is an Associate Professor at Carleton University, Ottawa (e-mail: ana_dammert@carleton.ca), Degnet Abebaw is an Independent Researcher (formerly at the Ethiopian Economic Policy Research Institute of the Ethiopian Economics Association), Addis Ababa, Ethiopia (e-mail: degnet06@yahoo.com). The authors gratefully acknowledge the financial support from IZA/DFID Growth and Labour Markets in Low Income Countries (GLM-LIC) program, grant agreement GA-C3-RA5-323. The authors would like to thank the editor and two referees, as well as seminar participants at the University of Maryland- College Park, 3ie-IFPRI Seminar, Ohio University, Dalhousie University, 2nd IZA/DFID GLM-LIC Research Network Conference, 2016 LACEA, 2017 Canadian Economics Association Meeting, 2018 ICABR Conference, and the 4th IZA/DFID GLM-LIC Research Network Conference for helpful comments and suggestions. Any errors and omissions are our own.

1. Introduction

Agricultural labor accounts for the largest share of child labor worldwide (60 percent). In Sub-Saharan Africa, the setting of this study, smallholder farming is the only source of income and employment for most households and accounts for 85 percent of child labor (ILO 2017). Yet, measurement of farm labor is a challenging topic due to its inherent seasonality, variable and irregular work schedules, multiple plots, intercropping, and varying saliences of individuals' work activities (Arthi et al. 2018), and the gender division of farm labor (Beegle et al. 2017), as culturally defined roles determine the capacity of men and women to allocate labor time across economic activities (Palacios-Lopez et al. 2017). While the measurement of child labor is essential to our understanding of the main factors that drive children to work and the development of sound policy, research on data quality and data collection methods is an overlooked topic in development economics (Beegle et al. 2012).

In this study, we implement three randomized survey design interventions over the agricultural coffee calendar in rural Ethiopia to address whether response by proxy rather than self-report has effects on the measurement of child labor statistics within and across seasons. Specifically, our experimental survey design intervention consists of the random manipulation of the survey respondent in the application of the same survey instrument to 1200 Fairtrade coffee households in three different seasons of the coffee production calendar: the *Meher* season (main rainy season), the *Belg* season (short rainy season) and the harvest season.

Over the past years, the increasing availability of multi-topic household surveys in agricultural settings has highlighted systematic challenges concerning the link between agricultural information that is commonly captured during a single visit to the household and the measurement of farm labor that is variable and irregular due to the salience of seasonal work activities (Arthi et al. 2018; Gaddis et al. 2019). One dimension of this measurement problem is

whether child labor statistics depend on to whom the survey questions are asked (Dillon et al. 2012, Dammert and Galdo 2013, Janzen 2018). While the ILO guidelines for survey design and measurement of children's work suggest that the child should answer the labor module him or herself (ILO 2008), substantial variation across household surveys exist. For instance, and although different surveys might entail different definitions of child labor, the Statistical Information and Monitoring Programme on Child Labour (SIMPOC) stand-alone child labor surveys collect labor information from both children and proxies, while the Multiple Indicator Cluster Surveys (MICS) gather labor information from the most knowledgeable adult member of the household, and UNICEF MICS surveys direct questions to the mother or primary caretaker. Furthermore, policymakers and researchers substantially rely on country-level labor force surveys (LFS), which provide in most cases either self or proxy reporting measures of child labor but not both measures due to the costs and logistics of fieldwork. A comprehensive review of LFS shows, for instance, that in half of the national-level LFS up to 50% percent of the responses are provided by proxy informants (ILO 2018). In Africa, the setting of our survey design intervention, the rate of proxy respondents ranges from 34% in Nigeria to 99% in Mali for the 15 to 24 age group (Desiere and Costa 2019).[1]

The effect of the type of respondent on child labor statistics is not clear a priori if both respondents recall different types of activities with different errors. Child-reported information may be more accurate than proxy responses if a child knows best how she allocates her time, while at the same time, recall of past activities may be cognitively burdensome if the child may not fully understand what "work" entails or how to use recall-count strategies to track her activities and hours. On the other hand, a proxy respondent may be familiar with the children's activities depending on the frequency and continuity of work activities, and the complementarity of child

and proxy activities across the agricultural production cycle. Thus, the main advantage that our survey design randomization offers is its ability to causally estimate survey design unbiased impacts of self/proxy reporting on child labor statistics and to assess potential drivers when these gaps emerge. Moreover, unlike some surveys that aim to capture two responses i.e., self - and proxy response for each child, the randomization of the survey instrument avoids potential strategic behavior where the last respondent (e.g., the child) tries to be consistent with earlier responses given by the proxy. Likewise, our controlled survey design instrument minimizes the self-selection of proxy respondents, a potential source of bias in standard survey designs.

The broader measurement error literature highlights the role of the salience of individuals and work activities, recall decay bias, and cognitive burdens of reporting when explaining recall bias relative to a known benchmark (e.g., Bound et al. 2001). Likewise, sources of measurement error in survey designs include question selection, sequencing and wording, and the data collection method. The use of a short module compared to detailed probing questions, for instance, has a statistically significant effect on child labor measurement (Dillon et al. 2012). Bias in statistics from surveys is also related to the representativeness of the sample due to non-interview rates and non-response rates. In this regard, the small literature on self/proxy reporting on labor statistics shows that self-respondents yield higher non-interview rates, while proxies yield higher item non-response rates (Biggs 1992).

Regular and predictable activities are easier to monitor and therefore recalling or counting the occurrence of events that are salient and regular is less costly for respondents. Thus, if farm labor performed by children is variable and irregular, proxy respondents are more likely to focus on information that is less prominent and perceptible, resulting in larger discrepancies of child labor reports. In the absence of recall-and-count information strategies, proxy respondents would

then rely on their general beliefs about their contextual circumstances in their search for answers to survey questions (Beegle et al. 2012). In the setting of sub-Saharan Africa, if the work of girls at the farm is more variable and has a less predictable pattern, we would expect to find that self- and proxy-reporting of child labor would be particularly prone to discrepancies. On the other hand, if farm work for boys is typically more frequent and regular, it would be easier for both proxy and boy respondents to rely on recall-and-count strategies, and thus we would expect to find less self/proxy reporting discrepancies for boys.

The seasonality of coffee production in the agricultural calendar not only affects the share of children who are working but it could also affect the extent of self/proxy reporting discrepancies. If a particular agricultural season involves child work activity with routinely steps and predictable timing, or involve the complementarity of child/adult work due to the physical strength requirements of farm activities, we would expect to observe smaller discrepancies in self/proxy reporting of child labor relative to other agricultural seasons where salient work activities do not require physical strength and, thus, can be done by children in their own. Likewise, when the labor demand is at its highest (so that the heads of households are actively busy), or at its lowest (so that the heads of households have to allocate time to outside off-farm activities to supplement income), we would expect to observe higher self/proxy discrepancies in child labor statistics as monitoring of child activities becomes more costly.

We find statistically significant gender variation in self/proxy discrepancies in child labor statistics in rural Ethiopia regardless of the age of the child or whether one uses 30-day or 7-day recall interviews. Girls report higher farm labor participation relative to proxy respondents within and across agricultural seasons. For boys, on the other hand, we do not find self/proxy information gaps across all seasons. These discrepancies in child labor statistics show variation across seasons

with higher gaps emerging in the main rainy and harvesting seasons relative to the short rainy season. However, we cannot reject the equality of reporting gaps across all three seasons. At the intensive margin of work, results are measured with less statistical precision as we do not find meaningful differential self/proxy survey design effects between girls and boys within and across seasons. A detailed power analysis shows that self-reporting appears to yield a good compromise between survey precision and costs for national surveys on child labor.

Our paper contributes to the measurement of agricultural data in some domains. First, we contribute to recent randomized studies in developing countries that analyze the effects of survey design on outcomes such as adult labor (Bardasi et al 2011), health (Das et al 2012), and education (Baird and Ozler 2012).[2] For child labor, Dillon et al. (2012) document the extent of variation in child labor statistics across respondent type and length of the child labor module in Tanzania. Dillon et al. (2012) reported that using a proxy or asking the child survey questions directly does not affect child labor statistics while using a short labor module generates lower child labor statistics by defining more precisely what work means, thus filtering out children that report labor market activities instead of domestic activities. While our experimental research design is motivated by and broadly similar to Dillon et al.'s (2012) study, we focused on agricultural activities and collected data in three different seasons of the agricultural calendar to account for the seasonality of self/proxy discrepancies.

Second, this study also contributes to the small but growing literature on seasonality and recall in agricultural data that, through its focus on adult labor, agricultural inputs and harvest measures, has cautioned that the degree of distortion in agricultural statistics depend, among other things, on the level of data aggregation and the seasonal timing of household surveys (e.g., Beegle et al 2012, Arthi et al. 2018, Gaddis et al 2019). In this study, we report variation in child labor

statistics across seasons, gender gaps in farm work that varies through the agricultural calendar, and self/proxy discrepancies in child labor statistics that vary according to the seasonal timing of the surveys. As informational constraints may be present in contexts where farms are mostly operated by families and monitoring is costly (Bharadwaj 2015), reporting of children's activities by proxy respondents could be affected, for instance, by the degree of complementarity of child's effort to adult labor activities in particular seasons.

Third, this study talks to the stream of literature that shows that women's work is poorly measured in developing settings (e.g., Reynolds and Wagner 2012, Mata and Greenwood 2000, Anker 1983), particularly in agricultural households where proxy respondents tend to rely on their general assumption about the state of the world in their search for answers to survey questions given the absence of recall-and-count information (Beegle et al. 2012). Our study shows the use of proxies becomes a source of gender bias in the measurement of child labor for girls in settings characterized by patterns of gender stratification of intrahousehold time allocation.

Fourth, this study may carry implications for the current debate on the measurement of agricultural productivity (e.g., Gollin et al. 2014, McCullough 2017), wherein child labor is used as an input in the production function. Our results suggest that eliciting labor information on child labor from a proxy respondent may affect the measured value-added per worker, and therefore productivity gaps by gender, or between agricultural and non-agricultural sectors in countries where child labor has a high prevalence.

This study is organized as follows. Section 2 describes the context of this study, the sampling procedure, and survey design intervention. Section 3 presents the empirical strategy, while Section 4 presents the main results. Section 5 assesses further impacts and potential channels that explain the main results. Finally, Section 6 presents the conclusions.

2. Setting and Experimental Design

2.1 Study Area and Sample

This study focuses on Ethiopia, which has one of the highest rates of child labor in the world, with 54 percent of rural 5 to 14-year-old children involved in economic activities, mostly as unpaid workers in family farms (Guarcello and Rosati 2007).[3] Our survey design experiment is carried out in two different regions in Ethiopia (Jimma and Sidama) that produce two different varieties of Arabica coffee with high demand in international markets. We focus on coffee cultivation since it is a child labor-intensive crop due to the characteristics of the tasks associated with the pre-harvesting and harvesting production process (Kruger 2007). Before the harvest season, it has been documented that children participate actively in pruning, weeding, and fertilizing. At harvest, coffee producers employ children mainly as pickers of red coffee beans which must be picked immediately upon ripening to maximize their quality (ILO 2004).

Our population framework is based on 5100 smallholder farmers who are active members of four Fairtrade coffee cooperatives that are spread out over 12 different districts (Kebeles). Within each selected region, we selected two representative Fairtrade Coffee Cooperatives, one characterized as of 'high' productivity and the second as of 'low' productivity to improve the external validity of the sample.[4] Sample selection is based on a 2x2 stratified random design: we split the population of farmers into high- and low-production groups according to whether they were above or below household median coffee production in 2014 based on administrative records. As variation in household coffee production could entail different combinations of adult/child work, this approach yields a sample that is representative of low and high coffee production

household units. Likewise, we stratified the population according to the gender of the household head as there may be gender differences in preferences and attitudes toward child labor. However, more than 90% of heads of households are males since we did not oversample female heads of households to keep the same observed male-female head of household ratio as in the population. Our representative sample is comprised of 1203 households.

One-third of the sample is randomly allocated to the self-response survey design group, while the remaining two-thirds to the proxy-response survey design group. The selection of the proxy respondent is limited to the household head or the spouse thereof to reflect the common practice of interviewing an informed household member. This statistical design yields 401 and 802 household units in the self-reported and proxy-reported groups, respectively. A statistical power analysis of a two-sample mean difference based on a two-sided 5 percent-level test and an effect size of 10 points shows a statistical power above the conventional threshold of 80 percent.[5]

2.2 Survey Design and Questionnaire

To investigate the effects of survey design on child labor statistics, we apply the same survey instrument to randomly selected respondents: children answer the questions themselves in the self-reported group and heads of households or spouses in the proxy group. Field surveyors ask the randomly selected respondent a specific labor-market module about child labor activities in the last 30 days before the survey (a child is defined as 6-14 years of age). There is no distinction in the wording nor in the sequencing of the questions across groups to avoid the possibility of bias in the way information is elicited. We ask the same questions in three agricultural seasons and make sure that the selected respondent in each group does not change over time. The overall compliance

with the survey instrument was above 98%.[6] In the second and third surveys, we also elicited information for a shorter reference period, i.e., last week before the survey. [7]

As the measurement of child labor statistics could be affected by how child labor definitions are operationalized in the survey design questionnaire (e.g., Bardasi et al. 2012), we employ a relatively long rather than a short questionnaire design. Specifically, the survey contains 12 questions that aim to elicit information about the specific farm and non-farm labor activities. The most important questions refer to work at the household farm, as this activity accounts for most children work in rural Ethiopia (Guarcello and Rosati 2007). A typical labor question asks: "*Did [name] work any time in the household farm in the last [..] days?"* Since the keyword "work" can have a different meaning for respondents, we supplement this question with a detailed, standard explanation of the concept of work by using a set of typical farming activities which is read aloud to the respondents: planting, watering, weeding, mulching, seeding, fertilizing, handpicking cherry coffee, cattle herding. This question is followed by an 'intensive margin' question, "*how many hours in the last [..] days did [name] spend working on the household farm?"* [8] It is worth noting that household chores such as fetching water and/or firewood, house cleaning, cooking, and child and elderly care are explicitly treated and explained as activities that do not belong to farm work. This approach aims to provide children and proxy respondents a clearer description of what constitutes child work at the farm.

2.3 Timeline and Sample Attrition

Fieldwork took place from July 2015 to January 2017. As the variation in work activities across the year depends on the agricultural calendar, we implemented the same survey design experiment

for the same households over three different coffee seasons coinciding with Ethiopia's rainfall seasons:

i) First survey: July-August 2015, during the *Meher* or main rainy season. This is the period of final coffee fruit development and the sowing of other crops (Moat et al 2017). This period is also known as the lean season due to the relatively low agricultural activity and its negative impact on agricultural income. This is also the time of year during which children are out of school.

ii) Second Survey: during April-May 2016, the *Belg* season or short rainy season. This is the period that corresponds to coffee planting, seeding, weeding, and early development of the coffee fruit. Land preparation for other crops takes place as well.

iii) Third Survey: during December 2016-January 2017, harvest season or dry season, the busiest agricultural season for coffee-growing households (Dercon and Krishnan 2000). Red cherry coffee crops are harvested and coffee processing and selling take place.

Importantly, the levels of attrition over time are very low. Out of 1,198 households surveyed in 2015, only one household was not surveyed in early 2016 and 10 households were not surveyed in late 2016. Regarding children aged 6 – 14 years, we surveyed 1,890 children in the first survey. We observe some children dropping out of the sample in the second and third surveys due to reaching the age of 15 during that year (100 children) and other children not present in the first survey but returning to the household in later surveys (46 children).[9]

2.4 Sample Characteristics

Out of the 1,200 households that form the experimental sample, a total of 1,197 were interviewed in July/August 2015, including 405 in the self-reported households and 792 in the proxy-reported households, respectively. As shown in Table 1, the average household head is male, 50 years of age, and has 4 years of formal schooling. On average, household size is 5.6 with 1.5 children aged 6 to 14 per household. These children are 10 years of age on average, with 2.5 years of formal schooling. The average monthly household income was around 1200 birr (about US$52), which was somewhat higher relative to the income of the average farmer in Ethiopia. Nonetheless, our sample participants live in houses with poor infrastructures such as mud-based floors (70 percent), no electricity (78 percent), and no sanitary services (80 percent). On average, the extent of the household plot is close to one hectare, of which 58 percent is used for coffee production while the remaining land is mainly used for cultivation of enset and maize. On average, households reported their plots as being located a 22-min walk from the closest primary school. The p-values of the t-test for the equality of means between the self-reported and proxy groups show statistical balance across all variables for the overall experimental sample and the sub-sample of households that have at least one child in the 6-14 age category. It is worth noting that around 90% of proxy respondents are males.

To address potential concerns about the representativeness of these data concerning non-Fairtrade smallholder farmers, we compare key characteristics of our sample with that of a nation-wide representative sample of rural households by using the 2015/2016 Ethiopian Socioeconomic Survey (ERSS) that was implemented by the Central Statistics Agency (CSA) as part of the Integrated Surveys on Agriculture program in close collaboration with the World Bank. The ERSS data is a representative survey of 4,954 households living in rural areas, small-town, and medium and large-sized towns.[10] By looking at Appendix Table A.1.1, one observes small differences in

socio-demographic variables for the household, heads of households, and children. Importantly, we also report descriptive statistics for child labor participation. The reference period for the child labor questions in the ERSS survey is the last 7 days before the survey date, and the age cohort refers to children 7 to 14. The labor module is answered by children aged 10 or older, while caretakers answer the questionnaire on behalf of children aged 7 to 9 years. With these caveats in mind, the child labor statistics emerging from the ERSS survey are similar (52%) to the number that emerges from our sample in the short rainy season.

3. Empirical Approach

We measure the mean effects of the survey design assignment following two approaches. First, we report descriptive (mean) statistics by treatment assignment for child labor at the extensive and intensive margins for both 30-day and the 7-day recall periods. This variation in the length of the recall period would allow us to assess whether the child/proxy discrepancies in child labor statistics is due to recall decay bias. We present results for the pooled data and each season separately across boys and girls subsamples.

Next, we estimate survey design treatment effects for pooled specifications to maximize power and test whether information gaps between self- and proxy-respondents vary according to the gender of the child. Irregular and less predictable activities are difficult to monitor and therefore recalling the occurrence of events that are less salient or unusual is more costly for respondents (e.g., Bound et al. 2001). Thus, if work schedules for girls are less salient and more variable across the agricultural calendar, proxy respondents are more likely to focus on information that is less prominent, resulting in larger discrepancies of child labor reports. On the other hand, if farm work

for boys is typically more permanent and regular across the agricultural calendar, it is more likely that both proxy and boy respondents would rely more on recall-and-count mental strategies to report it. As a result, we expect that self/proxy reporting of child labor would be less prone to discrepancies for boys.

In the absence of recall-and-count information, respondents are more prone to rely on their beliefs about the state of the world in their search for answers to survey questions (Beegle et al. 2012). Indeed, it has been documented that culturally defined roles affect the female labor share in crop production (e.g., Palacios-Lopez et al. 2017), as culture-specific gender roles determine the capacity of men and women to allocate labor time across economic activities (e.g, Ilahi 2000, Blackden and Wodon 2006), which has consequences for the measurement of socio-economic variables (Beegle et al. 2017).

For individual *i,* in Fairtrade cooperative *j* at season *t,* we then estimate the following equation,

$$y_{ijt} = \alpha + \beta_1 SR_{ij} + \beta_2 G_{ij} + \beta_3 (SR_{ij} \times G_{ij}) + X_{ijt}'\gamma + S_t + c_j + \varepsilon_{ijt}, \qquad (1)$$

where y_{ijt} denotes labor force participation or work hours, SR_{ij} is an indicator that takes the value 1 for self-reporting and 0 for proxy reporting, G_{ij} equals 1 for girls, and 0 for boys. β_3 shows whether the gender of the child affect the discrepancy of the reports and $\beta_1 + \beta_3$ shows whether the child provides a different response than the proxy. X_{ijt} is a vector of individual and household characteristics. S_t denotes agricultural season indicators with the short rainy season as the base category, while c_j is a Fairtrade cooperative fixed-effect that controls for potential institutional differences that could be correlated with the outcome of interest. ε_{ijt} is the idiosyncratic mean-zero error term. Robust standard errors are clustered at the household level.

Moreover, we use an extended regression framework that includes interactions of survey design status, gender, and seasonal indicators to test whether gender differences in self/proxy reporting vary across seasons since children perform different activities across seasons. In the main rainy season, children do not follow the daily school routine as the timing of the survey coincides with the school year's vacation[11]. Also, this period corresponds to the lean (or hungry) season in which farm income is scarce and, therefore, the heads of households might allocate effort to other off-farm activities to supplement income. As a result, proxy monitoring of child activities is more difficult to achieve, and thus proxies may be more likely to forget the occurrence of less salient events. The short rainy season, on the other hand, is a relatively quiet period during which children follow routine school days and the main agricultural activities consist of weeding, land preparation, and planting. These tasks follow specific steps with predictable timing and involve the joint participation of adults and their children (Admassie and Bedi 2008). Indeed, land preparation and planting cannot be done by children alone due to the physical strength requirements of these activities.[12] Thus, the specificity of these tasks made them salient and relatively easier to monitor by the heads of households.

At harvest, the busiest agricultural season that only occurs once per year and where farm labor demand is at its highest, children are dedicated to collect the dried coffee crops from the ground and pick the cherry-coffee bean from the trees. These activities do not require physical strength and can be done by children on their own. In this season, proxy respondents are actively engaged in picking, transporting, marketing, selling, storing, and drying coffee cherry crops as their labor demand is at its peak (Dercon and Krishnan 2000). This period is also characterized by active social festivities in coffee towns as coffee farmers receive cash windfalls from selling their

cherry-coffee crops. Thus, and similar to the main rainy season, there are fewer opportunities for monitoring child labor activities in the harvesting season relative to the short rainy season.

We thus estimate the following multivariate equation,

$$y_{ijt} = \alpha + \beta_1 SR_{ij} + \beta_2 G_{ij} + \beta_3 (SR_{ij} \times G_{ij}) + \beta_4 (SR_{ij} \times rain_{it}) + \beta_5 (SR_{ij} \times harv_{it}) + \beta_6 (G_{ij} \times rain_{it}) \quad (2)$$
$$+ \beta_7 (G_{ij} \times harv_{it}) + \beta_8 (SR_{ij} \times G_{ij} \times rain_{it}) + \beta_9 (SR_{ij} \times G_{ij} \times harv_{it}) + X_{ijt}'\gamma + S_t + c_j + \varepsilon_{ijt}$$

where $rain_{it}$ is and indicator for the main rainy season and $harv_{it}$ is an indicator for the harvesting season. The coefficients of interest are β_8 and β_9. We formally test the null that $\beta_8 = \beta_9 = 0$

4. Results

4.1 Differences in child labor statistics by respondent type and by season

Table 2 presents differences in means for child labor indicators by treatment assignment and disaggregated by gender for the pooled data and each season separately. The upper panel shows means at the extensive margin while the lower ones at the intensive margin. In each case, we test for differences in means between self- and proxy statistics based on a 30-day recall period. Descriptive statistics for the 7-day recall period depict similar patterns and thus are not discussed but are available in the Appendix Table A.1.2.

By looking at the left panel for the pooled data, we observe that participation in farm work obtained from the child (60.4%) is statistically significantly different from the one obtained from the proxy-respondent (56.1%). This discrepancy is driven by the girls' subsample that reports 7.2 percentage points (or 15%) higher rate of participation than the mean participation obtained from the proxy. On the contrary, the difference in reports is marginal and it is not statistically significant

for boys. By looking at each season separately, we observe significant differences in the share of children involved in farm labor activities regardless of the respondent type. The overall rate of labor force participation is 30 and 23 percentage points higher in the harvesting season compared to the main and short rainy seasons. Regardless of the gender of the child, corresponding self/proxy reporting differences in child labor show seasonal variation: 6.6 percentage points (15%) in the main rainy season, 2.7 (5.3%) in the short rainy season, and 4 percentage points (5.4%) in the harvesting season.

Moreover, we observe that labor participation for boys is typically more frequent and regular across seasons than that for girls. The coefficient of inter-variability of the self-reported measures of child labor, which gives a sense of how close the experimental values are to each other across seasons, is twofold for girls (34%) relative to boys (17%).[13] As a result, the gender gap in farm work participation varies across seasons as it reaches 20, 16, and 10 percentage points in the main rainy, short rainy, and harvesting seasons, respectively. While these numbers may reflect a pattern of gender stratification of intrahousehold time allocation with the salience of farm work for boys and more irregular work schedules for girls across the agricultural calendar, we consistently observe higher self/proxy discrepancies in child labor reporting for girls than that for boys across all seasons, reaching 9.2 percentage points (28%) in the main rainy season, 4.4 percentage points (11%) in the short rainy season, and 8 percentage points (12%) in the harvesting season. These results are statistically significant at 5% for all but the short rainy season.

The second panel reports child labor statistics at the (unconditional) intensive margin. Results are mostly consistent with the patterns emerging from the extensive margin of work: boys work at the farm several more hours than girls regardless of respondent type. For the pooled data, this difference reaches 10 monthly hours, or 50% higher monthly hours than the mean for girls.

This gender gap in hours of work holds across all seasons: 11hr in the main rainy season, 13hr in the short rainy season, and 7hr in the harvesting season. Moreover, and similar to the extensive margin results, we observe that children reported on average 2.2 (or 9.5%) higher worked hours than the mean hours obtained from the proxy. These self/proxy reporting differences also show variability across seasons: 2.8hr (18%) in the main rainy season, -0.05hr (-0.2%) in the short rainy season, and 4hr (12%) in the harvesting season. These results are statistically significant at 10% for all but the short rainy season. While the unconditional hours spent on-farm activities reported by proxies are lower than the self-reported for both boys and girls, however, we do not observe statistically significant child/proxy discrepancies for girls relative to boys either in the pooled data or in each season separately. Indeed, the coefficient of inter-variability of the self-reported measures of worked hours is twice the size of the coefficient of inter-variability computed for farm labor participation for boys and girls, suggesting the noisier nature of reporting worked hours.

4.2 Regression Results

Tables 3 and 4 depict the OLS results for the 30-day and 7-day recall periods at the extensive (columns 1-3) and intensive (columns 4-6) margins following equations 1 and 2.[14] We test directly whether differences in the reported participation in farm activities are affected by the gender of the child and whether gender survey treatment effects vary across seasons. Results show statistically significant gender gaps for self/proxy reporting in child labor statistics. Column 2 in Table 3 shows that girls have 6.4 percentage points higher discrepancy in the participation of farm activities report compared to boys, or a 13.5% higher rate of participation than the mean participation obtained from the proxy.

We test in column 3 whether these gendered survey impacts vary across seasons when the self-report indicator interacts with the gender and agricultural seasons indicators. The magnitude of the coefficient on this interaction term is non-negligible and, as expected, higher in the main rainy season (2.3 percentage points, or 5%) and harvesting season (3.7 percentage points, or 8%) relative to the short rainy season, suggesting that the differential child/proxy discrepancies in child labor statistics for girls vary across seasons. However, we do not have enough power to reject the equality of these interaction coefficients as shown by the p-values at the bottom of the table.

The panel at the right of Table 3 depicts results at the intensive margin of work. While in column 4 we observe meaningful overall child/proxy discrepancies in reporting of worked hours equal to 1.99hr, or 11% from the mean hours reported by the proxy, we do not observe statistically significant differential survey impacts for girls relative to boys (column 5), or differential gendered impacts that vary across seasons (column 6).

Turning our attention to the 7-day recall period in Table 4, column 2 shows that girls have a statistically significant 5.5 percentage points higher discrepancy in the participation of farm activities report compared to boys, or a 10.3% higher rate of participation than the mean participation obtained from the proxy. Likewise, column 3 shows that the higher child/proxy discrepancies in reporting for girls that that for boys increases in the harvesting season for 5.7 percentage points (10.7%), relative to the short rainy season. Moreover, as the 30-day recall results, the point estimates for the interaction of self-report indicator, child's gender, and season indicator are somewhat larger in the harvesting season than in the short rainy season but it is not measured with statistical precision. Furthermore, by looking at the intensive margin of work for the 7-day recall period in columns 4-6 in Table 4, we observe the same patterns concerning the 30-day recall period. We do not observe meaningful differential self/proxy discrepancies in child labor reporting

by gender and by season lines. Since reporting hours of work is inherently noisy for farm labor, discrepancies may tend towards a mean zero, and thus, gender differences are not large enough to be statistically meaningful.

Overall, we do not find support for the length of the recall period as the mechanism that explains our results. If forgetting is a mechanism that would explain child/proxy discrepancies, then one would expect to observe higher discrepancies for the 30-day recall period than 7-day interviews.

4.3 Further Results

The analysis of self and proxy reporting on household chores and schooling offers a way to assess reporting discrepancies for activities that, unlike agricultural farming for girls, are arguably frequent, regular, and continuing, and thus, equally salient for boys and girls.[15] Household chores (i.e. fetching water, firewood, house cleaning, cooking, laundry, childcare, and elderly care), which are performed mostly within the household premises, are relatively easy to monitor as they involve frequent and routinely activities over the agricultural calendar. Likewise, schooling is a well-defined activity that follows a regular yearly schedule with around 5 hours of daily classes during the school year.

Following similar specifications given in equations 1 and 2, we report self/proxy treatment estimates for participation in household chores (columns 1-3), hours spent on household chores (columns 4-6), and school participation (columns 7-9) in Table 5. [16] Results show negligible and not statistically significant differences between the child and the proxy reports for either variable. Moreover, the point estimates are close to zero for the coefficient associated with the interaction

of self-report treatment and gender of the child. These results reinforce the idea that child labor statistics, which are arguably less salient for girls, are more prone to variation by respondent type than other activities performed in the household.

The measurement error literature reports that recall bias can also be exacerbated in respondents with lower cognitive and communicative skills (Borgers et al. 2000). To test whether cognitive burdens of reporting explains the self/proxy discrepancies in child labor statistics, we rely on the differential impacts of the age of the child. Columns 1-2 in the Appendix Table A.1.3 show the estimated coefficient for the interaction between treatment status and child's age at the extensive and intensive margins, following similar specifications as in equation (1). We find negligible and not statistically significantly differential child/proxy discrepancies by child's age. Thus, these results suggest that cognitive burdensome bias does not explain discrepancies in the measurement of child labor statistics in our setting.[17]

Furthermore, it is expected that the extent of complementarity of farming tasks between children and proxy respondents, or more broadly, the likelihood of monitoring child activities by the proxy respondent, would reduce information discrepancies in child labor statistics. We then investigate heterogeneous survey treatment effects for two variables of interest, the gender of the proxy respondent and land size. While we acknowledge none of these variables provide direct information on monitoring of child activities, nonetheless, they provide useful information. Based on a specification presented in equation (1) after considering interaction terms for the survey treatment indicator, the gender of the child, and gender of the proxy respondent (or land size), unreported results shows that the differential survey impact for girls relative to boys increases in 6.4 percentage points when the proxy respondent is male. However, we cannot measure this sizable heterogeneous impact with statistical precision which is expected given the small number of

female proxy respondents over which this sub-group analysis is executed. For land size, on the other hand, we find statistically significant heterogeneous impacts of 4.5 percentage points per additional hectare of land.

Finally, we address the potential concern that self/proxy discrepancies in child labor may be affected by Fairtrade cooperative certification. It is important to highlight that Fairtrade does not ban children's engagement in the household farm but aims to regulate it by allowing children to work in family farms as long it is outside school hours and free of risky activities. We investigated self/proxy discrepancies of child labor reporting due to variations in proxy respondents' knowledge of Fairtrade standards related to labor, environmental regulations, and Fairtrade pricing. Based on this information we computed a knowledge index using principal component analysis.[18] Unreported results show that the coefficient associate with the interaction between the survey treatment indicator and the knowledge index is negative, negligible, and not statistically significantly different from zero. These findings show that knowledge of Fairtrade standards does not have a positive and systematic differential effect on the self/proxy reporting of child labor statistics.

5. Costs and Statistical Power Implications

Results presented in section 4 have implications for survey costs and statistical power. One way to address this trade-off is to ask how many household surveys would be needed to detect meaningful impacts for a hypothetical randomized intervention that targets child labor as the outcome of interest and relate that finding to survey cost considerations. To implement that exercise, we use self- and proxy-reported pooled tabulates of child labor for girls as the baseline

benchmark. Table 6 provides the sample size one would need under three impact scenarios: 5%, 10%, and 15% reduction in child labor.[19] For small impacts on child labor (5%), the top panel shows that one would need a large number of household observations to detect meaningful impacts according to the self-reported (n=4988) and proxy-reported (n=6769) statistics of a baseline measure of child labor. For expected moderate impacts (10%), the middle panel of Table 6 show the sample size requirements shrink to n=1295 and n=1684 for self and proxy reporting, while for relatively large impacts (15%) the bottom panel shows one would need sample sizes of n=583 and n=765, respectively. Thus, across these three different impact scenarios, the sample differences account for 1781, 389, and 182 additional observations if one uses the proxy reported child labor statistics as the baseline measure, relative to the self-reported child labor statistics.

Connecting these sample differences to survey costs allows one to assess the trade-offs between survey precision and costs. In rural Ethiopia where communications infrastructure is mostly absent, the costs associated with surveys fieldwork are high as any local research organization needs to move its operations from the capital city of Addis Ababa to rural areas for several days or weeks. The overall costs of collecting information in the field through multi-topic survey questionnaires that comprise a standard number of sections and questions were approximately 40US dollars per household. As indicated in columns 4 and 5 in Table 6, a large share of these costs is fixed (around 25US dollars) and thus, the variable costs associated with the additional time it would demand seeking information from each child within the household is not proportional to the number of children in the house.[20] Since self-reporting involves, on average, two respondents per household, as opposed to one adult respondent in proxy-reporting households, and since children are less mobile and siblings tend to be together, based on a detailed breakdown

of the budget data we calculated that additional survey costs per household are around 20% higher for self-reporting than that for proxy-reporting.

Next, we proceed to make a valuation of the trade-off between power analysis and survey costs. For girls, power analysis shows that one would need hundreds of additional observations to detect meaningful impacts if one uses the proxy reporting results rather than self-reports as the baseline, control-group statistic. Thus, given the high costs of field surveys in rural areas of underdeveloped countries, one would incur higher costs if planning a hypothetical RCT intervention by using proxy-reporting rather than self-reporting particularly when one expects small or moderate impacts on child labor. Put differently, the extra costs that would be incurred by surveying 1781 (5% impact) or 389 (10% impact) additional households under proxy reporting exceed the extra costs of having a self-reporting survey with the same fewer number of household units. For boys, on the other hand, self/proxy reporting does not affect child labor statistics, and thus there seem to be no benefits for carrying out self-reporting interviews given the somewhat higher costs associated with it. Self-reporting appears to yield a good compromise between survey precision and costs for national surveys on child labor, particularly in settings where gender stratification of work and social lives is the norm. Further research is necessary to generalize and validate this preliminary claim.

6. Conclusions

Over the past decades, the increasing availability of household surveys has helped to create a large body of research on the main determinants of child labor (e.g., Edmonds 2009) and the effects of social protection policies on children's allocation of time (e.g., Dammert et al 2018). However,

this progress has also raised questions about the survey methods by which child labor statistics are collected. While the agricultural sector accounts for by far the largest share of child labor, agricultural data is particularly vulnerable to recall errors due to the inherent seasonality of the activity, variable and irregular work schedules, multiple plots, and saliences of individuals and work activities (Beegle et al. 2012, Arthi et al. 2018).

This study addressed the measurement of agricultural child labor in rural Ethiopia with attention to the respondent type, seasonality, and gender-based information gaps. Our survey design intervention consisted of the random manipulation of the survey respondent in the application of the same survey instrument, which allows us to causally estimate survey design unbiased impacts of self/proxy reporting on child labor statistics. This design avoids potential strategic behavior where self- and proxy respondents within the same household might try to provide similar answers. To account for seasonality effects, we implemented the same survey design experiment in three different agricultural seasons of coffee production as the salience of farm activities, the complementarity of work across household members, and the overall demand for household labor vary across seasons.

The main findings show statistically significant child/proxy reporting gaps in child labor statistics for girls but not for boys. These results are prominent at the extensive margin of work but not at the intensive one. Although both self and proxy reports are measured with error, this result is in line with research that has shown that measurement in agricultural data is affected by the gender division of farm labor across agricultural seasons (e.g. Beegle et al. 2017), and by norms and beliefs about female employment (Reynolds and Wagner 2012, Lee and Lee 2012). We did not find support for recall decay bias as the mechanism that explains our results. Neither did we find support for cognitive burdens of reporting bias in child labor measurement.

While our results do not have enough power to statistically distinguish varying self/proxy discrepancies in child labor statistics over different agricultural seasons, the point estimates are ordered: self/proxy discrepancies in child labor reporting are at the lowest when the agricultural season involves routine steps and complementarity of tasks between children and adults i.e., planting and fertilizing usage, while higher discrepancies emerge in the lean (or hungry) season in which farm activities are less salient, income is scarce and, therefore, the heads of households might allocate effort to other off-farm activities to supplement income, or in the coffee harvest season, when farm labor demand is at its highest and farm activities can be done by children in their own. Future studies with greater power can potentially provide more definitive evidence on the seasonality of survey treatment effect for child labor statistics.

Our results are different from Dillon's et al. (2012) experimental survey design on child labor statistics in Tanzania. We find that given variable and irregular work schedules across the agricultural calendar for girls, it is plausible to find higher self/proxy discrepancies in farm child labor statistics. The difference in results highlights the need to understand the type of activities children perform and the context in which the surveys are implemented.

Our results also offer information for survey costs and statistical power trade-offs. A detailed breakdown of survey costs showed that for potential interventions that target child labor as the outcome of interest, self-proxy reporting offers a good compromise between survey precision and survey costs mainly if the expected impact on child labor is small or moderate.

Overall, this survey design intervention in rural Ethiopia points out that studies of child labor in agricultural settings should explicitly acknowledge and discuss survey design gender gaps in child labor statistics for a better understanding of its determinants and conditions, and for the design of social protection programs and policies.

Appendix A1: Additional Tables

Table A.1.1: ERSS Household and Demographic Characteristics

	2015/2016 Ethiopian Socioeconomic Survey	2015 Ethiopian Agricultural Labor Survey
Panel A: Household socio-demographics		
Household size	6.19	5.61
Children aged 6-14	1.54	1.60
Panel B: Head of Household		
Gender (% Male)	0.75	0.89
Age	47.86	50.22
Years of schooling	2.34	3.92
Married	0.76	0.86
Panel C: Children aged 6-14		
Gender (% Male)	0.51	0.49
Age	9.93	10.14
Years of schooling	1.68	2.50
Panel D: Child Labor		
Labor force participation	0.52	0.52

Source: Authors' analysis based on an agricultural household survey conducted in 2015 and the 2015/2016 Ethiopian Socioeconomic Survey (ERSS)

Table A.1.2: Child Labor Statistics by Survey Assignment and Season: 7-day Recall Period

	Pooled Data			Short Raining Season			Harvest Season		
	Self-reported	Proxy-report	Diff	Self-reported	Proxy-report	Diff	Self-reported	Proxy-report	Diff
Participation in Household Farm Activities (%)									
All	0.629	0.609	0.020	0.489	0.484	0.004	0.773	0.734	0.039
	[0.017]	[0.012]		[0.025]	[0.017]		[0.022]	[0.016]	
Boys	0.677	0.688	- 0.010	0.567	0.575	0.008	0.792	0.800	-0.008
	[0.022]	[0.015]		[0.031]	[0.022]		[0.025]	[0.018]	
Girls	0.580	0.532	0.048*	0.410	0.397	0.013	0.753	0.668	0.085**
	[0.022]	[0.015]		[0.033]	[0.023]		[0.031]	[0.024]	
Hours Spent on Household Farm Activities									
All	8.571	7.809	0.762*	6.167	6.338	- 0.171	11.043	9.292	1.751***
	[0.317]	[0.228]		[0.395]	[0.314]		[0.484]	[0.328]	
Boys	9.984	9.175	0.809	7.952	8.286	-0.334	12.089	10.062	2.028***
	[0.477]	[0.322]		[0.624]	[0.460]		[0.676]	[0.393]	
Girls	7.151	6.483	0.668	4.358	4.463	-0.105	10.000	8.538	1.462**
	[0.349]	[0.268]		[0.450]	[0.361]		[0.550]	[0.421]	
Hours Spent on Household Chores									
All	11.526	11.447	0.079	11.733	11.428	0.305	11.315	11.467	-0.152
	[0.365]	[0.248]		[0.525]	[0.353]		[0.435]	[0.328]	
Boys	9.550	9.375	0.175	9.802	9.547	0.255	9.290	9.205	0.085
	[0.407]	[0.297]		[0.590]	[0.447]		[0.482]	[0.400]	
Girls	13.515	13.447	0.067	13.695	13.216	0.478	13.332	13.683	-0.351
	[0.527]	[0.335]		[0.727]	[0.460]		[0.619]	[0.433]	

Source: Authors' analysis based on an agricultural household survey conducted in 2016.
Note: Standard errors are in brackets. Proxy respondents include head of households and spouses.
*p<0.1, **p<0.05, ***p<0.01

Table A.1.3: Age of Child and Cognitive Burdens for Recall Survey data

	Farm activities	Worked Hours
	(1)	(2)
Self-reported (SR)	0.010	-1.573
	(0.083)	(4.940)
Age	0.069***	3.946***
	(0.004)	(0.283)
Girls	-0.046	4.732
	(0.063)	(3.645)
SR*Girls	0.047	4.796
	(0.108)	(6.237)
SR*Age	-0.000	0.339
	(0.007)	(0.486)
Girls*Age	-0.013**	-1.401***
	(0.006)	(0.359)
SR*Girls*Age	0.002	-0.457
	(0.010)	(0.617)
Observations	5142	5409

Source: Authors' analysis based on agricultural household surveys conducted in 2015 and 2016.
Notes: Robust standard errors clustered at the household level in parenthesis. Control covariates are described in Table 3. Proxy respondents include head of households and spouses. *p<0.1, **p<0.05, ***p<0.01

ENDNOTES

[1] Likewise, the increasing number of available impact evaluations that study child labor as one of their outcomes of interest in areas as diverse as microfinance, training and education, conditional cash transfers, and public works, rely mostly on proprietary data that is collected mainly from the program beneficiary or head of household in the targeted households (Dammert el at. 2018).

[2] There is a wealth of evidence on measurement error on adult labor statistics in developed countries, particularly in the U.S., where labor statistics from household surveys is compared to administrative data (e.g., Bound et al 2001).

[3] In contrast to other developing settings, there are no social stigmas or negative perceptions of child labor in Ethiopia, as child labor historically has been viewed positively as training for children (Bass 2004).

[4] In Sidama, we worked with two Fairtrade cooperatives, each with around 1,500 active smallholder farmer associates. One cooperative reported a yearly average production of 1,122 kgs of coffee crops per associate, while the other reported average production of 789 kgs in 2014. In Jimma, we also worked with two Fairtrade cooperatives. The first cooperative had 800 active farmer associates and yearly average production of over 1,600 kgs of coffee per associate, while the second cooperative had 1,100 active farmer associates and yearly average production of 600 kgs of coffee per associate in 2014.

[5] The size effect considered for power analysis is based on a conservative approach to the survey design results found by Dammert and Galdo (2013) in Peru and Janzen (2018) in Tanzania.

[6] In the first survey, the compliance rate is around 97% as proxy respondents provided the information for 27 children in the self-report group, while other household members provided

information on behalf of 20 proxy respondents. Compliance rates are close to 100% in the second and third surveys.

[7] In the presence of widespread seasonal activities, the choice of the length of the reference period is important. A short length of the reference period (e.g., a day or a week) may not capture seasonal work depending on the precise timing of the survey work if labor inputs vary considerably across weeks (Levison et al 2007, Arthi et al 2017), or if the chosen day or week is atypical due to religious holidays or community celebrations (Matta-Greenwood 2000, Comblon and Robilliard 2017). A longer length of the reference period (e.g., 'last year'), on the other hand, could introduce bias in the measurement of variables due to recall or aggregation errors that operate firstly through decay in memory as the likelihood of forgetting an event increases as time passes. Thus, there are different layers of trade-offs with no firm conclusion on what time length is the most appropriate.

[8] The subsequent questions use the same length of the reference period to capture information on work in non-farm household business, non-farm wage work, Fairtrade coffee cooperatives, coffee plantations, and other households' farms. A negligible percentage (2%) of both self and proxy responses report children working outside the domain of household farm and, thus, this information is not used in the empirical analysis.

[9] Comparability of labor statistics across these three surveys follows from the similarity of the survey design, data collection and processing procedures. We used the same group of field surveyors across all rounds of data collection and the same software and personnel for data entry and coding. The wording and type of questions, the length of labor modules and the reference period are the same across the three survey rounds.

[10] ERSS data collection took place between February and April 2016, a timing that overlaps with our second survey intervention (short rainy season).

[11] 80 percent of children aged 6-14 attend formal school in our sample. No differences are observed for school enrollment between boys and girls.

[12] The use of plough shaft, ploughshare, plow, beam, and animal force is used for these activities.

[13] The Coefficient of Variability (%CV) for the self-reported measures of child labor across three seasons is estimated as $CV = (SD / Mean) \times 100.$ The lower the %CV, the less inter-variability.

[14] Corresponding Tobit specifications were also implemented at the intensive margin as the dependent variable has limited support with a mass point of zero. Qualitative findings do not change and thus we do not report these estimates.

[15] Descriptive statistics show that boys and girls spend the same number of hours in household chores and school activities.

[16] We did not collect information on household chores and school enrollment in the first household survey (main rainy season) due to time constraints as this multi-topic survey was extensive.

[17] This result does not change if we include a triple interaction of treatment indicator and child's gender and age.

[18] Questions about Fairtrade knowledge were asked in the first survey round (main rainy season). Households with response of "don't know" were treated as missing observations. For ease of interpretation, we normalized the estimated indices by using the mean and standard deviation of the control group

[19] We set significance level at $\alpha=0.05$, power $1-\beta=0.8$, and evenly split samples for treatment and control groups.

[20] A detailed breakdown of the budget data for this survey design intervention reveals that the share of training costs, transportation costs, field supervisors' and local senior researchers' honorariums, and institutional fee is mostly independent of self- or proxy-reporting survey, while it is

disproportionally higher with respect to variable costs associated to days of lodging and per-diems of surveyors.

References

Admassie, A. and A. Singh Bedi. 2008. "Attending School, Reading, Writing and Child Work in Rural Ethiopia." In *Economic Reform in Developing Countries*. Chapter 6. Edward Elgar Publishing.

Anker, R. 1983. "Female labour force participation in developing countries: A critique of current definitions and data collection methods". *International Labour Review*, 1226):709-23.

Arthi, V, K. Beegle, J. De Weerdt, and A. Palacios-Lopez. 2018. "Not your average job: Measuring farm labor in Tanzania" *Journal of Development Economics* 130:160-172

Baird, S., and B. Ozler. 2012 "Examining the Reliability of Self-reported Data on School Participation". *Journal of Development Economics*, 98(1):89-93

Bardasi, E., K.Beegle, A.Dillon, and P.Serneels. 2012. "Do Labor Statistics Depend on How and to Whom the Questions Are Asked? Results from a Survey Experiment in Tanzania." *World Bank Economic Review* 25 (3):418–47.

Bass, L..2004. "Child Labor in Sub-Saharan Africa". Lynne Rienner Publishers. London.

Beegle, K., G. Carletto, K. Himelein. 2012. "Reliability of recall in agricultural data", *Journal of Development Economics*, 34-41.

Beegle, K., J. Goldberg, and E. Galasso. 2017. "Direct and Indirect Effects of Malawi's Public Works Program on Food Security." *Journal of Development Economics* 128:1–23.

Bharadwaj, P. 2015. "Fertility and rural labor market inefficiencies: Evidence from India" *Journal of Development Economics*, 115:217-232

Biggs, B. 1992. “Self/proxy informant rules and data quality”, Research Paper, Income Research Paper Series, Ottawa: Statistics Canada.

Blackden, C.M. and Wodon, Q. (2006), “Gender, Time Use, and Poverty in Sub-Saharan Africa”. Vol. 73, World Bank Publications, Washington DC.

Borgers, N., E. de Leeuw, and J. Hox. 2000. “Children as Respondents in Survey Research: Cognitive Development and Response Quality.” *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique* 66:60–75.

Bound, J., 2001. “Measurement error in survey data” *Handbook of Econometrics*, Vol 5, pp. 3705-3843

Comblon, V., and A.S. Robilliard. 2017. “Are Female Employment Statistics More Sensitive than Male Ones to Survey Design? Evidence from Cameroon, Mali and Senegal.” Working Paper DT/2015-22, DIAL.

Dammert, A. and J. Galdo. 2013. “Child Labor Variation by Type of Respondent: Evidence from a Large-Scale Study.” *World Development* 51 (11):207–20.

Dammert, A., J. De Hoop, E. Mvukiyehe, and F. Rosati. 2018. “The Effects of Public Policy on Child Labor: Current Knowledge, Gaps, and Implications for Program Design.” *World Development,* 110:104-123.

Das, J., J. Hammer, C. Sanchez-Paramo. 2012. “The impact of recall periods on reported morbidity and health seeking behavior” *Journal of Development Economics* 98 (1):76-88

Dercon, S. and P. Krishnan. 2000. “Vulnerability, Seasonality, and Poverty in Ethiopia.” *Journal of Development Studies* 36 (6):25–53.

Desire, S. and V. Costa. 2019 “Employment Data in Household Surveys. Taking Stock, Looking Ahead” World Bank Policy Research Working Paper No. 8882

Dillon, A., E. Bardasi, K. Beegle, and P. Serneels. 2012. "Explaining Variation in Child Labor Statistics." *Journal of Development Economics* 98 (1):136–47.

Edmonds, E. 2009. "Child Labor." In *Handbook of Development Economics*. Vol. 4. Elsevier Science, Amsterdam, North Holland.

Gaddis, I., G. Oseni, A. Palacios-Lopez, and J. Pieters. 2019. "Measuring Farm Labor: Survey Experimental Evidence from Ghana" World Bank Working Paper No 8717

Gollin, D., D.Lagakos, M. Waugh. 2014. "The Agricultural Productivity Gap" *The Quarterly Journal of Economics*, 129(2): 939–993,

Guarcello, L. and F. Rosati. 2007. "Child Labor and Youth Employment: Ethiopia Country Study." SP Discussion Paper 0704, The World Bank.

Ilahi, N. 2000, "The Intra-household Allocation of Time and Tasks: What Have We Learnt from the Empirical Literature?" Policy Research Report on Gender and Development WP 13. The World Bank.

ILO. 2004. "Safety and Health Fact Sheet Hazardous Child Labour in Agriculture: Coffee." International Programme on the Elimination of Child Labour.

———. 2008. "Resolution II: Resolution concerning statistics of child labour", in Report of the Eighteenth International Conference of Labour Statisticians" (Geneva, 24 November– 5 December 2008). Document ICLS/18/2008/IV/FINAL. Geneva, pp. 56–67.

———. 2017. "Global Estimates of Child Labour. Results and Trends 2012-2016." International Labour Organization, Geneva.

———. 2018. "19th ICLS implementation: National LFS practices and implementation plans". Presentation prepared for the 20th ICLS. Last accessed on November 29, 2019 https://www.ilo.org/wcmsp5/groups/public/---dgreports/---

stat/documents/meetingdocument/wcms_646789.pdf

Janzen, S.. 2018. "Child Labor Measurement: Who Should We Ask?" *International Labour Review*. Volume 157, Issue 2

Kruger, D.. 2007. "Coffee Production Effects on Child Labor and Schooling in Rural Brazil." *Journal of Development Economics* 82 (2):448–63.

Lee, J. and S. Lee. 2012. "Does It Matter Who Responded to the Survey? Trends in the U.S. Gender Earnings Gap Revisited." *Industrial and Labor Relations Review* 65:148–60.

Levison D., J. Hoek, D. Lam, S. Duryea. 2007, "Intermittent child employment and its implications for estimates", *International Labour Review*. Volume 146, Issue 3-4.

Matta-Greenwood, A. 2000. "Incorporating gender issues in labour statistics. ILO.

Moat,J., J. Williams, S. Baena, T.Wilkinson, T. Gole, Z.Challa,S. Demissew and A. Davis 2017 "Resilience Potential of the Ethiopian Coffee Sector under Climate Change." *Nature Plants*. 3. 17081

Palacios-Lopez A., L. Christiaensen, and T. Kilic. 2017. "How Much of the Labor in African Agriculture is Provided by Women" *Food Policy* 52-63

Reynolds, J. and J. Wenger. 2012. "He Said, She Said: The Gender Wage Gap According to Self and Proxy Reports in the Current Population Survey." *Social Science Research* 41:392–411.

Table 1: Household and Demographic Characteristics by Survey Assignment: July 2015

	Household Random Assignment				Households w/ children aged 6-14	
	Self-reported	Proxy-report	Difference	p-value	Difference	p-value
Panel A: Household socio-demographics (N=1197)						
Household size	5.61	5.59	0.02	0.90	-0.03	0.81
Children aged 6-14	1.60	1.57	0.04	0.65	0.05	0.47
Christian (%)	0.57	0.55	0.02	0.57	0.01	0.81
House Mud floor (%)	0.71	0.71	0.00	0.92	0.01	0.70
Owns a mobile phone (%)	0.67	0.65	0.02	0.51	0.03	0.37
Walking distance to prim school (min)	22.39	21.74	0.65	0.47	0.80	0.44
Last month total income (Birr)	880.13	819.83	60.30	0.61	114.04	0.44
Yearly average monthly income (Birr)	1266.41	1258.54	7.87	0.93	-8.95	0.94
Land size (hectare)	1.11	1.06	0.05	0.49	0.04	0.63
Cultivated coffee % total land size	0.58	0.58	0.00	1.00	0.00	0.79
Yield of coffee per hectare	3323.31	3273.56	49.75	0.83	100.09	0.72
Share of red cherry sold to Fairtrade Coop (%)	94.47	94.55	-0.08	0.94	0.03	0.98
Panel B: Head of Household (N=1197)						
Gender (% Male)	0.89	0.87	0.02	0.24	0.01	0.53
Age	50.22	49.65	0.57	0.53	0.41	0.65
Years of schooling	3.92	3.82	0.11	0.63	0.09	0.73
Married	0.86	0.84	0.02	0.44	0.01	0.59
Panel C: Children aged 6-14 (N =1880)						
Gender (% Male)					0.01	0.69
Age					0.14	0.25
Years of schooling					0.05	0.65
Currently attending school					-0.01	0.44

Source: Authors' analysis based on an agricultural household survey conducted in 2015.
Note: Sample means computed from the first survey carried out in July 2015. P-values refer to the null hypothesis of equality of means between self-reported and proxy-reporting measures. Proxy respondents include head of households and spouses.

Table 2: Child Labor Statistics by Survey Assignment and Season: 30-day Recall Period

	Pooled Data			Main Rainy Season			Short Raining Season			Harvest Season		
	Self-reported	Proxy-report	Diff	Self-reported	Proxy-report	Diff	Self-reported	Proxy-report	Diff	Self-reported	Proxy-report	Diff
Participation in Household Farm Activities (%)												
All	0.604	0.561	0.044**	0.498	0.432	0.066**	0.536	0.509	0.027	0.786	0.746	0.040
	[0.015]	[0.010]		[0.026]	[0.018]		[0.024]	[0.017]		[0.021]	[0.016]	
Boys	0.665	0.653	0.013	0.580	0.543	0.037	0.613	0.608	0.005	0.809	0.810	-0.002
	[0.019]	[0.013]		[0.031]	[0.023]		[0.030]	[0.021]		[0.024]	[0.018]	
Girls	0.543	0.472	0.072***	0.418	0.326	0.092**	0.458	0.414	0.044	0.763	0.683	0.080**
	[0.020]	[0.013]		[0.035]	[0.022]		[0.033]	[0.023]		[0.032]	[0.023]	
Hours Spent on Household Farm Activities												
All	25.190	23.007	2.182*	18.195	15.371	2.824*	19.902	19.956	- 0.054	38.030	34.025	4.004*
	[0.939]	[0.624]		[1.300]	[0.868]		[1.391]	[1.028]		[1.757]	[1.287]	
Boys	30.400	28.056	2.344	23.877	20.888	2.989	25.768	26.167	- 0.399	42.079	37.301	4.779*
	[1.408]	[0.875]		[1.979]	[1.332]		[2.153]	[1.518]		[2.422]	[1.526]	
Girls	19.996	18.132	1.864	12.619	10.100	2.519	13.942	13.980	-0.038	34.007	30.816	3.191
	[1.008]	[0.707]		[1.429]	[0.925]		[1.441]	[1.157]		[2.032]	[1.632]	

Source: Authors' analysis based on agricultural household surveys conducted in 2015 and 2016.
Note: Standard errors are in brackets. Proxy respondents include head of households and spouses.
*p<0.1, **p<0.05, ***p<0.01

Table 3: Survey Design Average Effects: 30-day Recall Period

	Participation in Household Farm Activities (%)			Hours Spent on Household Farm Activities		
	(1)	(2)	(3)	(4)	(5)	(6)
Self-reported (SR)	0.041**	0.009	-0.005	1.997**	2.043	-1.001
	(0.017)	(0.020)	(0.033)	(1.008)	(1.477)	(2.418)
Girls	-0.154***	-0.176***	-0.190***	-9.509***	-9.478***	-11.777***
	(0.012)	(0.015)	(0.026)	(0.757)	(0.898)	(1.656)
Rainy Season	-0.053***	-0.053***	-0.056*	-3.043***	-3.043***	-4.720**
	(0.019)	(0.019)	(0.029)	(1.036)	(1.036)	(1.977)
Harvest Season	0.241***	0.241***	0.198***	15.416***	15.416***	10.982***
	(0.019)	(0.019)	(0.025)	(1.350)	(1.351)	(2.034)
SR*Girls		0.064**	0.045		-0.091	0.958
		(0.026)	(0.044)		(1.623)	(2.755)
SR*Rainy Season			0.033			3.364
			(0.048)			(3.236)
SR*Harvest Season			0.006			5.802
			(0.041)			(3.749)
Girls*Rainy Season			-0.024			1.301
			(0.037)			(2.241)
Girls*Harvest Season			0.068*			5.666**
			(0.038)			(2.479)
SR*Girls*Rainy Season			0.023			-0.774
			(0.063)			(3.667)
SR*Girls*Harvest Season			0.037			-2.364
			(0.065)			(4.353)
p-value of (SR*Girls*Rainy Season= SR*Girls*Harvest Season)			0.8219			0.6911
Observations	5,412	5,412	5,412	5,409	5,409	5,409
R-squared	0.214	0.215	0.217	0.199	0.199	0.201

Source: Authors' analysis based on agricultural household surveys conducted in 2015 and 2016.
Notes: Robust standard errors clustered at the household level in parenthesis. Control covariates include child characteristics (age and gender) and household characteristics (gender and schooling of the head of household, household size, and indicator variables for quartiles of household wealth). Cooperative fixed effects are included. Pooled regression includes time indicators. Proxy respondents include head of households and spouses. *p<0.1, **p<0.05, ***p<0.01

Table 4: Survey Design Average Effects: 7-day Recall Period

	Participation in Household Farm Activities (%)			Hours Spent on Household Farm Activities		
	(1)	(2)	(3)	(4)	(5)	(6)
Self-reported (SR)	0.020	-0.007	-0.018	0.669*	0.713	-0.561
	(0.019)	(0.023)	(0.034)	(0.344)	(0.498)	(0.701)
Girls	-0.133***	-0.152***	-0.176***	-2.577***	-2.548***	-3.703***
	(0.015)	(0.018)	(0.027)	(0.267)	(0.328)	(0.505)
Harvest Season	0.262***	0.262***	0.221***	3.588***	3.588***	1.701***
	(0.019)	(0.019)	(0.026)	(0.371)	(0.371)	(0.570)
SR*Girls		0.055*	0.027		-0.086	0.384
		(0.030)	(0.045)		(0.561)	(0.843)
SR*Harvest Season			0.022			2.564**
			(0.044)			(1.054)
Girls*Harvest Season			0.049			2.322***
			(0.040)			(0.695)
SR*Girls*Harvest Season			0.057			-0.929
			(0.067)			(1.243)
Observations	3,557	3,557	3,557	3,557	3,557	3,557
R-squared	0.213	0.213	0.215	0.205	0.205	0.211

Source: Authors' analysis based on agricultural household surveys conducted in 2016.

Notes: Robust standard errors clustered at the household level in parenthesis. Information on the last 7 days was not recorded from both the child and the proxy in the first survey. Control covariates are described in Table 3. Proxy respondents include head of households and spouses.

*p<0.1, **p<0.05, ***p<0.01

Table 5: Survey Design Average Effects on Household Chores and School Participation

	Weekly Participation in Household Chores (%)		Weekly Participation in Household Chores (%)		School Enrollment (%)	
	(1)	(2)	(3)	(4)	(5)	(6)
Self-reported (SR)	-0.005	-0.002	0.009	-0.045	0.011	0.011
	(0.011)	(0.017)	(0.392)	(0.454)	(0.012)	(0.017)
Girls	0.052***	0.054***	4.278***	4.241***	0.013	0.013
	(0.010)	(0.012)	(0.292)	(0.350)	(0.010)	(0.013)
Harvest Season	0.011	0.011	-0.102	-0.102	0.028***	0.028***
	(0.010)	(0.010)	(0.381)	(0.381)	(0.009)	(0.009)
SR*Girls		-0.006		0.109		-0.001
		(0.020)		(0.629)		(0.021)
Observations	3,564	3,564	3,553	3,553	3,547	3,547
R-squared	0.090	0.090	0.196	0.196	0.069	0.069

Source: Authors' analysis based on agricultural household surveys conducted in 2016.
Notes: Robust standard errors clustered at the household level in parenthesis. Information on the last 7 days was not recorded from both the child and the proxy in the first survey. Control covariates are described in Table 3. Proxy respondents include head of households and spouses.
*p<0.1, **p<0.05, ***p<0.01

Table 6: Power Analysis and Costs for Field Survey

Type of Respondent	Baseline proportion of child labor	Effect size on child labor	Sample size (households)	Fixed costs per household survey ($)	Variable costs per household survey ($)
Proxy	0.472	5%	6769	~25	~10
Self	0.543	5%	4988	~25	~17
Diff Proxy-Self			1781	0	-7
Proxy	0.472	10%	1684	~25	~10
Self	0.543	10%	1295	~25	~17
Diff Proxy-Self			389	0	-7
Proxy	0.472	15%	765	~25	~10
Self	0.543	15%	583	~25	~17
Diff Proxy-Self			182	0	-7

Source: Authors' analysis based on agricultural household surveys conducted in 2015 and 2016.

Notes: Power for two-sample proportion randomized test. This analysis assumes power=80%, significance level α=5%, ratio Ntreat/Ncontrol=1, and two children per household. Fixed costs include mostly training expenses, transportation from city to rural areas, honoraria for field supervisors and senior researchers, and institutional fee. Variable costs include expenses associate with days of lodging and per-diems of surveyors. Cost data (in US dollars) is calculated based on a multi-purpose household survey in rural areas of Ethiopia.