Harvesting rice in Tappita district, Liberia. Credit: Tommy Trenchard/Oxfam

MEASURING IMPACT

A Meta-Analysis of Oxfam's Livelihoods Effectiveness Reviews

ROB FULLER

INDEPENDENT CONSULTANT FOR OXFAM GB

Oxfam's Effectiveness Reviews evaluate the impact of the organization's projects on the lives of those they are intended to help. This paper uses statistical meta-analysis to summarize the results of 23 Effectiveness Reviews of livelihoods projects carried out between 2011 and 2016. The projects are found to have had a statistically significant positive impact on the welfare of participants, measured by household consumption or wealth. The paper also analyses changes in intermediate outcomes, such as crop production and sales, and investigates what can be learned about the measurement approach applied in the Effectiveness Reviews.

www.oxfam.org/effectivness

OXFAM

CONTENTS

Executive Summary **3**

1 Introduction **5**

2 Data **6**

2.1 Selection of evaluations 6

2.2 Evaluation methodology 6

2.3 Measurement of outcomes 7

2.4 Meta-analysis methodology 8

3 Meta-analysis results **10**

3.1 Overall impact on household consumption 10

3.2 Overall impact on indicators of household wealth 12

3.3 Differences in effect size by gender of the household head 15

3.4 Differences in effect size by pre-project wealth level 17

4 Effects on intermediate outcomes **19**

5 Exploring measurement approaches **23**

5.1 How sensitive are the results to household equivalence scales? 23

5.2 Is food consumption a good proxy for overall consumption? 27

5.3 Is dietary diversity a good proxy for consumption? 29

5.4 Should food security indicators be used as an outcome measure? 31

5.5 What can be learned from the comparison of results based on household income and those based on wealth indicators? 32

5.6 Should subjective welfare assessments be used as an outcome measure? 33

6 Learning from evaluations **36**

7 Conclusions **38**

Appendix 1: Evaluations included in the meta-analysis 39

Appendix 2: Methodology used for propensity-score matching 41

Appendix 3: Histograms of consumption per adult equivalent (standardized) 43

Appendix 4: Histograms of post-project wealth index (standardized) 44

Appendix 5: Robustness of meta-analysis estimates 45

Appendix 6: Differences between intervention and comparison groups in baseline wealth status 47

References **48**

Notes **51**

EXECUTIVE SUMMARY

In 2011, Oxfam GB established its Global Performance Framework (GPF) to enable the organization to deliver on its commitments to be accountable to its wide range of stakeholders and improve its ability to both understand and communicate the impact of its programmes in thematic outcome areas. The GPF is comprised of two key elements: a Global Output Report, which details what the organization is doing to bring about a world free of poverty, inequality, and injustice; and Effectiveness Reviews, intensive evaluation processes that consider the extent projects have contributed to change in relation to the particular global outcome indicator that it has been selected under. Closing and sufficiently mature projects contributing to five of Oxfam GB's seven global outcomes (livelihoods, resilience, women's empowerment, citizen voice, and policy influence) are randomly selected each year and rigorously evaluated.

This paper presents the findings of a meta-analysis of the 23 projects that have been evaluated for their impact on household livelihoods. Statistical meta-analysis has been used to examine what general lessons can be learned from pooling data from across the evaluations.

A conventional meta-analysis, such as those carried out by the Cochrane Collaboration or the International Initiative for Impact Evaluation (3ie) involves consolidating multiple studies of a single type of intervention, with the aim of making general conclusions about the effectiveness of that intervention. In contrast, this meta-analysis consolidates data on the effectiveness of a single organization, Oxfam GB, across the 23 livelihoods support projects. The projects evaluated were implemented in various locations around the world and were diverse in their scale and activities, but they all aimed to support participants in improving their livelihoods and material welfare. Each evaluation involved comparing project participant households to a group of comparison households that did not benefit from the project, with propensity-score matching being used to control for the main observable differences between the two groups.

Overall, the projects are estimated to have had a positive effect on the livelihoods of participant households. Household consumption is used as a proxy measure for household income. On average, the projects are estimated to have had a positive effect on household consumption by 0.12 standard deviations, with a 95% confidence interval ranging from 0.03 to 0.20 standard deviations. (A standard deviation is a measure of the variation in the outcome measure in each evaluation dataset. Calculating the project effects in terms of fraction of a standard deviation allows us to compare and aggregate results across the 24 evaluations.) This corresponds to an increase in household consumption of approximately 6.6%, with the 95% confidence interval ranging from 1.6% to 11.9%.

While some of the projects evaluated had more positive results than others, there are no clear differences in the average impact of projects between different regions of the world or between lower-income and middle-income countries. Surprisingly, the effect of projects also does not seem to depend consistently on the scale at which they were implemented, nor on the project's budget or duration. There is no evidence that the projects had greater impact among households that were initially poorer or wealthier than average. However, there is some evidence that female-headed households have tended to benefit less from the projects than male-headed households.

The evaluations also collected data on some intermediate outcome measures on which projects were expected to have an effect, particularly on agricultural production and sales. Projects that targeted a specific agricultural product or products (such as vegetables, coffee or dairy production) were generally found to be successful in promoting production and sales of those products. In some cases, this came at the expense of reducing the range of other crops grown by households, while other projects appear to have been successful in promoting crop diversification. Eight of the projects evaluated are found to have had an impact on the proportions of participants who were making sales of agricultural products. Not all of the

evaluations provide clear evidence of an impact on the revenue generated from these sales. However, in all four cases in which there is evidence of a positive effect on sales, the project is also found to have had a positive effect on household consumption; this suggests that the additional sales were profitable for the households making them.

The meta-analysis also provides an opportunity to test some important assumptions about the measurement approaches used in the livelihoods evaluations, and produced some interesting insights. For example, the results of the evaluations do not depend to any great extent on the way in which household consumption is calculated. A second finding is that, in most of the evaluations, food consumption is closely correlated with total consumption, though the projects tended to have a slightly smaller effect on food consumption than on total consumption. A test of whether relying simply on the diversity of food items consumed by the household would be a good proxy for overall consumption is also carried out, but the pattern of results for food diversity is found to be quite different to those for the value of food consumption. Data are also examined on households' experience of food security measures and on subjective evaluations of households' economic status: these measurements were routinely collected in the earlier evaluations, but were omitted in later evaluations because they had either been found to have low power to detect project effects, or to have produced anomalous results. The meta-analysis suggests that the decision to omit these more subjective measures in later evaluations was justified.

One alternative measure of material welfare included in all of the evaluations is an index of household wealth, based on households' ownership of assets and housing conditions. As well as recording this information at the time of the survey, respondents were asked to recall these details from a specific pre-project period, so the intervention and comparison groups can be compared in terms of the reported *change* in wealth indicators over time. The projects are estimated on average to have improved wealth indicators by 0.17 standard deviations, which is larger than the average project effect estimated from the household consumption data. There are important discrepancies between the consumption data and wealth indicators for the results of specific projects. One explanation for these discrepancies may be that the wealth index provides a longer-term measure of household welfare than consumption data, which may be more prone to fluctuations with short-term changes in income. However, this explanation has not been tested.

The results suggest that there is value in continuing to measure both income and wealth, but that there is further investigation to be done on the connections and complementarity between these measures. In future, efforts to move beyond aggregated household-level measures of welfare would help to understand intra-household project effects, and would allow for stronger differentiation and understanding of gender impacts of Oxfam's livelihoods projects.

Finally, the report reviews the 'programme learning considerations' that have been identified from the results of each of the evaluations. Eleven of the 23 evaluation reports recommended scaling up or replicating the project interventions, based on their positive findings. Most of the reports also included some constructive criticism. In a third of the reports, the evaluators were able to identify the stage of the expected chain of results at which the project logic broke down; in a few cases (particularly the more recent evaluations) they go on to suggest potential remedies. Only one report makes a fundamental criticism of the project logic (specifically, the contradiction between promoting livelihood diversification while also providing an incentive for producers to invest in a single product). Most of the reports (15 of the 23) make observations about the project structure, implementation or monitoring and evaluation arrangements that arose from the process of carrying out the evaluation, rather than from the results themselves.

1 INTRODUCTION

This meta-analysis considers the impact of all 23 projects that were evaluated by Oxfam for their impact on household livelihoods from 2011 to mid-2016. The aim of the meta-analysis is to examine what lessons can be learned – both about the impact of Oxfam projects and about the measurement approach used in the evaluations – by pooling data from across the 23 evaluations.

Five broad questions are addressed in this meta-analysis:

- What is the overall impact that Oxfam's projects have on the main global output indicator for livelihoods (total household consumption per adult equivalent per day)? Are there any systematic differences between Oxfam's impacts in certain geographical regions? Are there any contextual or project characteristics that explain differences in the effect size?
- Is it possible to identify different theories of change among the projects selected? If so, is there any difference in terms of impact on the main livelihood outcome indicator?
- What are the other most commonly used indicators measuring livelihoods? What is the overall impact on these indicators? Are there any contextual characteristics, project characteristics or theories of change that explain these differences?
- How are these other commonly used indicators correlated with the main global outcome indicator?
- What types of learning points commonly arise from the Effectiveness Reviews on livelihoods?

This paper is structured as follows. Section 2 describes the methodology and outcome measurements applied in the evaluations, and the methodology used in the meta-analysis. The main results of the meta-analysis – addressing the first bullet point in the list above – are then presented in Section 3. In order to respond to the second point in the list above, Section 4 goes on to review the evidence for effects on some intermediate outcomes considered in the various evaluations. The third and fourth points are discussed in Section 5, which considers alternative measures of welfare outcomes and their consistency with the household consumption measure used in the rest of the analysis. Section 6 analyses patterns in the learning points identified in the evaluation reports. Section 7 concludes with some observations on what can be learned from this set of evaluations as a whole.

2 DATA

2.1 SELECTION OF EVALUATIONS

The results analysed in this meta-analysis come from a series of impact evaluations (or 'Effectiveness Reviews') carried out each year by Oxfam GB on a sample of mature or recently closed projects. Since the start of this initiative in 2011, Oxfam GB has carried out 23 such evaluations of projects that were seeking to support household livelihoods.[1,2] The projects evaluated were selected largely at random from among all Oxfam GB's community-level projects that were seeking to support household livelihoods and that met a particular budget threshold.[3]

The 23 projects evaluated are described briefly in Appendix 1. Results from all of these evaluations are included in this meta-analysis. In contrast to a conventional meta-analysis, no further criteria were used to determine whether a study was eligible to be included in the meta-analysis. All 23 evaluations used a similar methodology, so criteria relating to the quality of the methodology would discriminate little between the studies. In addition, there is little potential for publication bias, since all the evaluations carried out since the start of the initiative in 2011 and completed by late 2016 are included in the meta-analysis.[4]

2.2 EVALUATION METHODOLOGY

Each of the impact evaluations involved carrying out a household survey, comparing outcomes for programme participants or beneficiaries to non-participants with similar baseline characteristics.[5] Surveys were carried out either towards the end of each project's implementation period, or up to two years after implementation ended. In order to minimize the possibility that the comparison respondents may have benefited indirectly from the project activities, in most of the evaluations the respondents interviewed for comparison purposes were sampled from different communities from those in which the project was implemented. In the five evaluations in which comparison respondents were sampled (wholly or partially) from the same communities as project participants, any indirect effects from the project on comparison respondents were thought to be minimal.[6]

The way in which comparison respondents were selected in each evaluation was intended to replicate, as far as possible, the methods that had been used to select the project participants or beneficiaries at the start of the project. The degree to which evaluations were successful in this respect varied. In many cases, the implementation sites (and often the individual project participants) were selected through an idiosyncratic targeting process that could not be replicated closely in the comparison areas. The consequence of this is that in many of the evaluations the possibility cannot be excluded that there are unobservable differences between the intervention and comparison groups that may bias the estimates of project effects upwards. However, the analysis discussed in Appendix 5 confirms that the meta-analysis estimates are robust in excluding evaluations in which there are particular concerns over the validity of the comparison group.

In each of the evaluations, propensity-score matching (PSM) was used to control for observable differences between the intervention and comparison groups. The approach used to implement PSM is described in Appendix 2. The variables used for matching include demographic characteristics, pre-project wealth indicators and, in some cases, indicators of pre-project engagement in particular livelihoods activities. Since pre-project data was not available in any of the evaluations, data on pre-project wealth indicators – and, in some cases, on pre-project sources of income – was based on information recalled by respondents during the single, post-

project, survey. The use of recalled data clearly creates potential for error. However, there is no evidence on the direction or size of such error, so it is not possible to determine whether this would be likely to result in project effects being under- or over-estimated.

It should be stressed that the evaluations included in this meta-analysis evaluated the impact only of project activities that were carried out at a household or community level. Many of the projects evaluated also sought to contribute to systemic change – for example, through advocating for changes in policy at a local or national level. The success of these higher-level activities is not assessed in the 23 evaluations, nor in this meta-analysis.

2.3 MEASUREMENT OF OUTCOMES

The evaluations included in this meta-analysis all sought to assess the extent to which the respective projects had enabled participant households to increase their income. Measuring income itself is often problematic in developing countries – particularly in rural contexts where much income is derived from agriculture and tends to be highly variable, both during the year and from one year to the next. For that reason, these evaluations did not attempt to collect data on income directly. Instead, the evaluations followed common practice in collecting data on household consumption (Deaton, 1997; Deaton and Zaidi, 2002). Consumption is usually thought to be closely correlated with household income, as well as being an important measure of welfare in its own right.

To construct the estimates of household consumption, survey respondents were asked for details of the food items that had been consumed in their household during the seven days prior to the survey[7] (including the approximate value of those items), as well as for details of the household's expenditure on non-food items (based on one-month, three-month or 12-month recall periods, as appropriate to the item in question). The household's total consumption was then calculated on a per-day, per-person basis, with an adjustment made to account for the lower consumption needs of children relative to adults and for economies of scale within the household.[8] This measure is referred to in this paper as consumption 'per adult equivalent'.

An alternative to the use of income or consumption data for measuring welfare is to examine indicators of household wealth. In each of the evaluations, data were collected on households' housing conditions, ownership of assets, and access to services, such as water and electricity. These data were then used to derive a wealth index, following the approach of Filmer and Pritchett (2001).[9] A wealth index is likely to be less sensitive than household consumption to small or short-term changes in economic status. However, the wealth measure has one important advantage in these evaluations: survey respondents were asked to provide information about those wealth indicators not only as applied at the time of the survey, but also to recall similar information from a notional pre-project period. This means that the wealth index can be calculated both for the pre-project period and for the time of the survey, allowing difference-in-difference analysis to be carried out. The impact of projects on the change in the wealth index, and the link between the wealth index and the household consumption measure, are discussed in Section 3.2 and Section 5.5 respectively.

Many of the evaluations also included some measures of household food security, or of the diversity of the household's diet. In addition, in several of the evaluations respondents were asked to provide a subjective assessment of their economic situation or of whether they had experienced an increase in income in recent years. The correlation of the dietary diversity, food security, and subjective measures with household consumption are discussed in Sections 5.3, 5.4 and 5.6.

Each of the evaluations collected data not only on these measures of overall welfare, but also on intermediate outcomes through which projects were seeking to have an effect. In each case, the evaluators and project implementers mapped out a logic model for the expected chain of results for each project, and identified indicators for some of the key steps in that chain. For

example, for projects that were supporting households' crop production, the evaluations generally collected data on investments in farming, the quantity of crops harvested and sold, and the revenue generated. Some of these measures of intermediate outcomes – particularly those related to agricultural production – were applied across multiple evaluations, so it is possible to carry out meta-analysis of the intermediate outcomes, and to examine how well they are linked with changes in welfare. This analysis is discussed in Section 4.

Finally, it is important to note that the projects evaluated were all seeking to have effects on outcomes other than simply household income or welfare. For example, many of the projects sought to change gender attitudes and promote women's empowerment, and some were designed to have effects on local-level governance. Most of the evaluations included assessment of projects' impacts on some of these complementary outcome measures. However, there is little commonality between the projects in the desired outcomes or in the indicators used to measure them, so these are not assessed in this meta-analysis.

2.4 META-ANALYSIS METHODOLOGY

The approach to meta-analysis adopted in this paper follows the guidance provided by Borenstein *et al.* (2009), Higgins and Green (2011) and Waddington *et al.* (2012).[10] In order to compare and aggregate results across the various evaluations, estimated project effects are expressed as a proportion of the standard deviation of the outcome measure in question.[11,12] The outcomes in this paper are therefore reported in terms of Cohen's *d*, a measure of standardized mean difference between groups. The interpretation of standardized effect sizes may not be intuitive, but it may be helpful to be aware that Cohen (1992) characterized standardized effects of 0.2, 0.5 and 0.8 as being 'small', 'medium', and 'large' respectively, with a 'medium' effect being one that is 'likely to be visible to the naked eye of a careful observer' (p. 156).

One disadvantage with the use of a standardized outcome measure is that the effect sizes appear relatively larger in datasets that have lower variation in the outcome measure. To check for robustness of the results, the main meta-analysis discussed in Sections 3.1 and 3.2 have also been conducted using non-standardized estimates of project effects on the logarithm of household consumption (which can be approximately interpreted as percentage differences in consumption between the intervention and comparison groups). The patterns of results derived in this way are very similar to those reported in Sections 3.1 and 3.2, and the qualitative conclusions are not affected.

The meta-analysis models are applied with study-level random effects, an approach that takes into account that the effect of the projects on the outcome measures may vary between contexts. This seems appropriate, given the wide variation in the types of interventions evaluated and the various environments in which these interventions were carried out. A test of this assumption is provided by the I^2 statistic, which represents the proportion of the variation between studies that is attributable to heterogeneity. For example, in Figure 1 below, the high value for the I^2 statistic of 73% (reported in the last line of the chart) provides strong evidence that the projects had heterogeneous effects on household consumption.[13]

The analysis in this paper follows common practice in weighting each evaluation according to the inverse of the variance of its outcome estimates, modified to account for the heterogeneity between studies. The sample sizes used in the 23 evaluations are of a similar size, so the standard errors of the outcome estimates tend to be reasonably homogeneous: this leads to each evaluation being given approximately equal weight in the meta-analyses. (The weights allocated to each evaluation in the meta-analysis of project effects on household consumption can be seen on the right-hand side in each of the results plots.) However, the projects evaluated varied widely in scale, as can be seen in Appendix 1. The consequence of this is that the aggregate estimates derived from the meta-analysis do not represent the effect on the average

project participant household.[14] However, as reported in Section 3.1, there is no evidence of a relationship between the scale of a project and the size of its effect on participant or beneficiary households. The meta-analysis is therefore considered to provide a reasonable guide to the size of the average effect across the 23 projects as a whole.

Duvendack *et al.* (2012) provide a warning about including quasi-experimental studies – of which Oxfam's Effectiveness Reviews are an example – in a meta-analysis. They have three concerns: firstly, that the imperfect identification strategies used in quasi-experimental methods may bias the conclusions; secondly, that there may be heterogeneity between studies in the methodology applied in different studies and the treatment effects being estimated; and thirdly, that 'researcher allegiance' can lead to publication bias or other forms of positive bias in results. The second of these concerns does not apply in our case, since all the evaluations applied a common PSM approach, and they all estimate the average treatment effect on the treated. However, the other two points warrant some discussion.

The concern about bias in the underlying evaluations is clearly relevant to this meta-analysis: the results discussed here are valid only insofar as the identification assumptions made in each of the underlying evaluations are valid. As discussed in Section 2.2, the evaluation teams attempted to select the intervention and comparison observations in a way that would minimize both observable and unobservable differences between them. Remaining observable differences were controlled for using PSM models at the analysis stage, but the extent to which they were successful in controlling for unobservable characteristics cannot be known. However, the analysis discussed in Appendix 5 confirms that the meta-analysis results are robust to excluding evaluations that have particular concerns around the validity of the comparison group.

With respect to the third of Duvendack *et al.*'s concerns, this meta-analysis is unlikely to be affected by publication bias for the reason discussed in Section 2.1. Whether there is potential for the results to be affected by subtler forms of 'researcher bias' is more difficult to assess. However, despite being employed by or contracted by Oxfam, the evaluators have a high degree of autonomy within the organization and have frequently published results that are less favourable than programme managers may have hoped.

What can be said with confidence is that the Effectiveness Reviews provide the most robust data available on the impacts of Oxfam GB's projects on household livelihoods. This meta-analysis therefore represents the organization's best attempt to systematically analyse and learn from those impacts.

3 META-ANALYSIS RESULTS

In this section, the main results of the meta-analysis are presented. As discussed in Section 2.3, the key outcome indicators included in each of the evaluations are, firstly, household consumption, and secondly, indicators of household wealth status (asset ownership and housing conditions). The average impact that the projects evaluated had on both of these outcome measures is examined, as well as whether those patterns vary by region, by project characteristics, by the gender of the household head, or by the household's economic position.

3.1 OVERALL IMPACT ON HOUSEHOLD CONSUMPTION

Figure 1 shows the results of a meta-analysis for the estimated impact of the 23 different projects on the income of participant households. For each evaluation, the difference between intervention and comparison households in household consumption is estimated through PSM. In Figure 1 and other plots in this paper, the estimated effect size of each evaluation is shown as a point, with the horizontal bars representing the corresponding 95% confidence interval. The diamond shapes represent the 95% confidence intervals for the average effects across evaluations, aggregated through meta-analysis.

Overall, the 23 projects are estimated to have resulted in an increase in the consumption of participant households of 0.12 standard deviations on average, with a 95% confidence interval ranging from 0.03 to 0.20 standard deviations. Carrying out the equivalent meta-analysis using non-standardized project effect estimates, the projects are found to have increased the consumption of participant households on average by approximately 6.6% (with the associated 95% confidence interval ranging from 1.6% to 11.9%).

The results in Figure 1 are divided into four regions of the world (Africa, Asia, the Caucasus, and Latin American and the Caribbean), but the estimated average effect on consumption is of a similar size across each of the four regions.[15] There is no evidence of a difference in project effects between lower-income and middle-income countries.[16]

As discussed in Section 2.4, the projects evaluated varied widely in scale, but they are each given approximately equal weight in the meta-analysis. Weighting the projects equally in this way may seem counter-intuitive; in particular, it may be objected that projects that have focused their resources on working with a small number of participants are likely (all else being equal) to achieve greater impact on the average participant than those that have sought to work with a much larger number of participants. In fact, the data do not provide evidence for any such relationship between the scale of the project and the effect size. This can be seen in Figure 2, which shows the same data as Figure 1, but with projects ordered by size. The results for smaller projects (at the top of the chart) are no more positive on average than those for larger projects (lower down in the chart).[17] Furthermore, there does not seem to be any relationship between the amount invested by Oxfam in each project (either in total or in per-household terms) and the size of its impact.[18]

It can be seen from the list of projects in Appendix 1 that there was also considerable variation in the duration of the projects evaluated. It may be natural to expect that projects of a longer duration would have greater impact. However, our data do not provide evidence that this is the case.[19]

One way to interpret the size of the aggregated project effect – that the 23 projects have increased household income by 0.12 standard deviations on average – is to compare it with the

effects found in studies of other livelihoods interventions. There do not appear to be any organizations that have carried out a process of aggregating data on the impact of a sample of livelihoods projects that can be compared directly to the approach in this paper. However, reviews of interventions that have elements in common with the livelihoods projects evaluated here do exist. In particular, a systematic review of agricultural certification schemes – such as Fair Trade and organic labelling – found a positive (but not statistically significant) effect on income for producer households of 0.13 standard deviations, equivalent to a 6% increase (Oya *et al.*, 2017). An aggregated analysis of 'graduation' schemes for ultra-poor households in six countries found a positive effect on consumption of 0.12 standard deviations, when measured immediately after the end of the intervention or 12 months later (Banerjee *et al.*, 2015).[20] These results suggest that the impact achieved by Oxfam's livelihoods projects is broadly in line with the results found in studies of somewhat comparable interventions.

Figure 1: Project effects on household income by region and country (on logarithm of household consumption per adult equivalent per day)

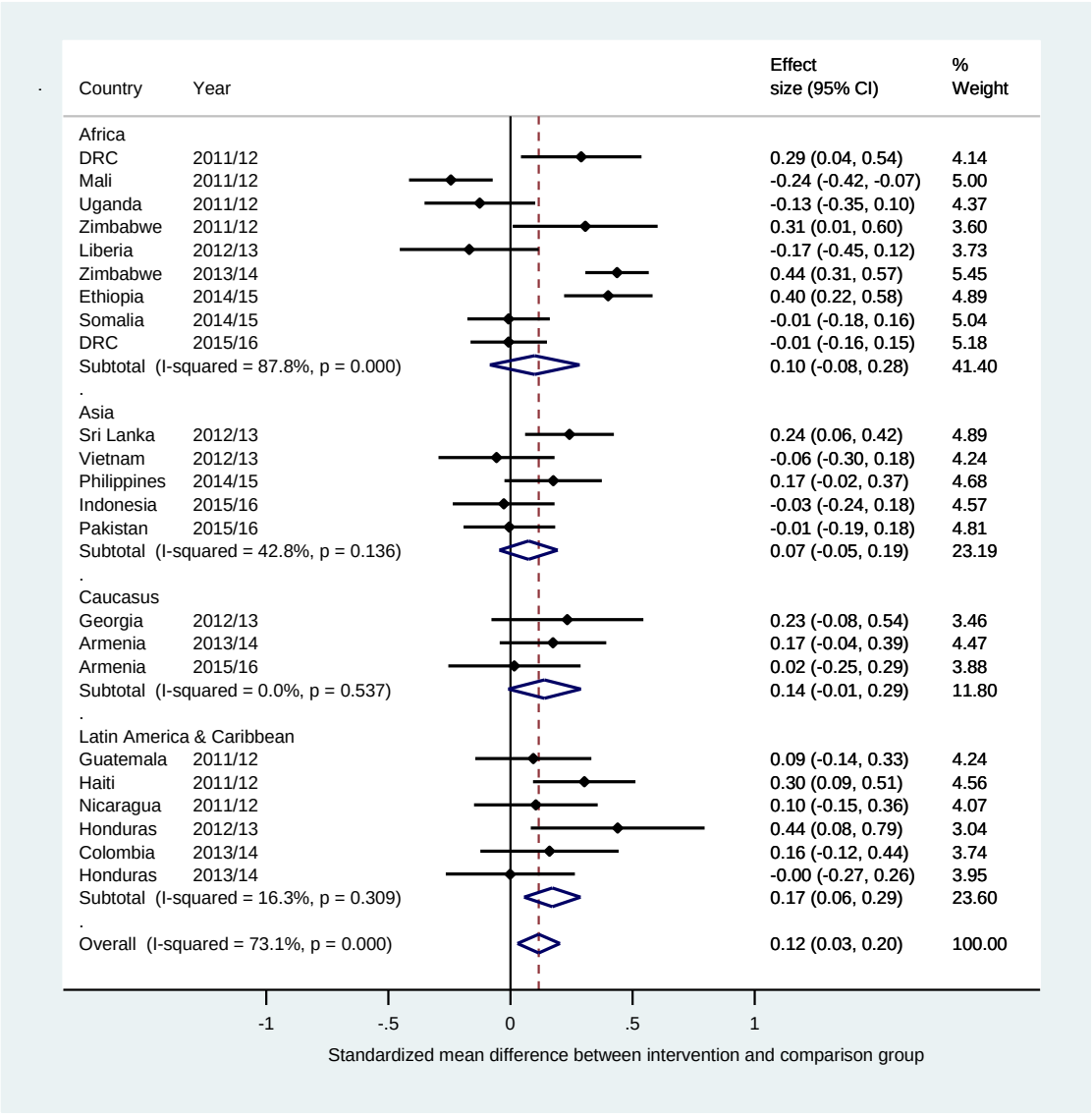
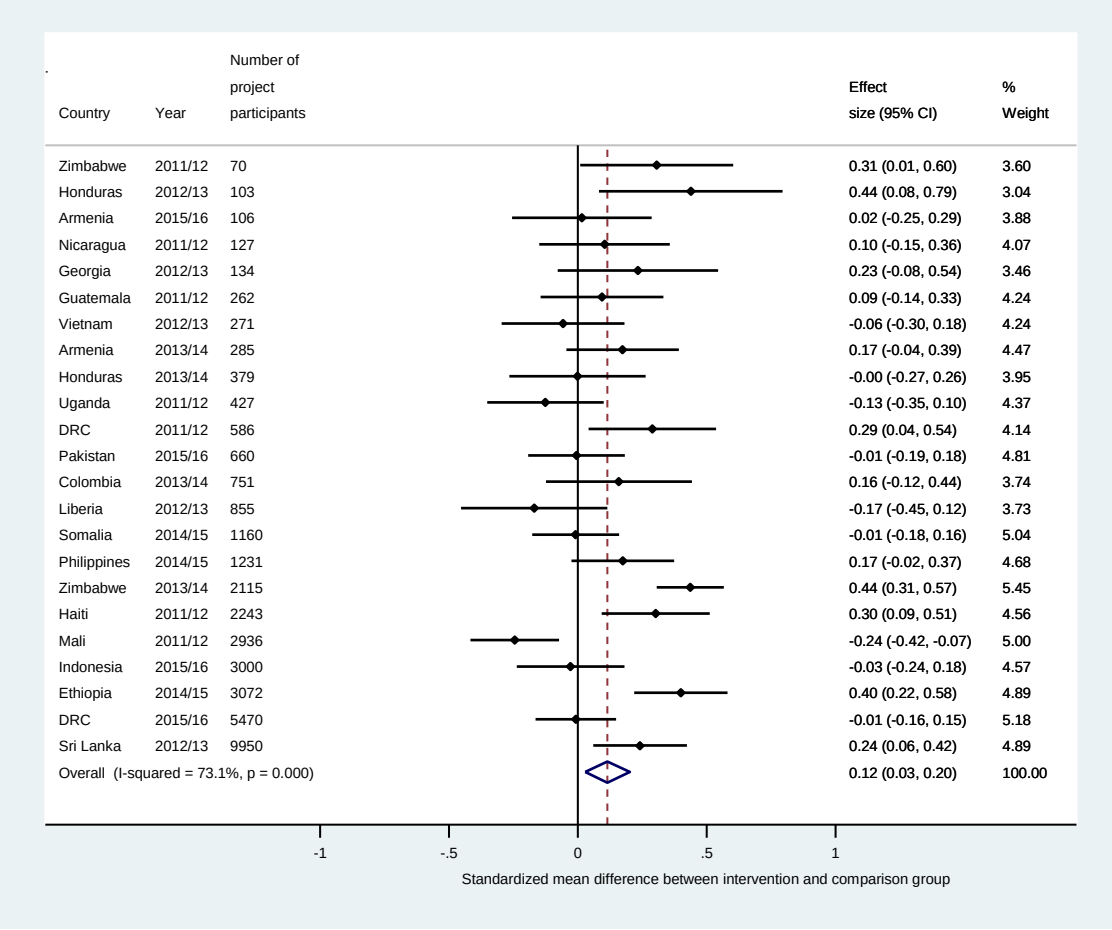


Figure 2: Project effects on household income by scale of project (on logarithm of household consumption per adult equivalent per day)



3.2 OVERALL IMPACT ON INDICATORS OF HOUSEHOLD WEALTH

As discussed in Section 2.3, each of the evaluations collected data not only on household consumption, but also on indicators of household wealth – including ownership of assets, housing conditions, and access to services such as water and electricity. Respondents were asked to provide information about these characteristics not only at the time of the survey, but also to recall them from a specific pre-project period. This allows us to compare the intervention and comparison groups in terms of the (recalled) change in the resulting index of household wealth over time – that is, using a difference-in-difference measure.

It should be noted that several of the 23 projects involved distributing assets to some or all of the project participants. For example, many of the participants in the project in Georgia received donations of livestock under the project, while those in the first project evaluated in the Democratic Republic of Congo (DRC) received fishing nets or equipment for preparing fish for sale. If these assets were still in participants' possession at the time of the survey, then they would have directly resulted in an increase in the wealth index among the project participants. However, the types of assets that were distributed under each of the projects make up only a small proportion of the wealth indicators included in the index.

The results of the meta-analysis of the wealth index are shown in Figure 3. It can be seen that the projects overall are estimated to have had an effect on the change in wealth index of 0.19 standard deviations. This is substantially larger than the effects of the projects that was

estimated from the household consumption data. Figure 4 goes on to compare the point estimates of the results derived from the consumption data and the wealth index for each of the 23 evaluations. It is clear that there is an overall correlation between the two measures: a project's impact on consumption explains approximately a quarter of the variation of its impact on the wealth index, and vice versa.[21] However, the results for specific projects can appear quite different, depending on which of the measures is examined. Specifically, the wealth indicators show a much smaller effect from the projects in Zimbabwe (in 2013/14), Ethiopia, the Philippines, Guatemala, and Haiti (the latter two of which have estimated effects close to zero in terms of the wealth index). On the other hand, the wealth indicators provide much stronger evidence of a positive effect from the projects in Somalia, DRC (in 2015/16), Pakistan, Nicaragua, Colombia, and Honduras (in 2013/14), as well as reducing the size of the negative estimate in Mali.

The discrepancies between the results derived from the consumption data and those derived from the wealth indicators have been noted in most of the corresponding evaluation reports. One possible explanation for the divergence is that wealth indicators generally provide better evidence of a household's long-term economic situation, whereas consumption measures are more subject to shorter-term fluctuations. If so, then projects that are found to have had a positive effect on consumption but not on wealth indicators, should be understood as having resulted in recent increases in welfare, but that it is not yet clear whether these will be sustained over time. In contrast, if a project is seen to have had a positive effect on wealth indicators but not on consumption, this may imply that the positive effects of the project are confined to the past. However, this conclusion is at odds with what is known of some of the projects (such as those in Nicaragua and Colombia), where the project activities were still in the process of scaling up, and should be expected to have had greater impact on welfare at the time of the survey than in previous years. The connection between consumption measures and wealth indicators is discussed further in Section 5.5.

As in the previous section, these results can be compared to those found in studies of some similar interventions. The aggregation of the results for the ultra-poor 'graduation' schemes found an effect on asset wealth of 0.26 standard deviations when measured at the end of the intervention period, or 0.25 standard deviations 12 months later (Banerjee *et al.*, 2015) – that is, a similar magnitude to the average effect estimated for the Oxfam livelihoods projects.[22] The review of results for agricultural certification schemes found a (non-statistically significant) effect on wealth indices of 0.05 standard deviations, but this result is an average of only two underlying studies that provided such data.

Figure 3: Project effects on wealth by region and country (on change in index of wealth indicators)

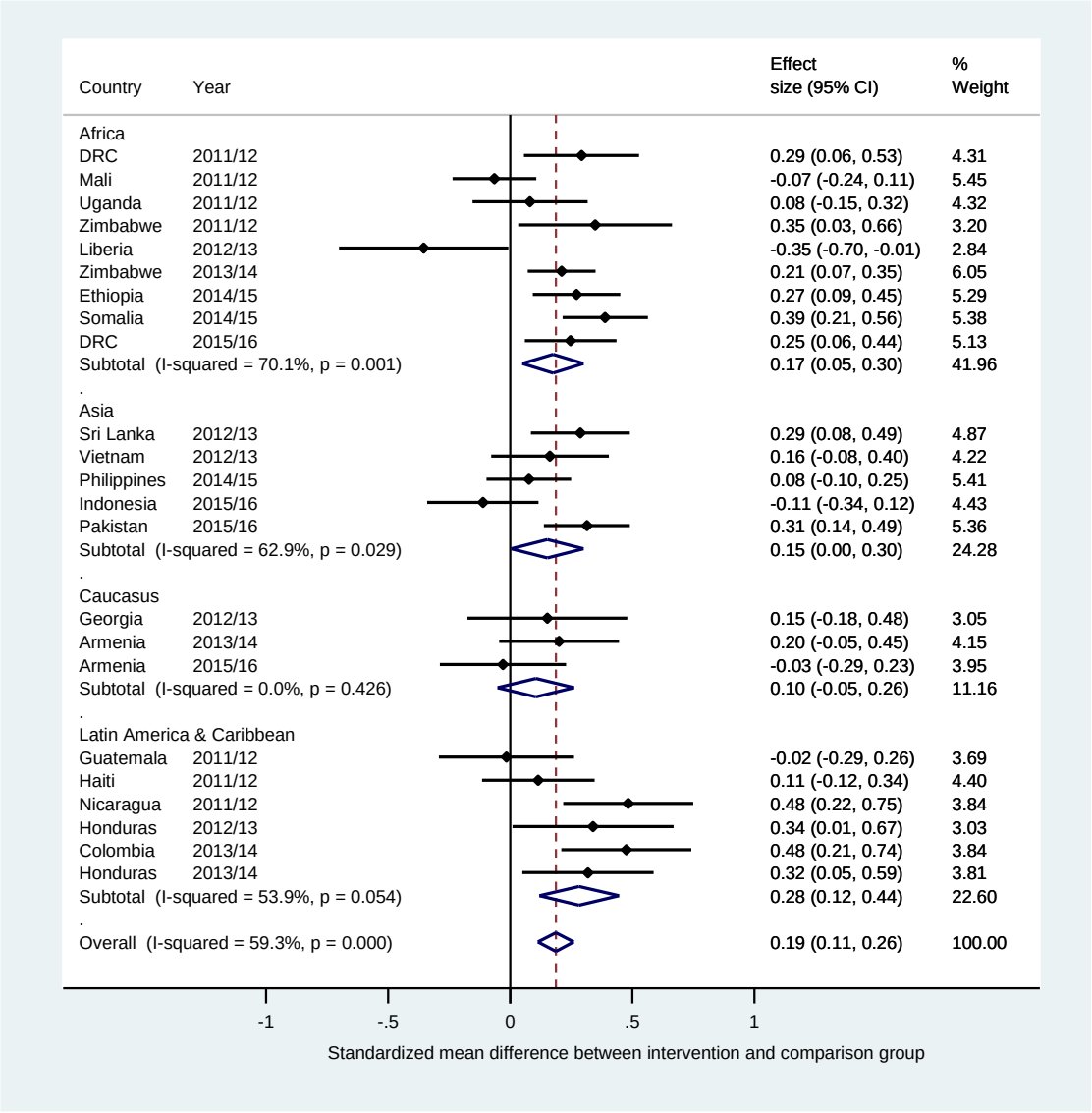
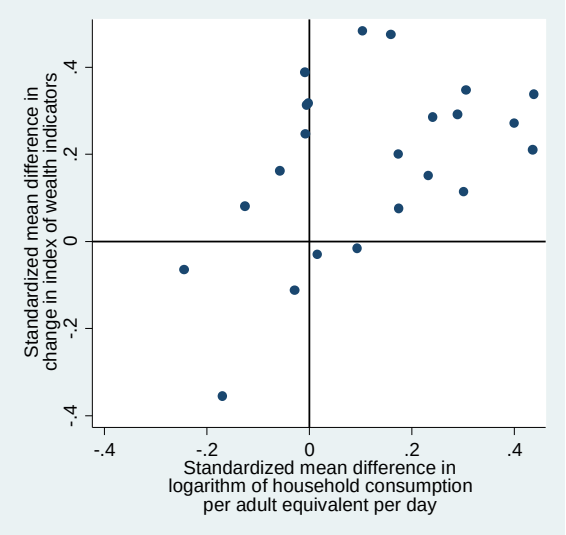


Figure 4: Comparison of project effects on consumption and wealth



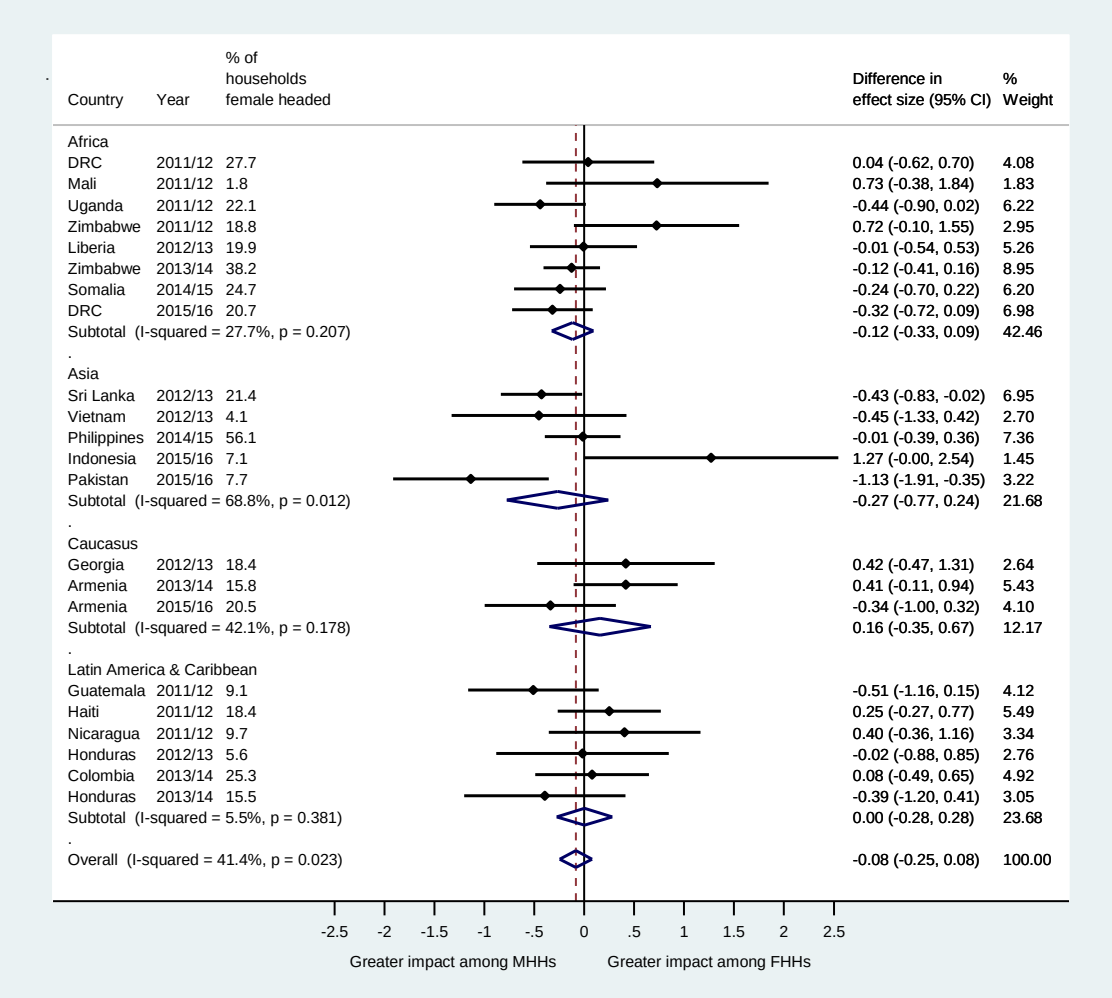
3.3 DIFFERENCES IN EFFECT SIZE BY GENDER OF THE HOUSEHOLD HEAD

Oxfam seeks to promote women's rights and women's empowerment throughout its programmes. Several of the projects evaluated included components that specifically involved working with women. The dependence on household-level data for these evaluations means that it is not possible to differentiate the effects that projects had on women and men as individuals. However, it may be instructive to examine the effects on female-headed households against male-headed households.

This question is analysed in Figure 5, where the results represent the estimated difference between female-headed and male-headed households in the size of projects' effects on household consumption.[23] The decision as to which household member to identify as the head was made by the respondent at the start of each interview.[24] Overall, female-headed households are estimated to have benefited less from the projects than male-headed households, but this difference in not statistically significant.

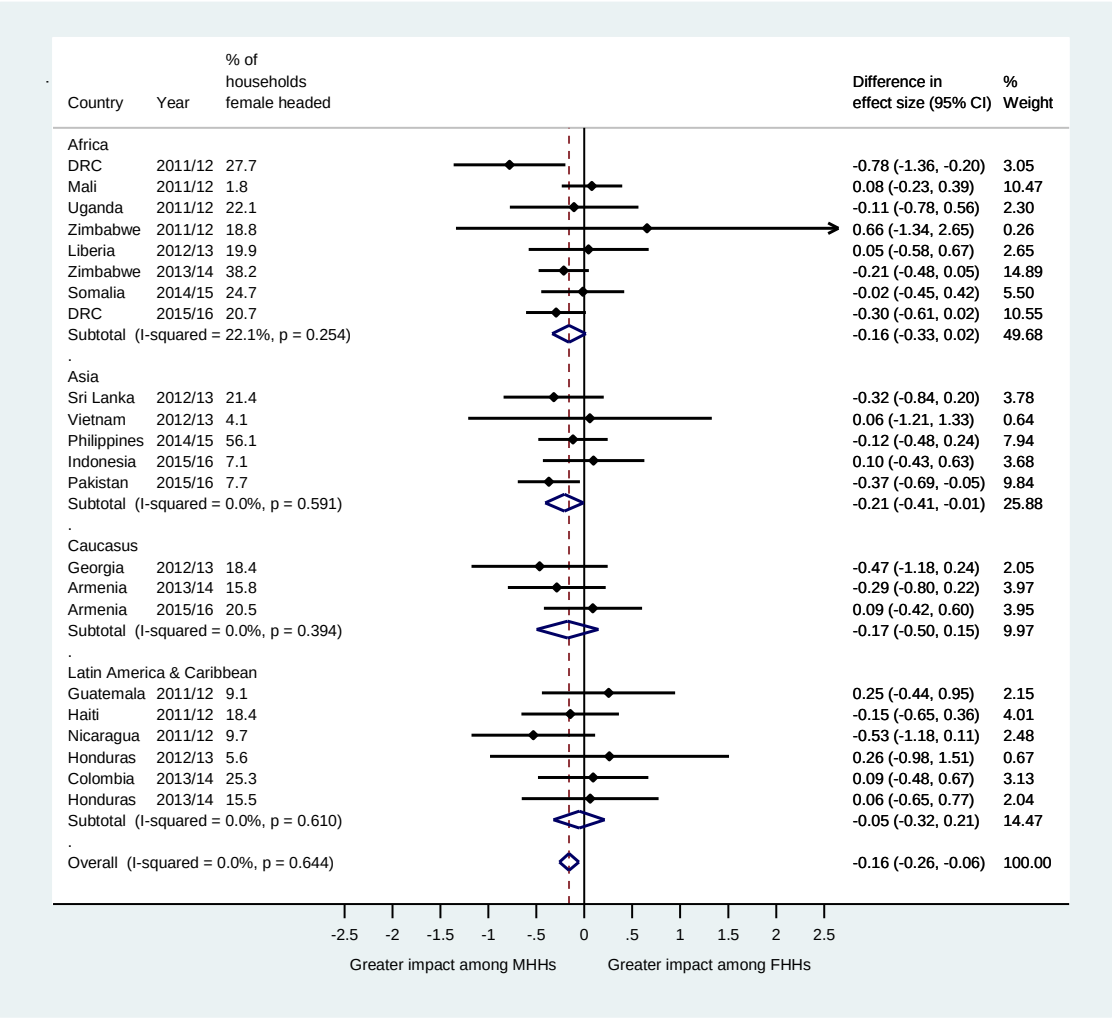
The same analysis is repeated in Figure 6, but this time with the change in the index of wealth indicators as the outcome variable. Using this measure, the difference between female-headed and male-headed households in the effects of the projects is larger – at 0.16 standard deviations – and statistically significant. The low value of the I^2 statistic (zero) implies that this difference is consistent across the various projects.[25]

Figure 5: Project effects on household income by gender of household head (on logarithm of household consumption per adult equivalent per day)



One factor to consider is whether the smaller project effects among female-headed households is related to their relative poverty.[26] However, the effect observed in Figure 6 is changed little even after controlling for the households' pre-project wealth levels. The fact that female-headed households appear consistently to benefit less from livelihoods projects than do male-headed households, according to one of the two main measures in which outcomes are analysed, should therefore be of concern.

Figure 6: Project effects on wealth by gender of household head (change in index of wealth indicators)



3.4 DIFFERENCES IN EFFECT SIZE BY PRE-PROJECT WEALTH LEVEL

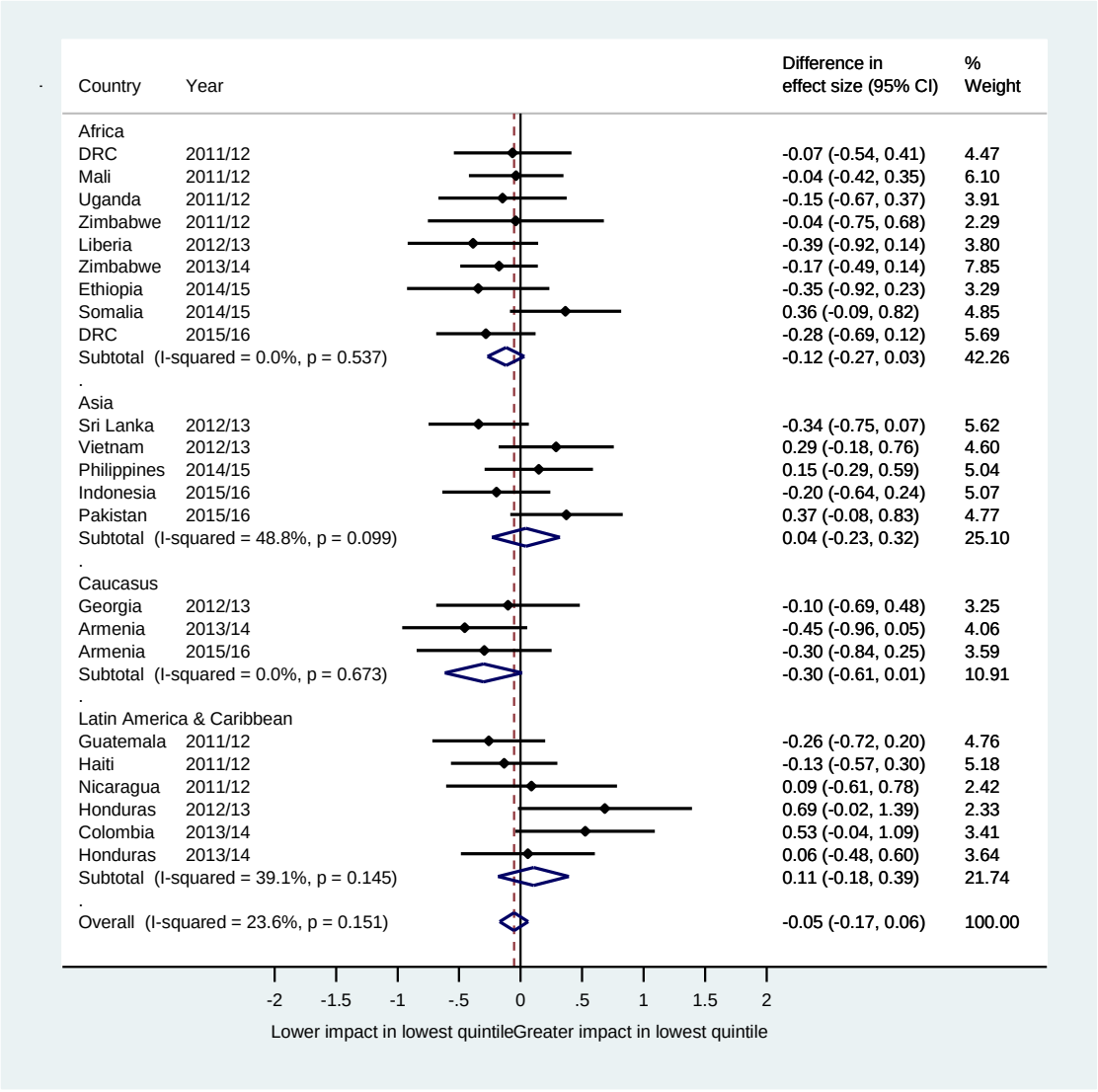
The projects evaluated were all implemented in geographic areas and in communities that were identified as poor and vulnerable. Furthermore, many of the projects evaluated sought to target particularly poor households to work with. Nevertheless, there was some variation in economic levels across the group of participant households in each of the projects evaluated. This allows for investigation of whether households that were initially poorer benefited more from these projects.

To that end, project participants were divided into quintiles, based on their recalled pre-project wealth indicators (asset ownership and housing conditions), using the approach described in Section 2.3.[27] The effect of each project on households in each of the quintiles was then compared to the effect of households in the other four quintiles. For example, Figure 7 shows the meta-analysis for the differential effect on households in the lowest pre-project wealth quintile compared to households in the other four quintiles. Overall, households in the lowest quintile do not appear to have benefited either more or less than households in the other quintiles. The same applies when impact among households in each of the other four quintiles is considered. The forest plots relating to the other four quintiles are not included here, but the summary of the meta-analysis for each is shown in Table 1. All of the meta-analysis estimates in Table 1 are close to zero, suggesting that the 23 projects evaluated had approximately equal effects across the pre-project wealth distribution. Similar results are obtained when carrying out this analysis in terms of the wealth index rather than in terms of household consumption.

Table 1: Project effect on household income by pre-project wealth quintile (logarithm of household consumption per adult equivalent per day)

Quintile of pre-project wealth index	**Differential project effect on household consumption per adult equivalent (standardized)**
Lowest	-0.05 (-0.17, 0.06)
Second	0.04 (-0.07, 0.14)
Third	0.05 (-0.04, 0.15)
Fourth	0.00 (-0.10, 0.10)
Highest	0.01 (-0.14, 0.17)

Figure 7: Project effects on household income by pre-project wealth quintile (logarithm of household consumption per adult equivalent per day)



4 EFFECTS ON INTERMEDIATE OUTCOMES

The 23 evaluations collected data not only on consumption and other indicators of welfare, but also on intermediate outcomes that the projects were seeking to affect. By examining the evidence for a positive effect from projects on these intermediate outcomes, the evaluators sought to understand the causal chains by which projects did or did not produce an effect on households' overall welfare.

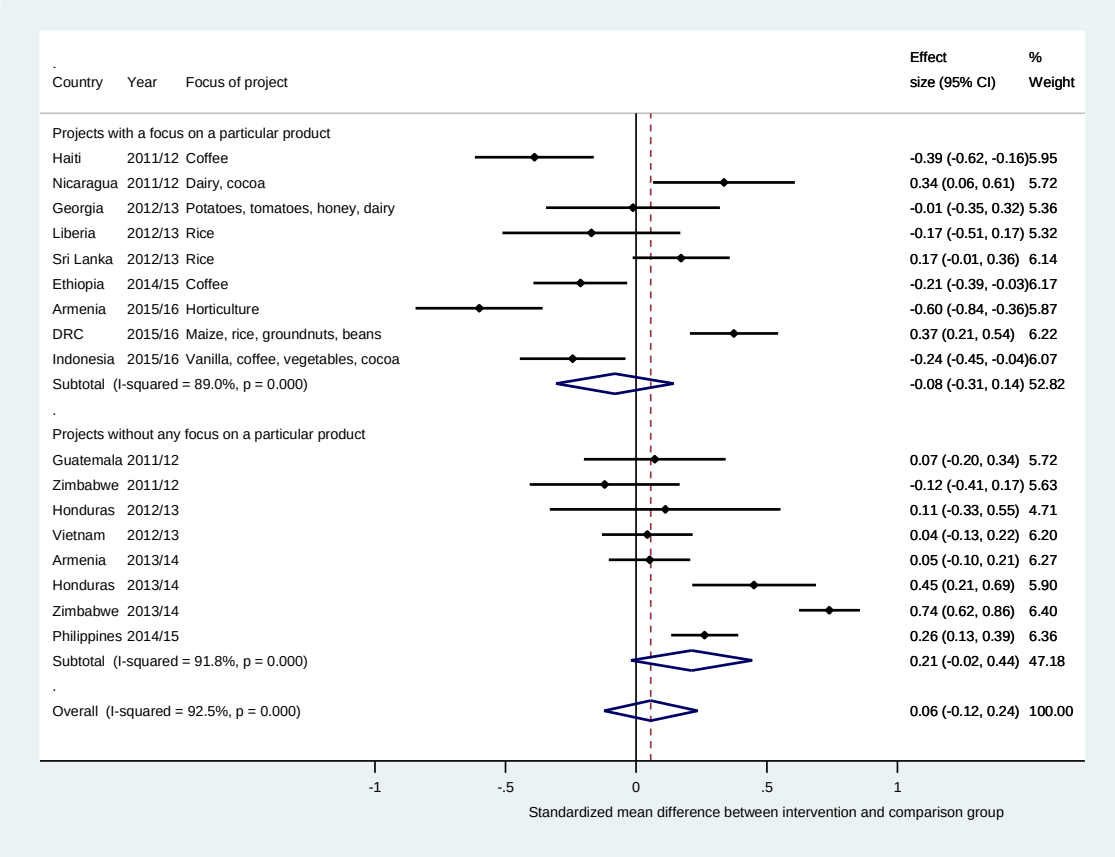
Most of the evaluations collected data on the proportions of households that said they had received the types of support or been involved in the types of activities carried out under the project. For example, they may have been asked about whether they had received distributions of agricultural inputs or training or support in agricultural production. In most cases, the evaluations found clear differences between the project participants and comparison respondents in these respects; this acted as confirmation that the activities of the projects reached the intended beneficiaries and that there was little evidence of the project activities having produced indirect benefits within the comparison group. Since the specific variables analysed differed across the evaluations, there is little potential for carrying out meta-analysis on these first-level indicators of project effects.

Where there is more potential for meta-analysis to provide insight is in terms of the effects of those project activities on other intermediate outcomes. In particular, most of the projects (with the exception of those evaluated in Pakistan and in DRC in 2011/12) were concerned wholly or partly with supporting crop production, so it is possible to analyse agricultural outcomes across the datasets. (In contrast, the other sectors in which projects have focused their activities – such as livestock rearing, non-agricultural household businesses, or access to finance – involved only two or three projects each, so meta-analysis of the results would not be meaningful.)

One useful indicator of the impact of the projects on households' agricultural activities is the diversity of crop types cultivated. Figure 8 shows the estimated effect of each project in terms of this measure, for all the evaluation datasets for which data on crop diversity was collected. In this figure, the projects are categorized by whether they were supporting participants in production of a specific type of agricultural product (in the upper half of the figure), or supporting agricultural production in general (in the lower half).

Figure 8 shows that the projects overall have not had any clear effect on the range of crops grown by participant households. However, there is some indication of a difference between those projects that focused on a specific crop type and those that did not. In some cases, projects with a focus on a specific product seem to have had the effect of reducing participant households' investment in other agricultural products; examples are the projects in Haiti and Ethiopia, both of which focused on supporting production and sales of coffee. (In fact, the most recent phase of the project in Haiti had sought to promote crop diversification; in this respect, it did not appear to have been successful by the time of the evaluation.) On the other hand, the projects in Nicaragua and DRC appear to have had a positive effect on the diversity of crop types produced. There is also a clear differentiation in terms of agricultural projects that were *not* focused on specific crop types: three appear to have had a positive effect on crop diversity, while the other five did not.[28]

Figure 8: Project effects on the number of crop types produced by the household



Several of the evaluations collected data on the quantity of crops harvested or on the volume of other agricultural products produced, but it is, of course, difficult to aggregate production volumes across multiple crop types. However, meaningful analysis can be carried out on the *sales* of agricultural products, since these are all recorded in monetary terms. Figure 9 shows a comparison between the proportions of the intervention and comparison households that reported having made any sales of agricultural products during the year prior to the survey. Some of the projects had a clear, positive effect on whether households had made any sales, while others did not. In most of the cases in which there was no apparent effect on the proportion of project participant households engaging in sales, this was because high proportions of all households surveyed (including those in the comparison group) were engaged in sales of their products, so there was little potential for the projects to have an effect in this respect. (The main exceptions are the projects in Liberia and Vietnam, in which only around half of households were selling agricultural products, but where the projects appear to have had little or no effect.)

Figure 10 goes on to consider the total value of sales of agricultural products reported by households during the year prior to the survey. In general, these data were collected only in evaluations of projects that had an emphasis on increasing sales revenue – so the results for the 13 projects shown in Figure 9 are unlikely to be representative of the 10 projects for which that data was not collected.

It is important to emphasize that the figures used for the analysis in Figure 10 represent gross revenue from sales: that is, they do not account for costs of production or costs involved in making the sales themselves. It is likely that households that made greater sales incurred greater costs in doing so. An illustration of this is that the projects seen in Figure 10 to have had clear, positive effects on sales (those carried out in Georgia, Sri Lanka, Ethiopia, and Zimbabwe) generally had smaller effects on household consumption. Overall, though, the size of the effect on agricultural sales among the 13 projects shown in Figure 9 of 0.13 standard deviations, is similar to the size of the effect on household consumption among those 13

projects, also estimated at 0.13 standard deviations. This is reassuring in that it implies that the extra sales being made are generally profitable for participant households.

Figure 9: Project effects on agricultural sales (the proportion of households engaging in the sales of any agricultural products)[29]

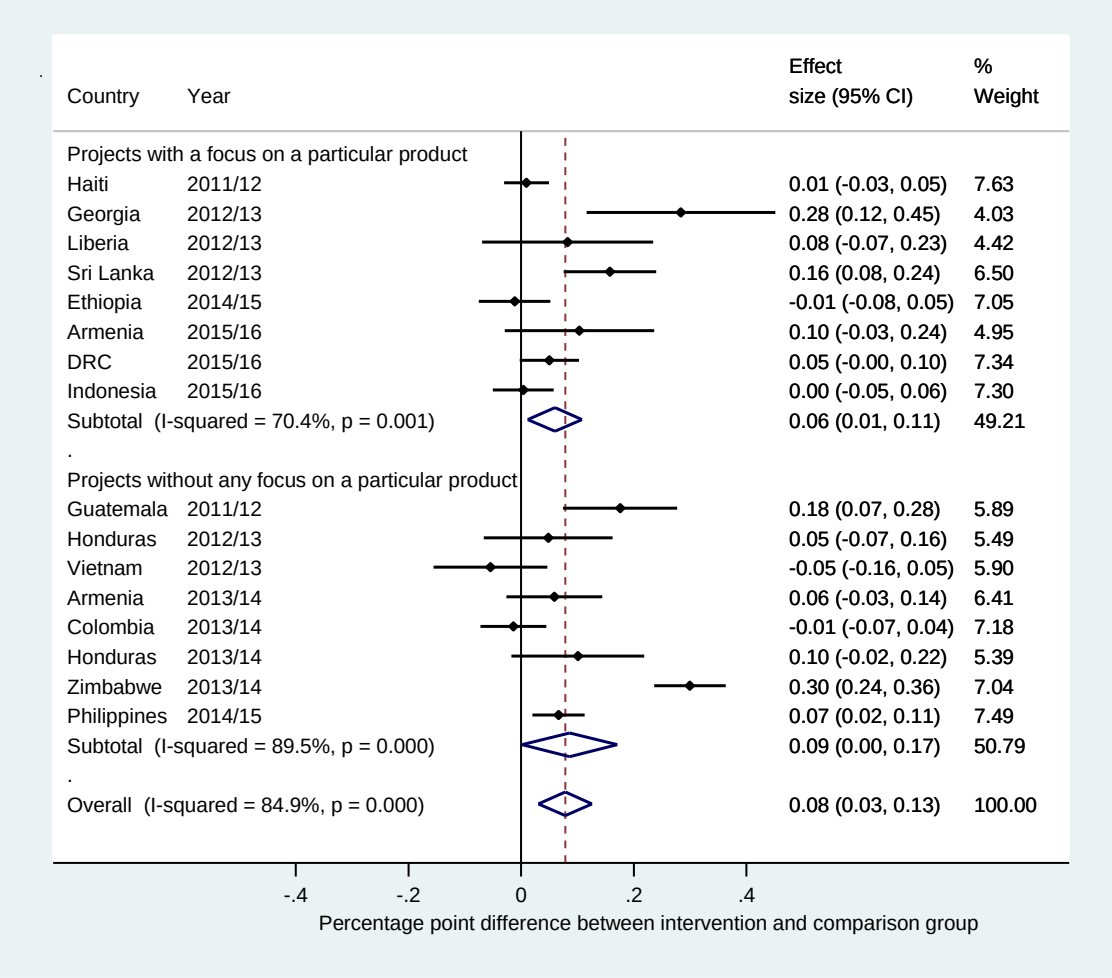
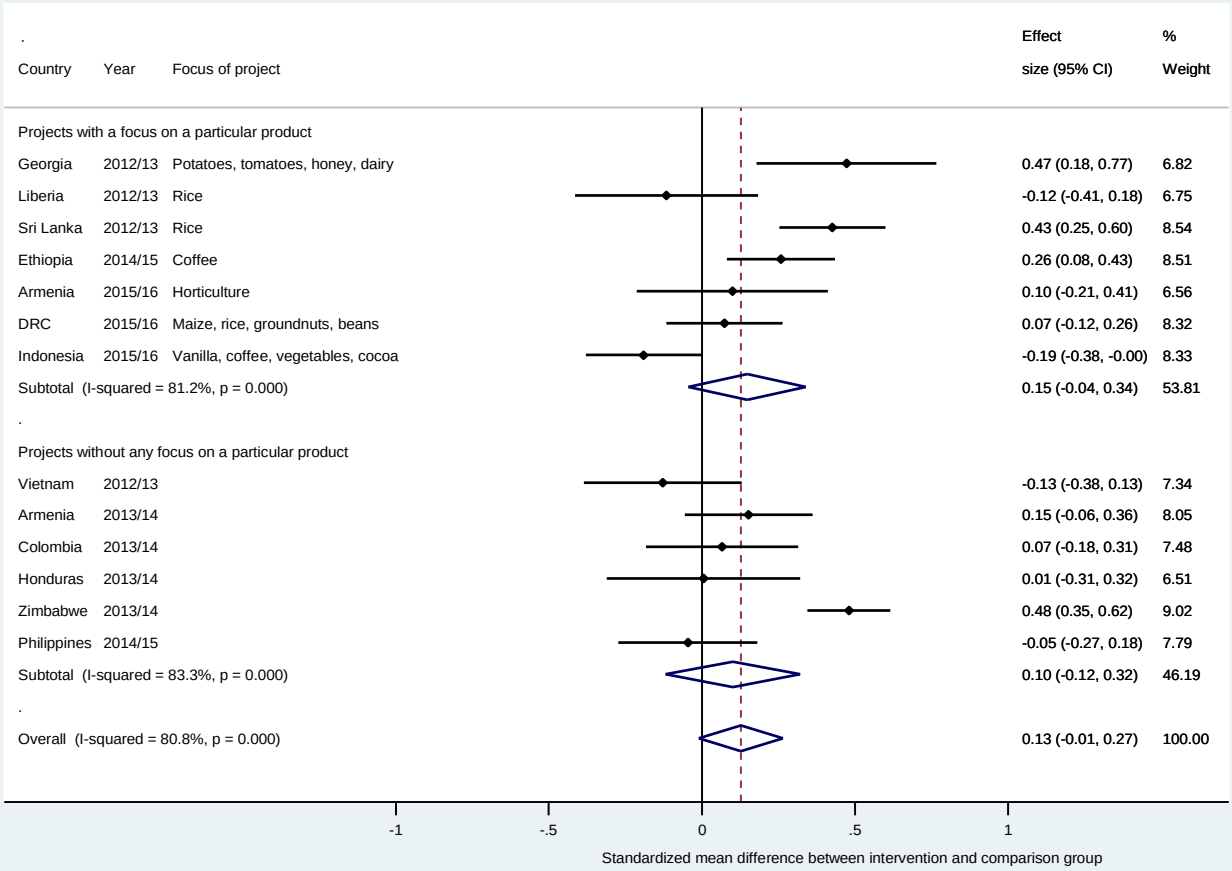


Figure 10: Project effects on the value of agricultural sales (sales of agricultural products in the year prior to the survey)



5 EXPLORING MEASUREMENT APPROACHES

This meta-analysis provides an opportunity to investigate the robustness and consistency of the various outcome measures used in the 23 evaluations. The key results presented in Section 3 of this paper were based on data on overall household consumption and the index of wealth indicators. This section examines how the results of the evaluations change when alternative outcome measures – including alternative measures of consumption, but also measures of food security and subjective welfare assessments – are considered.

In the following discussion, the various outcome measures are compared in two respects: firstly, the extent to which the measures agree in their assessment of the welfare of any particular household, and secondly, the extent to which they agree on the estimated impact of the projects evaluated. The latter assessment is made by calculating project effects using the same PSM models used to derive the results in Sections 3 and 4. How to assess the agreement between outcome measures in their measurement of each household's welfare is more complicated. One approach would be to examine the correlation coefficient between any two outcome measures. However, correlation coefficients do not take into account differences in the distribution of those measures and hence can obscure systematic bias (Howe *et al.*, 2008). Instead, the observations in each dataset are divided into quintiles according to each outcome measure and examine whether the different outcome measures agree as to the quintile in which households are placed. In particular, Cohen's kappa statistic (Cohen, 1960) is used to quantify the extent of this agreement. The kappa statistic takes its maximum value of one when the outcome measures are in complete agreement as to the allocation of households between quintiles, and zero when the measures are in no more agreement than would be expected by chance.[30]

5.1 HOW SENSITIVE ARE THE RESULTS TO HOUSEHOLD EQUIVALENCE SCALES?

As discussed in Section 2.3, the measure of household consumption used in the 23 evaluations was derived by converting total household consumption into a 'per adult equivalent' figure. This conversion accounts for children having lower consumption requirements than adults, and for economies of scale in the household. Of course, the specific assumptions made in that calculation are open to question. In Table 2 below, four alternative means of calculating the household consumption aggregate are compared with the standard method. All the consumption measures that control in some way for the size of the household can be seen to be closely correlated: the majority of respondents are allocated to the same quintile of the distribution whichever measure is chosen, and the kappa statistics are around 0.8 or higher. Agreement is weaker between those measures and the total consumption of the household (that is, without accounting for household size).

The right-hand column of the table shows that the choice of denominator for the consumption aggregate makes little difference to the overall effect estimated through meta-analysis. In fact, changing the denominator has little effect on the estimated effect size of the results from any of the individual evaluations, or to their statistical significance. This is a natural consequence of the fact that the PSM models used to estimate these effects mostly used household size and composition as matching variables, so that the intervention and comparison groups in each evaluation are generally well balanced in terms of these characteristics. Nevertheless, it will be

reassuring for future evaluations to be aware that the estimates of project effects are not sensitive to the choice of denominator.

Table 2: Comparison of project effects across alternative measures of household consumption

Denominator of household consumption measure	Denominator in calculation of consumption aggregate	Agreement with standard per adult equivalent consumption measure:		Estimated effect size (95 per cent confidence interval)
		Proportion of households in the same quintile	Kappa statistic (95 per cent confidence interval)	
Per adult equivalent	$(A + \frac{1}{3}K)^{0.9}$	–	–	0.12 (0.03, 0.20)
Per adult equivalent (modified)	$(A + \frac{2}{3}K)^{0.9}$	79%	0.87 (0.86, 0.88)	0.11 (0.02, 0.20)
Per adult equivalent (modified)	$(A + \frac{1}{3}K)^{0.5}$	70%	0.80 (0.79, 0.82)	0.13 (0.04, 0.22)
Per person	$A + K$	64%	0.76 (0.74, 0.78)	0.10 (0.01, 0.18)
Per household	1	47%	0.57 (0.54, 0.60)	0.14 (0.05, 0.22)

A represents the number of adults in the household, and *K* the number of children (defined as those under 16 or 18 years of age, depending on the dataset).

Table 3: Agreement between outcome measures

Outcome measure	Agreement with total consumption per adult equivalent:		Agreement with food consumption per adult equivalent:		Agreement with wealth index:		Estimated project effect (95% confidence interval)	Number of evaluation datasets
	Proportion of households in the same quintile	Kappa statistic (95% confidence interval)	Proportion of households in the same quintile	Kappa statistic (95% confidence interval)	Proportion of households in the same quintile	Kappa statistic (95% confidence interval)		
Total consumption per adult equivalent	–	–	54%	0.65 (0.61, 0.70)	28%	0.21 (0.18, 0.23)	0.12 (0.03, 0.20)	23
Total consumption per household	47%	0.57 (0.54, 0.60)	38%	0.42 (0.36, 0.47)	33%	0.32 (0.30, 0.35)	0.14 (0.05, 0.22)	23
Food consumption per adult equivalent	54%	0.65 (0.61, 0.70)	–	–	25%	0.13 (0.10, 0.16)	0.09 (-0.00, 0.18)	23
Non-food consumption per adult equivalent	49%	0.59 (0.55, 0.63)	30%	0.27 (0.24, 0.30)	29%	0.23 (0.20, 0.26)	0.12 (0.05, 0.20)	23
Dietary diversity	33%	0.32 (0.27, 0.37)	34%	0.35 (0.30, 0.39)	28%	0.23 (0.19, 0.27)	0.15 (0.06, 0.24)	23
Food security	23%	0.07 (0.03, 0.11)	22%	0.06 (0.02, 0.10)	25%	0.12 (0.07, 0.17)	0.02 (-0.10, 0.14)	11
Wealth index	28%	0.21 (0.18, 0.23)	25%	0.13 (0.10, 0.16)	–	–	0.17 (0.09, 0.24)[a]	23

[a] Based on change in wealth indicators since the pre-project period.

Table 4: Agreement between subjective outcome measures and other outcome measures

Outcome measure	Distribution of observations with positive response for the outcome measure across quintiles of total consumption per adult equivalent:					Distribution of observations with positive response for the outcome measure across quintiles of food consumption per adult equivalent:					Distribution of observations with positive response for the outcome measure across quintiles of wealth index:					Estimated project effect (95% confidence interval)	Number of evaluation datasets
	Q1	Q2	Q3	Q4	Q5	Q1	Q2	Q3	Q4	Q5	Q1	Q2	Q3	Q4	Q5		
Economic situation[a]	18%	19%	20%	21%	22%	19%	19%	20%	21%	21%	15%	18%	20%	22%	25%	0.11 (-0.04, 0.26)	12
Increase in income[b]	16%	16%	20%	22%	25%	18%	18%	19%	22%	23%	12%	15%	17%	24%	31%	0.16 (0.02, 0.31)	5

[a] Respondents reporting that their household was 'doing well' or 'breaking even'. [b] Respondents reporting that their household's overall income had increased since the notional pre-project period.

5.2 IS FOOD CONSUMPTION A GOOD PROXY FOR OVERALL CONSUMPTION?

As discussed in Section 2.3, the collection of household consumption data was a major emphasis in the 23 evaluations. Detailed data were collected on food consumed by the household during the seven days prior to the survey, as well as on non-food expenditure by household members over a longer time-frame (during the past month, past three months or past 12 months, depending on the type of expense). The breakdown of the proportion of food and non-food consumption in the overall consumption measure in each dataset is shown on the left-hand side in Figures 11 and 12 respectively. In all the datasets, except those from Colombia and Pakistan, food consumption made up more than half of overall consumption. As would be expected, food consumption generally makes up a greater proportion of total household consumption in lower-income countries than in middle-income countries.

One useful question to consider is whether it is necessary in evaluations such as these to collect data on both food and non-food consumption. If project effects could be accurately estimated from food consumption alone, or from non-food consumption alone, then collecting data on both would be redundant. *A priori*, food consumption data may be expected to be a more useful guide to estimating project impacts than non-food consumption for four reasons. Firstly, non-food expenditure is expected to be more 'lumpy' than food consumption (especially given that investments, such as the purchase of durable goods, are included in non-food expenditure in these datasets), so should have higher variance. Secondly, recall accuracy is known to decrease as the recall period increases (see, for example, Beegle *et al.*, 2012): the recall of food consumption over seven days is likely to be more accurate than the recall of non-food expenditures over one-month or 12-month periods. Thirdly, the food consumption module of the survey instruments was structured to enable as accurate an assessment as possible of the value of food consumption; in contrast, the types of non-food expenditure varied widely, so it was not possible to structure the survey instrument in a way that would facilitate the estimation of expenditure. Finally, many types of non-food expenditure involve monetary values that are orders of magnitude greater than the value of any one food type consumed: the greater number of digits means that they are more likely to be subject to transcription or data-entry errors.

Table 3 quantifies the extent of agreement between the food consumption and non-food consumption measures and total consumption. There is only moderate agreement between these measures, with kappa statistics of 0.65 and 0.59 respectively for food and non-food consumption. The agreement between the food and non-food consumption figures is particularly weak, with only 30% of observations being allocated to the same quantile under both measures, and a kappa statistic of 0.27.

Figures 11 and 12 show meta-analyses for the estimated project effects on food and non-food consumption respectively. These charts can be directly compared to Figure 1. It can be seen that estimating project impact from the food consumption data only would generally result in lower estimates of impact than using data on overall consumption. The project effect is estimated at 0.09 standard deviations in terms of food consumption, but 0.12 in terms of total consumption. In some of the evaluations this would substantively change our interpretation of whether there was evidence of a positive impact from the project: relying on food consumption data would suggest that the effects of the projects evaluated in Ethiopia, Sri Lanka, and the Philippines are much smaller than was estimated from total consumption, and the results from DRC and Zimbabwe (both from in 2011/12) lose statistical significance. In contrast, the effect of the project in Colombia is much clearer when only food consumption is considered.

These results suggest that it is not possible to forgo the collection of data on non-food consumption completely without affecting evaluators' ability to detect project effects in some

contexts. In the 11 datasets in which food consumption makes up more than two-thirds of total consumption, there is (naturally) quite close agreement between the two measures in both the allocation of households between quintiles (the kappa statistic is 0.73) and in the estimated project effects. In these particular cases, the non-food consumption adds little value to the food consumption data – but the difficulty is in judging in advance which those cases are. There may be potential in some cases to use existing survey data to make this assessment, and not to collect data on non-food consumption if non-food items make up only a small proportion of total consumption. It may be possible that forgoing the non-food consumption data could allow survey teams to spend relatively more time during interviews on the detailed food consumption data and thereby increase its measurement accuracy.

Figure 11: Project effects on household food consumption (per adult equivalent per day)

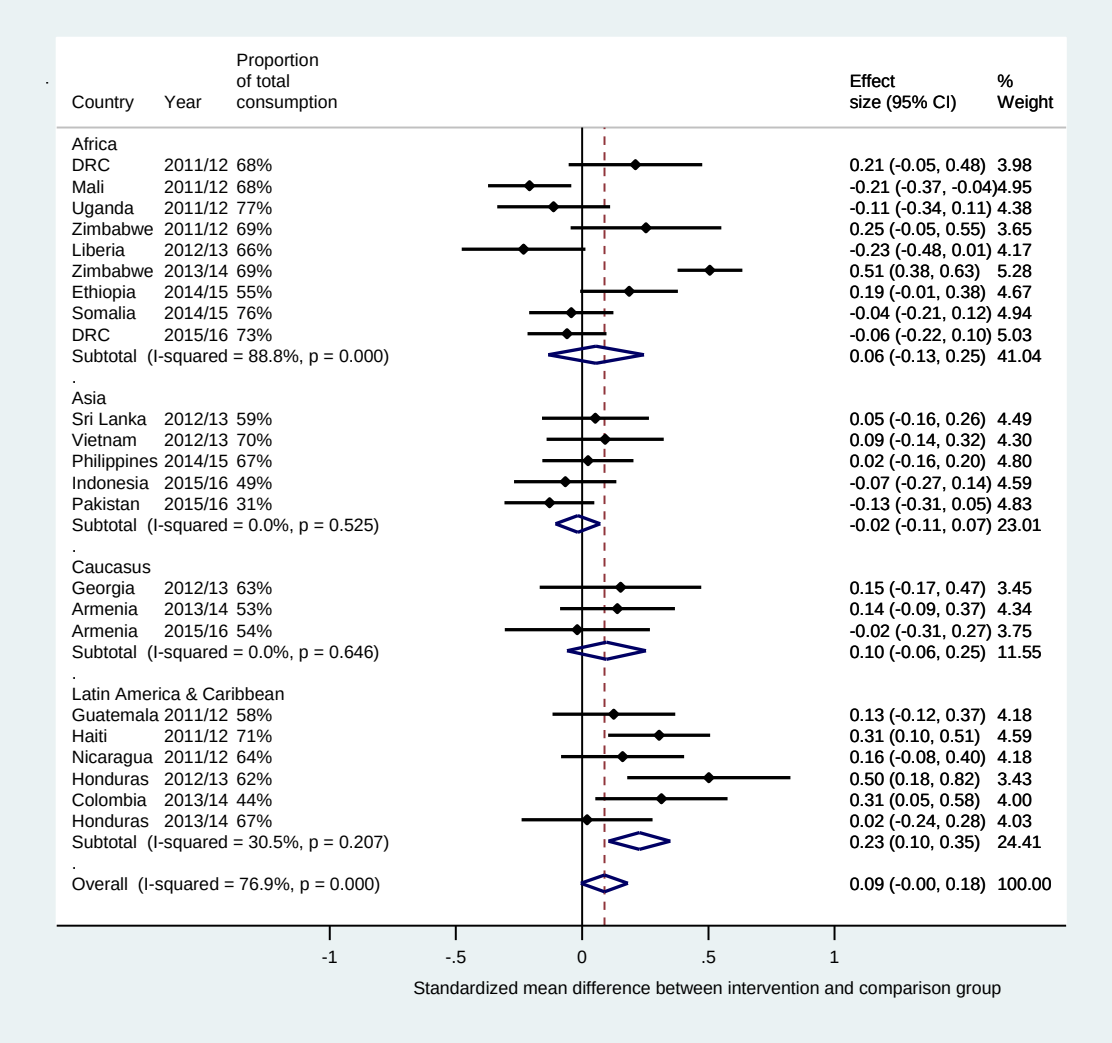
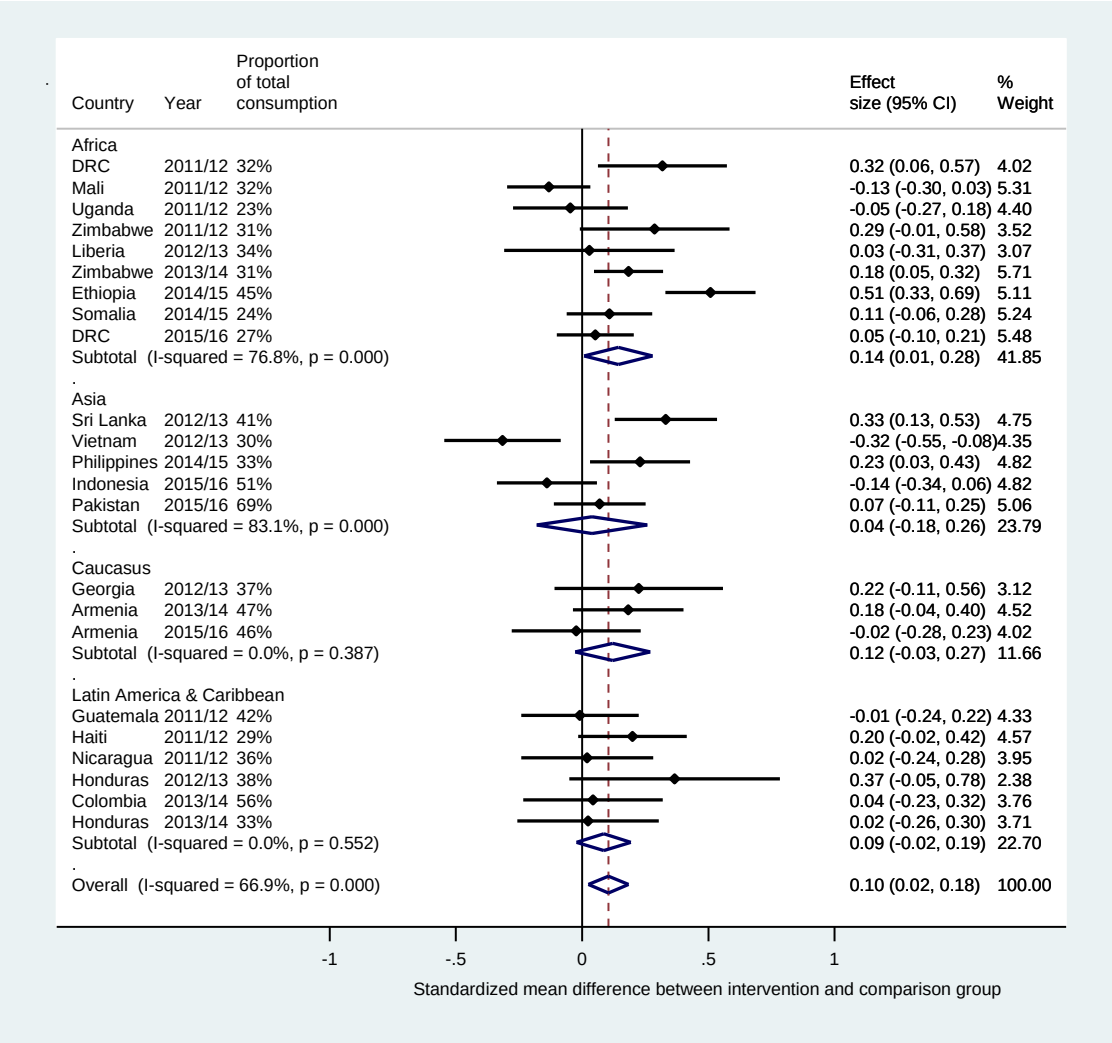


Figure 12: Project effects on household non-food consumption (per adult equivalent per day)



5.3 IS DIETARY DIVERSITY A GOOD PROXY FOR CONSUMPTION?

If it may be sufficient in some contexts to measure welfare using food consumption only, it is interesting to ask whether there are further simplifications that could be made. One question of interest is whether data based only on the diversity of food items consumed in the household can proxy for the total value of food consumption.[31]

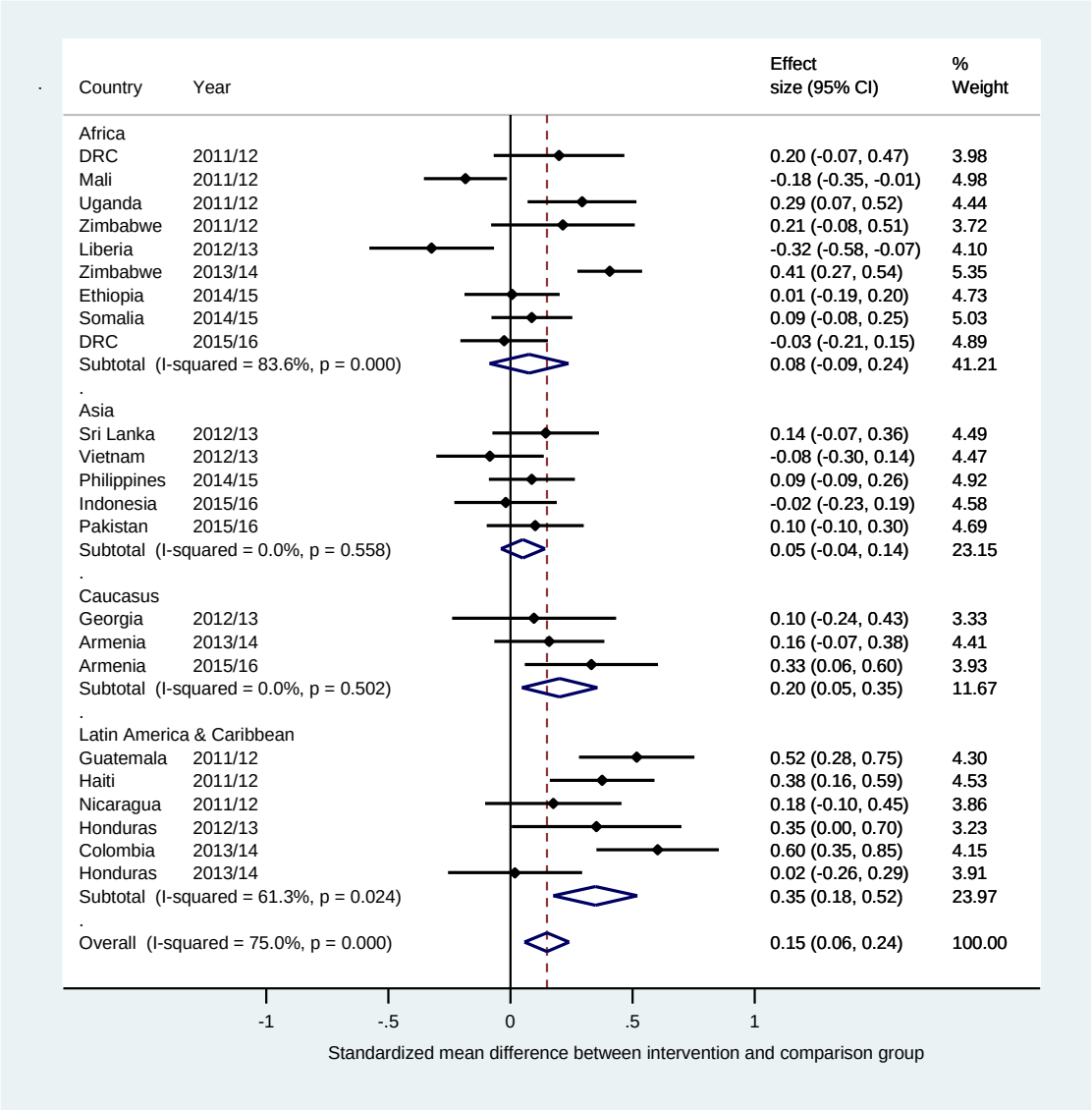
The correlation between dietary diversity and consumption measures is much lower than that between the various consumption measures discussed in the preceding sections: only 32% of observations are in the same quintile in terms of dietary diversity and total consumption per adult equivalent, with the kappa statistic being 0.32. The equivalent figures for the agreement with household *food* consumption per adult equivalent are only slightly higher: 34% are allocated to the same quintile, with the kappa statistic being 0.35.

Figure 13 shows the results of the meta-analysis for estimated project effects on the diversity of food items consumed. Comparing this with Figure 1 or Figure 11, it can be seen that the estimated effect sizes across the complete set of evaluations are similar (0.15 standard deviations, against 0.12 standard deviations for the effect on total consumption). However, there are important differences in the results derived for specific evaluations. The projects in Uganda and Guatemala show strong positive effects on dietary diversity, despite having had no

detectable effect on consumption, and the effect of the project in Colombia is much larger than was estimated from the value of food consumption or total household consumption. On the other hand, the figures for dietary diversity fail to show the positive effects of the projects in Ethiopia and Sri Lanka that were estimated from the consumption data. In each case, these results seem to be linked to whether the project itself had encouraged participants to diversity the range of crops they are growing.

It appears, then, that restricting data collection to the number of food items consumed would be likely to underestimate project effects in some cases and exaggerate project effects in others. Dietary diversity may, of course, be an important indicator of welfare in its own right. For example, Hatløy *et al.* (1998) find a correlation between the variety of food items consumed and a more sophisticated measure of the adequacy of nutrients in a household's diet.

Figure 13: Project effects on number of different food items consumed in the household



5.4 SHOULD FOOD SECURITY INDICATORS BE USED AS AN OUTCOME MEASURE?

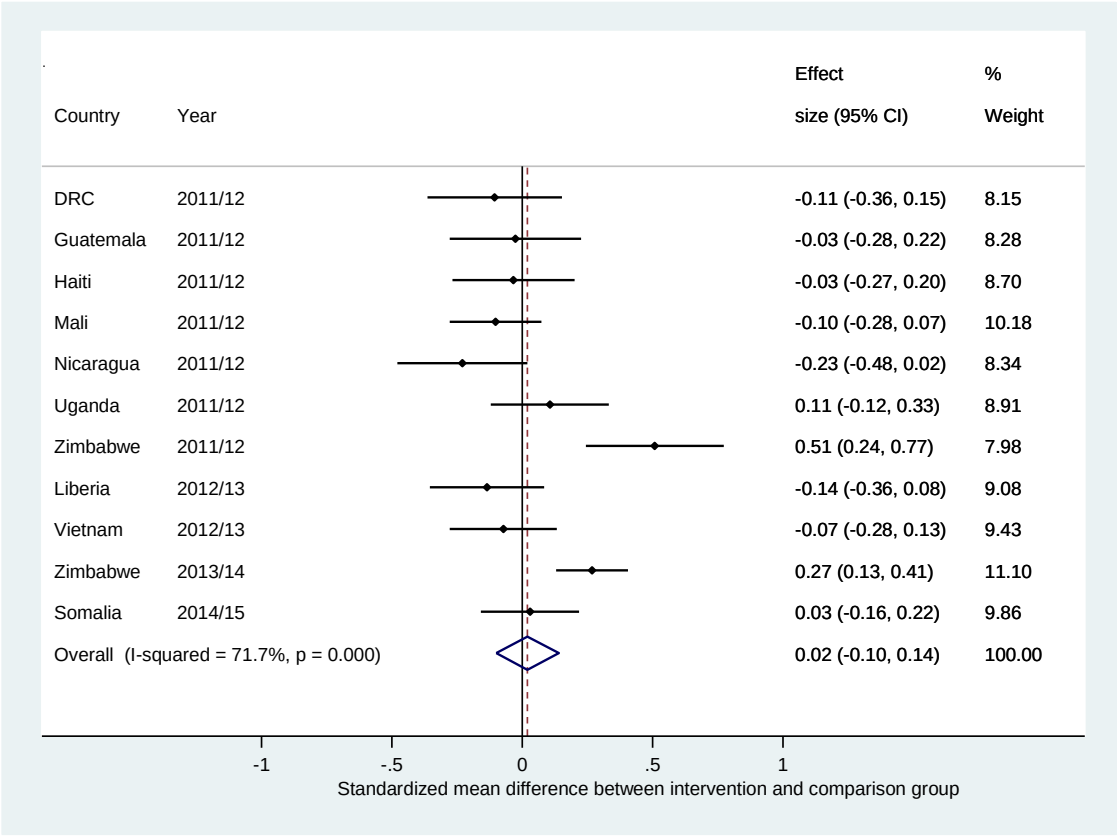
Apart from the detailed questions about food consumption, several of the evaluations also included questions about whether households had experienced any difficulties in accessing food in the recent past. The questions were based on those included in the Household Hunger Scale and the Household Food Insecurity Access Scale developed by the Food and Nutrition Technical Assistance (FANTA) projects (Ballard *et al.*, 2011). Most of the evaluations made use of six such 'experiential' indicators of food insecurity,[32] but the specific questions varied between evaluations, as did the recall periods and response categories. In spite of these differences, the resulting food security indices can be compared in standardized form.

Of course, asking about whether respondents and their households had experienced any difficulties in accessing food is not the same as asking about the total value of their food consumption – but a relation between these two measures may be expected. Indeed, Melgar-Quinonez *et al.* (2006) find correlations between food security and the total value of food consumption in Bolivia, Burkina Faso, and the Philippines. However, our data do not provide strong support for this conclusion. Across the 11 evaluations that collected data on food insecurity, the kappa statistic for agreement between the food security score and food consumption was 0.06, showing that agreement was little higher than would be expected if the two measures were completely unrelated.[33]

Figure 14 shows the results of a meta-analysis for project effects based on the food insecurity measures. It can be seen that there is no evidence that the 11 projects had any effect on the food security measure on average. In fact, there is evidence for a positive effect from the project only in the cases of the two evaluations from Zimbabwe. Unlike in the analysis of the detailed data on food consumption and on total household consumption, the food insecurity measure provides no evidence for a positive effect from the projects in DRC (in 2011/12) or Haiti. The estimated project effects on the food security score are generally closer to zero than are estimated effects in terms of consumption, dietary diversity or wealth indicators.

It is possible that these results reflect a genuine weakness in the projects evaluated – that is, that even the projects that led to the average participant household increasing their food consumption did not significantly reduce their risk of experiencing food security problems (except in Zimbabwe). Alternatively, it may be that the food security indicators have low sensitivity to detect changes in households' food security situation. Indeed, the reason that food security measures were not included in most of the evaluations conducted in later years was that they were found to have been of little use in detecting project effects. An important consideration is that the survey teams in several of the evaluations reported that many respondents found the questions about food security particularly intrusive and were reluctant to answer them. Apart from the undesirable consequence of leaving respondents with negative feelings about the survey, these factors would also tend to increase the inaccuracy in the measurement of food security and so reduce the statistical power available for detecting project effects. The results in Figure 14 suggest that the decision to discontinue use of the experiential food security indicators was justified, and that these indicators are of little value in evaluating project impact.

Figure 14: Random effects meta-analysis for project effects on food security score



5.5 WHAT CAN BE LEARNED FROM THE COMPARISON OF RESULTS BASED ON HOUSEHOLD INCOME AND THOSE BASED ON WEALTH INDICATORS?

As discussed in Section 2.3, the data on household consumption and other welfare measures was complemented in each of the evaluations by data on asset ownership and housing conditions, which allowed an index of wealth indicators to be constructed using the approach of Filmer and Pritchett (2001). A particular strength of the wealth indicator data in Oxfam's evaluations is that it is available for the pre-project period (albeit based on respondents' recall) as well as for the time of the survey, allowing difference-in-difference estimation of the project effects.

In Section 3, the aggregated effect of the projects evaluated was estimated to be larger when measured by the difference in the wealth index than by household consumption. However, the difference in effect size was not uniform across the evaluations: some of the evaluations found a significant effect in terms of consumption but not wealth indicators, and some found the opposite. In Table 3 it can also be seen that the wealth index and consumption measures show only low levels of agreement about the categorization of households into quintiles: the kappa statistic for the wealth index and consumption per adult equivalent is only 0.20. These results concur with existing literature that finds similarly low levels of agreement between wealth indices and consumption (Howe *et al.*, 2008; Howe *et al.*, 2009; Filmer and Scott, 2012). The analysis supports Filmer and Scott's observation that agreement is stronger when considering total consumption for the household (that is, without adjusting for household size) – but even so,

the kappa statistic is a modest 0.32, and only 33% of households are categorized in the same quintile using both measures.

Howe *et al.* (2009) observe that wealth indices tend to agree more closely with household consumption data in middle-income countries. However, there is no evidence for this pattern in the Oxfam data.[34] They also find that agreement is closer when the wealth index is based on a larger range of indicators, but again the Oxfam data do not support this.[35]

The fact that wealth indices do not agree closely with the household consumption aggregates does not imply that they are not valuable indicators of welfare in themselves. It has been argued (Filmer and Pritchett, 2001; Sahn and Stifel, 2003; Filmer and Scott, 2012) that wealth indices are to be preferred as a measure of long-term welfare than consumption measures, which are more likely to be subject to short-term fluctuations. Some of the evaluation reports – such as that from Haiti – have applied this interpretation when discussing reasons why the results derived from the two measures do not agree. However, there does not seem to be any clear evidence in the literature that wealth indices provide a better measure of long-term welfare than do consumption aggregates. Most of the evaluations included in the meta-analysis were anyway conducted while implementation of the project was ongoing, or shortly after implementation ended. For that reason, it could therefore be argued that a short-term change in welfare would be an important indicator of success. In any case, some of the evaluations (such as those carried out in Nicaragua and Colombia) find stronger evidence of an effect on wealth indicators than on consumption, even though the projects were still at a relatively early stage of implementation at the time they were evaluated.

The possibility of deriving pseudo difference-in-difference estimates of change in the wealth index provides an important check on the robustness of the results derived from household consumption – for which no pre-project data are available, and which it would not be realistic to expect respondents to recall with any accuracy. Unfortunately, though, it remains unclear how to interpret the results when the two measures are not in agreement.

5.6 SHOULD SUBJECTIVE WELFARE ASSESSMENTS BE USED AS AN OUTCOME MEASURE?

Most of the earlier evaluations included questions asking for respondents' subjective assessments of their overall economic situation, or of whether they had experienced an increase in income over the past several years.

In particular, all of the evaluations carried out in 2011/12 and 2012/13 included a question asking respondents to assess their household's economic situation, on a four-point scale ranging from 'doing well' to 'unable to meet household needs'. (The form of the question and the response categories were consistent across the 12 evaluations.) In the first row of Table 4, the distribution of those who responded positively (that is, those who responded that they were either 'doing well' or 'breaking even') is shown across the quintiles of the consumption and wealth measures. It can be seen that there is some correlation between the responses to this question and household consumption, though the strength of the correlation is low: 22% of those who responded positively are in the top quintile of consumption per adult equivalent, against 18% who are in the bottom quintile. The link between responses to the subjective economic welfare question and the wealth index is stronger.

When the estimated project effects on the subjective economic measure are examined (as shown in Figure 15), it can be seen that they vary quite widely from the estimated project effects measured by using household consumption or wealth (shown in Figure 1 and Figure 3). In particular, the subjective measure does not show a significant positive effect from the project in

Honduras (2012/13), nor the significant negative effect in Mali. The only cases in which the subjective measure produces a positive effect that is clearly statistically significant are for the projects in Zimbabwe (2011/12) and Sri Lanka, and in both of these cases it produces an estimate much larger than that derived from the consumption or wealth data. (For example, the estimated effect size on the subjective measure in Zimbabwe is 0.76 standard deviations, compared to 0.44 standard deviations for the household consumption measure.)

In five of the evaluations, respondents were also asked a general question about whether their overall household income had increased, decreased, or remained approximately the same over the past several years (since the pre-project period).[36] As can be seen from the second row of Table 4, the distribution of this measure is more clearly linked to consumption than was the economic welfare measure. The connection is even stronger between this measure and the wealth index: 32% of those who reported that their household had experienced an increase in income were in the highest quintile according to the wealth index, and only 12% were in the lowest quintile. However, the estimation of project effects using this indicator, shown in Figure 16, again demonstrates little correspondence to the results derived from the consumption or wealth data.

Respondents' subjective sense of their economic level and the trajectory of their income is clearly important to their welfare. The fact that the results from these subjective measures do not reflect the results derived from the consumption data does not, therefore, imply that they are of no value. It is possible that the particularly positive results for the subjective measures in Zimbabwe, Sri Lanka, and Georgia are a result of the project participants feeling a higher degree of confidence or optimism as a result of the project activities. However, we should be cautious in drawing such a strong conclusion from this data. As noted by Howe *et al.* (2011), survey respondents may be inclined to report that their subjective welfare is low if they believe that they could gain by doing so. In our case, a particular concern is that the survey respondents were aware that the interview was connected to the project being evaluated. It is possible that respondents may have felt that they should either (a) exaggerate their level of poverty or (b) overstate the successes they had experienced during the project's lifetime, in order that the project would continue or that they would benefit from other projects in the future.[37] It seems unwise to base assessments of projects' effects on simple subjective measures such as these unless the survey process is so separated from project implementation that respondents do not make any connection between them.

On the other hand, if it is accepted that the subjective welfare measures provide important information about how respondents perceive their own welfare, then the fact that responses to these questions are more closely linked to values of the wealth index than to consumption adds weight to the argument that the wealth index should be treated as an important outcome measure in itself.

Figure 15: Random effects meta-analysis for project effects on subjective economic welfare measure

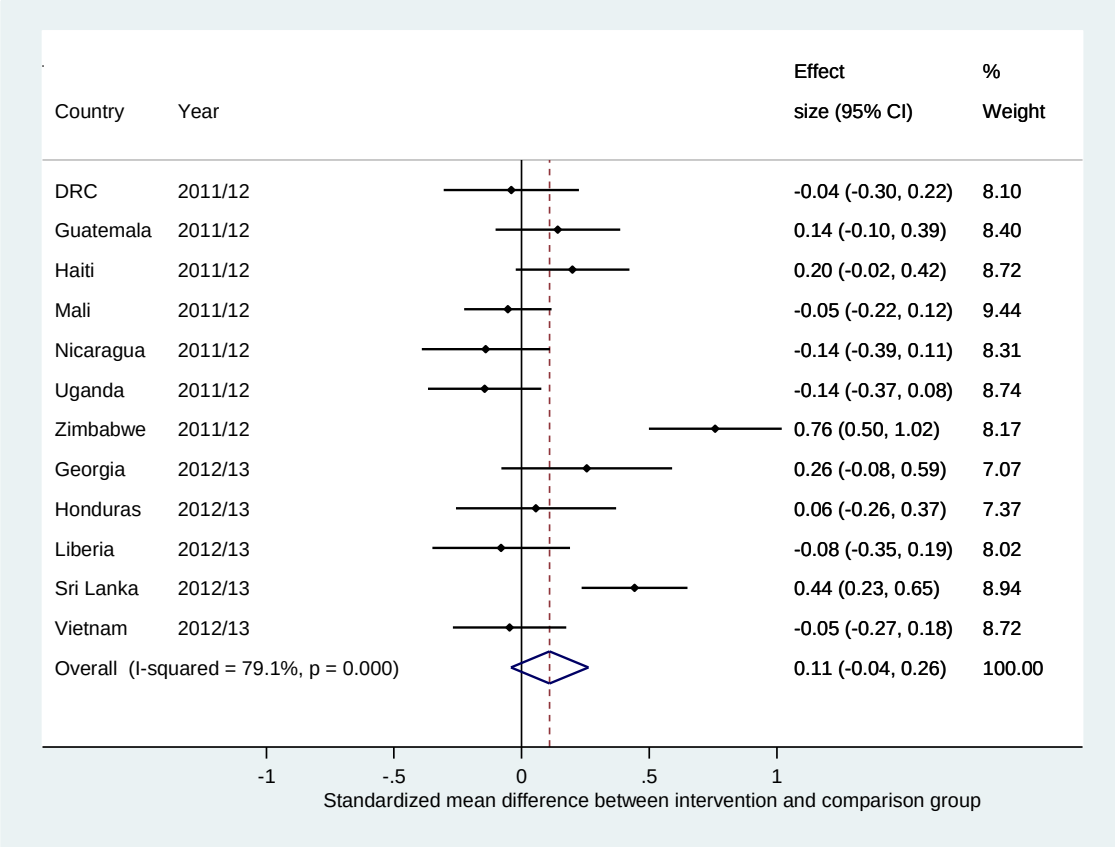
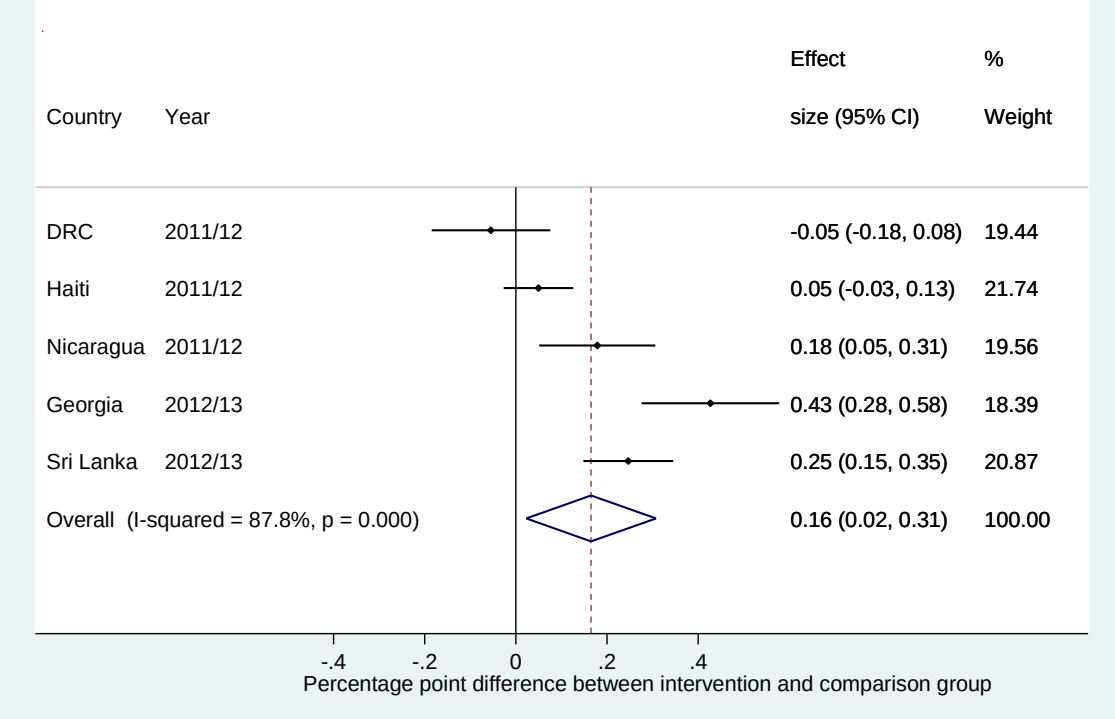


Figure 16: Random effects meta-analysis for project effects on subjective income change



6 LEARNING FROM EVALUATIONS

The evaluations were all carried out by Oxfam with the aim not only of understanding the effects of its projects, but also of generating learning about how to make those projects – and future projects – more effective. To that end, each of the evaluation reports concludes with a set of 'programme learning considerations', suggestions about what can be learned from the results or what further research would be useful. In earlier years these learning considerations were mostly proposed by the evaluators themselves; in more recent years, identifying the learning considerations has involved a collaborative process of discussion with the programme teams.

The most straightforward of the learning considerations are those in which the evaluations found clearly positive results from (some or all of) the project interventions. In these cases, learning considerations were often included that called for interventions to be replicated or scaled up, either within Oxfam programmes or through advocacy with partners. Ten of the 23 evaluations included such a call for replication or scale-up, either of particular interventions (such as the promotion of agriculture in Uganda, the mentorship programme in Indonesia, or the use of coffee as a marketing channel in Ethiopia) or of the overall project strategy (such as farmers' markets in Colombia or the 'basket' of interventions in the evaluation in Zimbabwe in 2013/14).

However, most of the evaluations (with the single exception of that carried out in Colombia) also provided constructive criticism of some aspects of the project. The learning considerations that provided such criticism can be broadly categorized into (a) those relating to the theory of change of the project, (b) those relating to how the project was organized or implemented, and (c) those relating to monitoring, evaluation, and learning processes.

As discussed in Section 4, most of the evaluations attempted to examine data on intermediate stages in projects' expected chains of results. Approximately a third of the evaluations (eight of the 23 cases) were able to identify points at which the results chain had broken down, thereby explaining why the project did not achieve more-positive results. In some cases, it is not clear why those breaks in the project logic occurred. For example, in Guatemala the increased use of fertilizer did not result in increased crop yields, while in the Philippines, increased revenue from agricultural sales seems not to have led to an overall increase in household income. In other cases, the evaluation results make clear what the nature of the problem is, if not how to remedy it. For example, in Armenia (the project evaluated in 2013/14) the producers had difficulty in accessing the agricultural inputs they had been trained to use on the local market, while in Nicaragua producers had not been able to realize higher prices when selling cocoa and dairy products, as had been intended. It is notable that the more recent evaluations have been more likely to propose a solution to problems that were identified, such as to focus support on a smaller number of value-chain enterprises in Indonesia, or to advocate for improved road networks in DRC as a means to reduce transportation costs. These more concrete recommendations probably reflect the closer involvement of programme staff in drafting them in recent years. In earlier years, the evaluation reports tended instead to raise questions to consider without proposing solutions.

Some of the evaluations involved analysing results separately in two or more regions or districts in which the project was implemented. In several of these cases the results were more positive in one of the areas than in the others. In those cases, the evaluators have suggested investigating whether learning can be transferred between the regions. In addition, two of the evaluations (in Haiti and the first evaluation in Honduras) suggest that there may be potential to learn from the comparison group, and one of the evaluations (Sri Lanka) suggests learning from a small group of individuals who had particularly positive results under the project.

However, there were very few cases in which the evaluations called into question the overall logic of the project. The only clear example of this was in highlighting a contradiction between promoting diversification at the same time as giving producers an incentive for the production of a single crop type. This was a key finding in the evaluation in Haiti, and arose also in the report of the evaluation in Honduras in 2013/14. The fact that the evaluations have rarely criticized the fundamental logic of the projects reflects the design of these evaluations: it is inherently difficult with an *ex-post* approach to test specific hypotheses about the interventions.

A second category of learning considerations focused on the structure and implementation of the projects evaluated. These involved raising questions about the number of partners involved in a project (Indonesia), whether the interventions had the potential to complement each other (Guatemala), whether the institutions established under the project were sustained after the project's end (in both evaluations in Honduras and the second evaluation in Armenia), and whether the project was adding value to government programmes (Vietnam) or to the activities of other actors (Pakistan). In two cases (Liberia and the Philippines) the evaluation data implied that the process of targeting had been poor: the households of project participants appear to have been significantly more wealthy before the project than the average household in their communities, contrary to the projects' intention to target more-vulnerable households.[38]

In another two evaluation reports (Georgia and Uganda) the learning considerations raised the question of whether the outcomes among what was a relatively small number of participant households were commensurate with the resources put into the project. On the other hand, the first evaluation in DRC and that in Mali highlighted that some of the interventions may have been too diffuse to achieve any detectable results for the average participant.

The final category of learning considerations is those relating to monitoring and evaluation procedures: ten of the evaluation reports mentioned either the need to improve monitoring during implementation, or included suggestions on how to integrate an evaluation plan into the project design.

7 CONCLUSIONS

This meta-analysis has provided evidence that the 23 projects had a statistically significant, positive impact on the material welfare of the average participant household. The size of this effect is estimated at 0.12 standard deviations in terms of household consumption, or 0.17 standard deviations on the index of wealth indicators. Some of the projects were more successful in this respect than others – and it should be recognized that a third of the projects do not appear to have had any significant impact either on consumption or on wealth indicators for the average household.

The analysis does not provide any evidence of systematic differences in project effects between regions of the world, by the scale, duration or budget of the project, nor by participants' pre-project economic level. There is some evidence (from the change in wealth indicators) that female-headed households have tended to benefit less from the projects than male-headed households. However, there remains considerable heterogeneity in project-effect sizes, and future research could examine the sources of this variation further.

The meta-analysis of intermediate outcomes confirms that the projects were generally successful in encouraging participants to engage in production and sales of the specific agricultural products promoted under these projects. In those cases where the project had a positive effect on revenue from agricultural sales, this was generally associated with increased household consumption (which is taken to be a proxy measure of overall net income). This may be taken to imply that the increased agricultural sales were generally profitable for the project participants. As discussed in Section 6, several of the evaluation reports used the analysis of intermediate outcomes to identify what went wrong in cases where the projects did not produce the expected results. However, it is important to have realistic expectations about the degree to which quasi-experimental evaluations such as these can assess projects' effects on intermediate outcomes (see Green *et al.* (2010) for a useful warning on this subject). In many of the evaluations, the key learning points arose from observations made in the course of conducting the evaluation, rather than from the results of the evaluation themselves.

The meta-analysis results also provide some valuable insights into the measurement approach used in these evaluations. The results provide some support for the decision to prioritize household consumption and wealth indicators as the key measures of welfare. Alternative outcome measures – such as dietary diversity or food security indicators – seem to be less sensitive, and so reduce the statistical power available for detecting project effects.

One important observation is that the household consumption data and the household wealth indicators provide markedly different views on the effects of several of the projects. It may be hypothesized that the consumption data provides a more sensitive indication of immediate changes in income, while wealth indicators may provide a more stable measure of longer-term economic level – but this idea is certainly open to question, and has not been tested. The data also provide some evidence that the wealth index is more closely linked to respondents' own perceptions of their welfare than is the consumption data.

The results suggest that there is value in continuing to measure both income and wealth, but that there is further investigation to be done on the connections and complementarity between these measures. In future, efforts to move beyond aggregated household-level measures of welfare would help to understand intra-household project effects, and would allow for stronger differentiation and understanding of gender impacts from Oxfam's livelihoods projects.

APPENDIX 1: EVALUATIONS INCLUDED IN THE META-ANALYSIS

Region	Country	Year of evaluation	Main project activities	Approximate number of participant households	Project duration
Africa	Democratic Republic of Congo (DRC)	2011/12	Training, inputs, and infrastructure for fishers and fish processors	586	1 year
	Mali[a]	2011/12	Training, inputs, and technical support to cotton farmers	2,936	4 years
	Uganda	2011/12	Training and inputs to women's groups for agriculture and livestock rearing	427	4 years
	Zimbabwe	2011/12	Provision of irrigation, training, and inputs for agriculture	70	1 year
	Liberia	2012/13	Provision of irrigation, training, inputs, and marketing support for rice production	855	2 years
	Zimbabwe	2013/14	Training and inputs for agriculture and livestock rearing	2,115	4 years
	Ethiopia	2014/15	Training, inputs, and marketing support for coffee production	3,072	3 years
	Somalia	2014/15	Provision of inputs and equipment for household businesses, Cash for Work	1,160	3 years
	DRC	2015/16	Training, inputs, and marketing support for agricultural production	5,470	2 years
Asia	Sri Lanka	2012/13	Provision of irrigation, training, inputs, and marketing support for rice production	9,950	4 years
	Vietnam	2012/13	Training, credit, and marketing support for agricultural production	271	4 years
	Philippines	2014/15	Training, inputs, and marketing support for agricultural production	1,231	6 years
	Indonesia	2015/16	Training and inputs for agricultural production	3,000	3 years
	Pakistan[b]	2015/16	Training and marketing support for dairy production	660	3 years

Region	Country	Year of evaluation	Main project activities	Approximate number of participant households	Project duration
Caucasus	Georgia	2012/13	Training, inputs, and marketing support for agricultural or livestock production	134	4 years
	Armenia	2012/13	Training, inputs, processing facilities, credit, and marketing support for agricultural production	285	3 years
	Armenia[b]	2015/16	Training, infrastructure, credit, and marketing support for horticultural production	106	2 years
Latin America and the Caribbean	Guatemala[b]	2011/12	Training, inputs, and marketing support for agricultural production and household businesses	262	2 years
	Haiti	2011/12	Support to network of Fair Trade coffee producers	2,243	14 years
	Nicaragua	2011/12	Training, inputs, and marketing support for dairy and cocoa production	127	1 year
	Honduras[b]	2012/13	Establishment of community banks and agricultural enterprise, provision of agricultural inputs and marketing support	103	14 years
	Colombia	2013/14	Establishment of farmers' markets	751	7 years
	Honduras	2013/14	Establishment of community banks, provision of training, inputs, credit, and marketing support for agricultural production	379	7 years

[a] Originally selected for evaluation under the 'resilience' theme, rather than as a livelihoods project.

[b] Originally selected for evaluation under the 'women's empowerment' theme, rather than as a livelihoods project.

Summary statistics on projects' funding allocations are given in endnote 18.

APPENDIX 2: METHODOLOGY USED FOR PROPENSITY-SCORE MATCHING

The sizes of project effects in the 23 evaluations included in this meta-analysis are estimated using propensity-score matching (PSM). The principle of PSM is to match households in the intervention group to those in the comparison group, based on their similarity in terms of pre-project observed characteristics. Following the guidance provided by Caliendo and Kopeinig (2008), variables were used for matching only if they were thought to influence selection into the project but not be affected by participation in the project. The specific characteristics used in each evaluation varied, but typically included:

1. Indicators of the size and composition of the household.
2. Indicators of the gender, age, and education level of the head of household (as defined by the respondent).
3. Indicators of the household's pre-project wealth level, based on respondents' recollections of wealth indicators (ownership or assets and housing characteristics) from a period before implementation of the project. The wealth index is constructed as described in Section 2.3.
4. Indicators of the livelihoods activities or sources of income that the household was engaging in before the project, if these were thought to be relevant to the participation decision and if the recall data were thought to be reliable.
5. Indicators of a household's access to infrastructure or markets, based on estimated travel time to the nearest market or major road.

In most of the evaluations, a list of potential matching variables was drawn up, and a stepwise regression procedure used to eliminate those that were not found to be significant predictors of participation, and (in some cases) that were also not found to be significant predictors of the main outcome variable, the value of household consumption.[39] In seven of the evaluations (mostly those conducted more recently), matching variables were instead chosen deliberately, based on evaluators' judgements of the most important factors determining participation.

It would be difficult to find exact matches for each intervention group household based on all of the variables selected for matching. Instead, these characteristics are used to estimate the propensity score for each household, the probability that a household is in the intervention group, conditional on all the matching variables. Rosenbaum and Rubin (1983) demonstrated that if the intervention and comparison groups are balanced in terms of their propensity scores, then they are also balanced in terms of each of the matching variables.

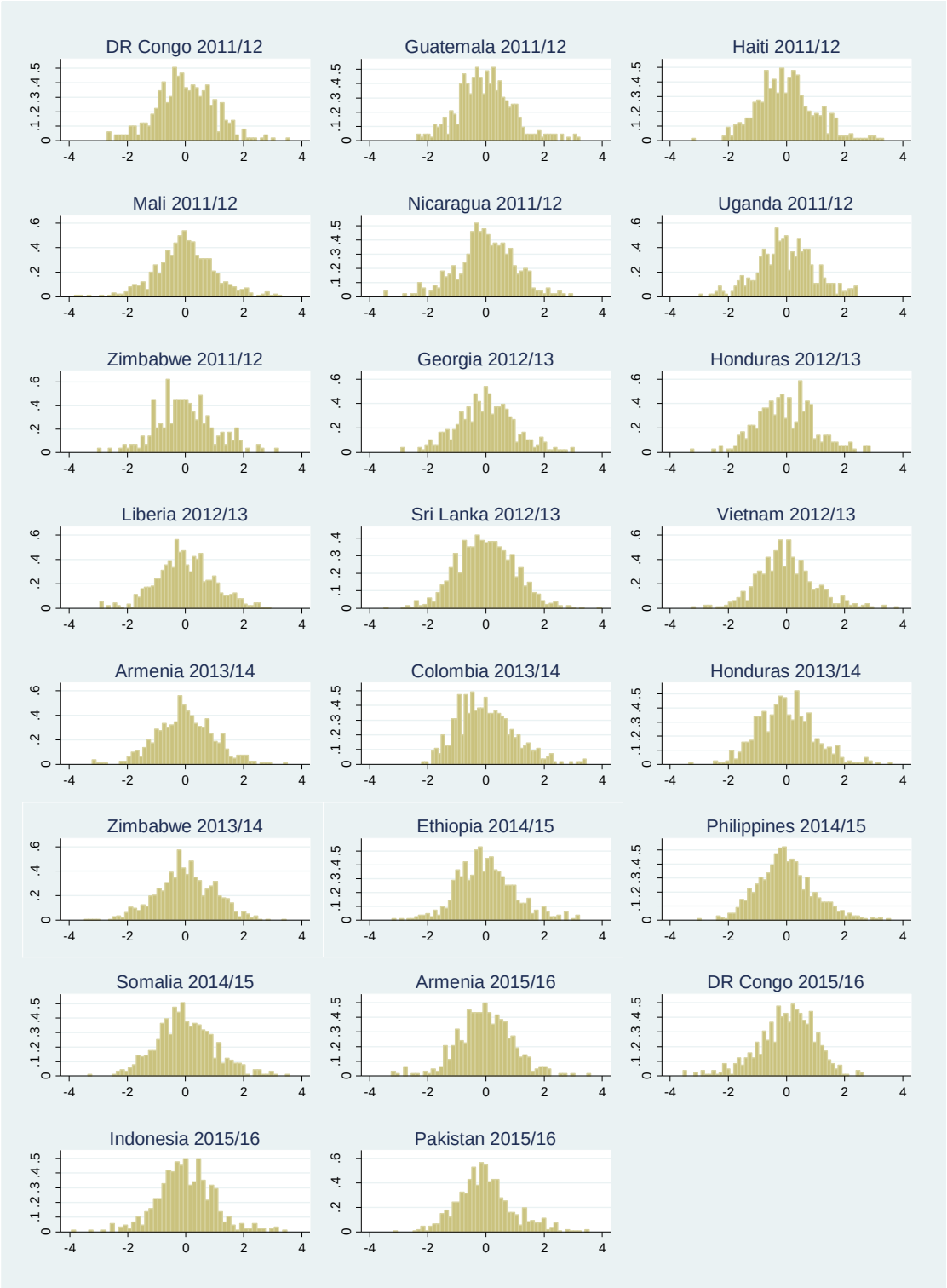
Within each evaluation dataset, propensity scores are estimated by regressing actual intervention status on the selected matching variables, and then using the resulting model to predict each observation's probability of being in the intervention group.[40] Analysis is then restricted to the area of common support, the region in which the propensity score distributions of the intervention and comparison groups overlap. Observations outside the area of common support are dropped from the analysis. In some cases, this results in a small number of intervention observations being dropped, meaning that the matched intervention group is not a fully representative sample of the project participants or beneficiaries. On average, approximately 5% of the intervention group observations are dropped for this reason, though this proportion is as high as 25% in one of the evaluations.

Within the area of common support, a kernel matching procedure is used to match each intervention observation with a weighted average of the comparison observations, with greater weight given to comparison observations with propensity scores close to the propensity score of the intervention observation. The *psmatch2* module in Stata was used to carry out this analysis (Leuven and Sianesi, 2003). The rate at which weights given to the comparison observations declined with distance from the intervention observation was adjusted to minimize the observed

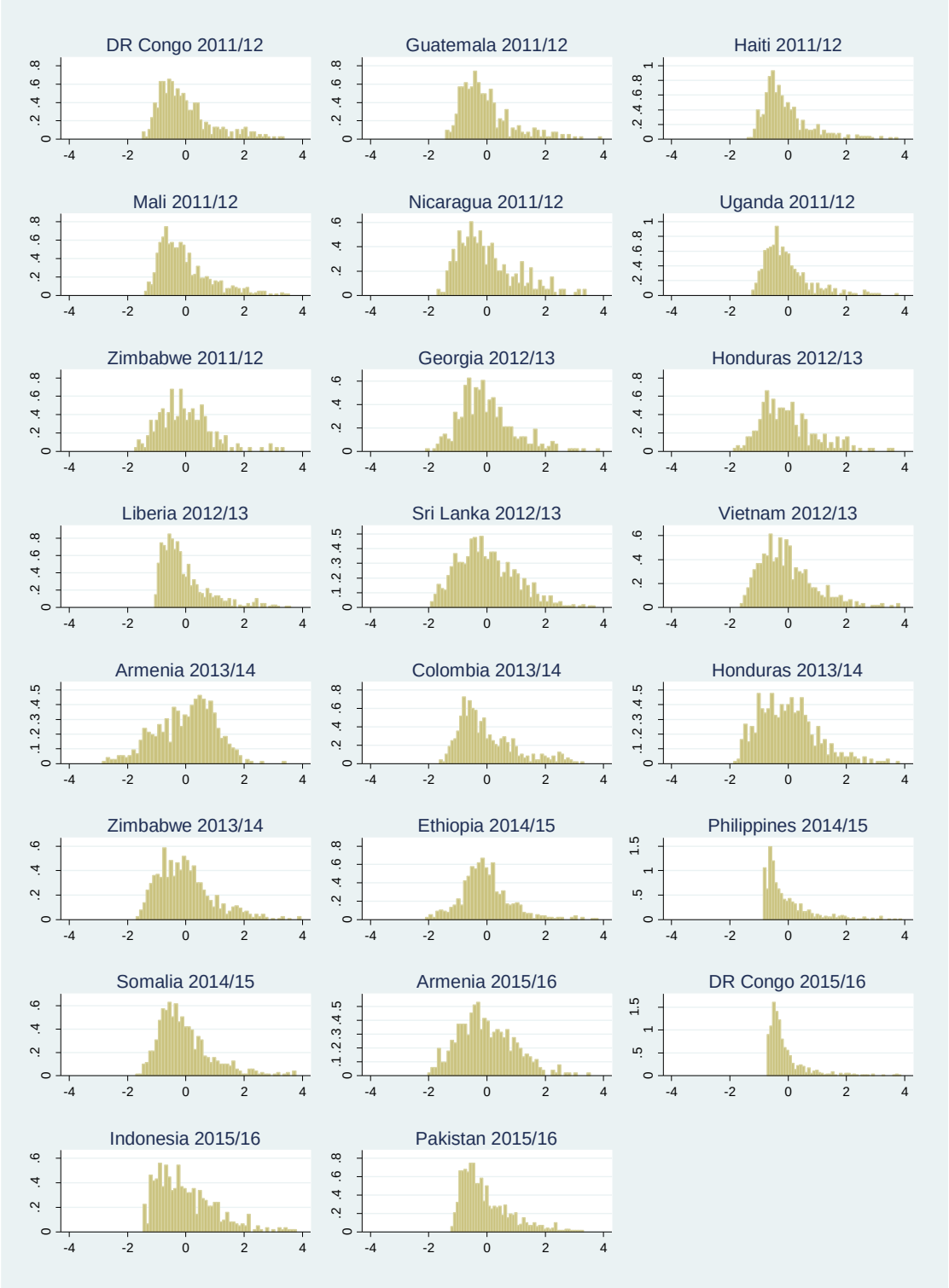
bias of the overall model in each case. The mean standardized bias in the matched model (suggested as a criterion for matching quality by Rosenbaum and Rubin, 1985; Caliendo and Kopeinig, 2008) was less than 10% in all but two cases, and less than 15% in all cases. The pseudo-R^2 of the matched model is less than 0.05 in all but one case, indicating that there are few differences in the distribution of the matching variables between the intervention and comparison groups.[41]

The estimates of the effect of each project on the various outcome measures are calculated as the difference in average outcomes between the matched intervention and comparison groups. Standard errors of the estimates are bootstrapped using 1,000 repetitions to account for the variation caused by the estimation of the propensity scores and the determination of the common support. Standard errors are not clustered during bootstrapping: based on the guidance of Cameron and Miller (2015), the generally small number of clusters over which respondents were sampled in these evaluations would make the calculation of clustered standard errors unreliable.

APPENDIX 3: HISTOGRAMS OF CONSUMPTION PER ADULT EQUIVALENT (STANDARDIZED)



APPENDIX 4: HISTOGRAMS OF POST-PROJECT WEALTH INDEX (STANDARDIZED)



APPENDIX 5: ROBUSTNESS OF META-ANALYSIS ESTIMATES

As discussed in Section 2.2, the evaluations included in this meta-analysis were based on identifying comparison groups *ex-post* – that is, after the project was implemented – and matching on the basis of data recalled from a notional pre-project period. There are clear limitations to this methodology, and the evaluators judged the identification of an appropriate comparison group to have been more successful in some cases than in others.

It is important to consider how the results of the meta-analysis are affected by excluding evaluations in which there are particular reasons to doubt the suitability of the comparison group. Four criteria are used to identify such evaluations:

1. Cases in which the comparison group consisted of households residing (wholly or partially) within the same communities as intervention group households. In these cases, there is potential for project effects to be underestimated if the comparison group had benefited indirectly from the project activities – that is, there may be 'spillover' effects of the intervention on the comparison group. Six of the 23 evaluations included in the meta-analysis have this characteristic.
2. Cases in which the intervention group consists of a deliberately selected (or, most often, self-selected) group of households in the project communities, but in which the comparison group consists of a random sample of households residing in comparison communities. This approach was taken in evaluations in which no credible method could be found to replicate the process by which project participants were selected. Since participation in projects is often thought to be driven by unobservable characteristics that are likely to be associated with positive outcomes (such as initiative, motivation or social connections), it is possible that project effects may be over-estimated in these cases. There are nine evaluations included in the meta-analysis to which this may apply.
3. Cases in which more than 10% of the intervention observations were excluded from the PSM analysis as being outside the common support area. In these cases, it is recognized that the evaluation results are not fully representative of the intervention group over which they were sampled, meaning that there is potential for the results to be biased. This applies to five of the evaluations included in the meta-analysis.
4. Cases in which the Effectiveness Review report specifically recognizes that there are serious concerns over the appropriateness of the comparison group. There are two such cases. Firstly, for the evaluation conducted in Liberia, the comparison group were found to be considerably different from the intervention group in terms of their pre-project wealth status and productive activities, and had also been benefiting from the project activities for several months at the time of the evaluation. Secondly, for the evaluation conducted in Honduras in 2012/13, the effects of the project evaluated are likely to have been conflated with the effect of municipal-level governance (given that the intervention and comparison communities were located in different municipalities), as well as with other actors that were active only in the intervention communities.

Table A5.1 shows how the main meta-analysis results presented in Section 3 of this paper differ after excluding evaluations that meet each of these four criteria. It is reassuring that the overall estimated effects change little after these exclusions. In none of these cases is there a statistically significant difference between the evaluations that are excluded and those that are retained (this is confirmed through meta-regression of the project effects on a dummy variable representing the set of projects excluded based on each of the criteria listed above). After excluding the nine evaluations that are thought to be particularly vulnerable to household-level selection bias (criterion number 2 in the list above), the meta-analysis estimate of the effects across the remaining projects increases; this is the opposite to what would be expected if the meta-analysis were being influenced by such bias.

In the final row of Table A5.1, evaluations that meet *any* of the four criteria listed above are excluded. The meta-analysis estimates across the remaining six evaluations are similar in size to those across the full set of 23 evaluations, although the confidence intervals are large and the effect on household consumption is not statistically significantly different from zero.

Table A5.1: Meta-analysis for the project effect on logarithm of household consumption per adult equivalent per day and on the change in index of wealth indicators

Evaluations included in meta-analysis	Project effect on household consumption per adult equivalent (standardized)	Project effect on change in index of wealth indicators (standardized)	Number of evaluations included in meta-analysis
All	0.12 (0.03, 0.20)	0.17 (0.09, 0.24)	23
Excluding cases with potential spillover effects	0.11 (0.00, 0.22)	0.17 (0.10, 0.25)	18
Excluding evaluations with potential household-level selection effects	0.14 (0.02, 0.26)	0.20 (0.09, 0.31)	14
Excluding evaluations with more than 10% of the intervention group outside the common support area	0.11 (0.01, 0.21)	0.15 (0.07, 0.24)	18
Excluding evaluations in which the original authors identified problems with the comparison group	0.12 (0.03, 0.20)	0.20 (0.13, 0.27)	21
Excluding all evaluations excluded by any of the four criteria listed above	0.11 (–0.11, 0.33)	0.14 (0.01, 0.28)	6

APPENDIX 6: DIFFERENCES BETWEEN INTERVENTION AND COMPARISON GROUPS IN BASELINE WEALTH STATUS

The table compares the intervention and comparison groups included in each of the Effectiveness Reviews based on the (recalled) baseline wealth index. These figures are shown before matching, to provide an indication of whether the intervention group sampled for the survey tended to be more or less wealthy at baseline than the comparison group.

Region	Country	Year of evaluation	Difference between intervention and comparison groups in:					
			Continuous wealth index (standardized)		Proportion of group in the lowest wealth quintile (percentage points)		Proportion of group in the highest wealth quintile (percentage points)	
Africa	DR Congo	2011/12	0.323***	(0.105)	-7.2*	(4.3)	6.7	(4.3)
	Mali[a]	2011/12	-0.029	(0.084)	-2.0	(3.4)	0.0	(3.3)
	Uganda	2011/12	0.036	(0.101)	- 13.0***	(4.0)	-1.8	(4.0)
	Zimbabwe	2011/12	0.393***	(0.141)	-6.5	(5.7)	12.5**	(5.7)
	Liberia	2012/13	0.159**	(0.074)	-10.0***	(2.9)	0.9	(2.9)
	Zimbabwe	2013/14	0.241***	(0.062)	-8.9***	(2.5)	5.1**	(2.5)
	Ethiopia	2014/15	0.588***	(0.080)	-17.6***	(3.3)	16.1***	(3.3)
	Somalia	2014/15	-0.087	(0.083)	4.1	(3.3)	-1.2	(3.3)
	DR Congo	2015/16	0.407***	(0.077)	-3.1	(3.4)	18.1***	(3.1)
Asia	Sri Lanka	2012/13	0.465***	(0.065)	-5.5**	(2.7)	14.9***	(2.7)
	Vietnam	2012/13	0.112	(0.105)	-7.1*	(4.1)	1.8	(4.3)
	Philippines	2014/15	0.309***	(0.073)	0.7	(3.0)	9.9***	(3.0)
	Indonesia	2015/16	-0.430***	(0.082)	-0.1	(3.4)	-22.9***	(3.3)
	Pakistan[b]	2015/16	-0.247***	(0.072)	4.2	(2.9)	-5.7	(2.9)
Caucasus	Georgia	2012/13	-0.242**	(0.105)	12.1***	(4.3)	1.6	(4.2)
	Armenia	2012/13	0.042	(0.087)	-4.6	(3.5)	-6.0	(4.9)
	Armenia[b]	2015/16	0.039	(0.111)	3.3	(4.4)	4.8	(4.4)
Latin America and the Caribbean	Guatemala[b]	2011/12	0.168	(0.108)	-5.4	(4.4)	4.4	(4.3)
	Haiti	2011/12	0.224**	(0.090)	-3.1	(3.7)	13.5***	(5.0)
	Nicaragua	2011/12	0.263**	(0.110)	-11.4**	(4.5)	8.4*	(4.4)
	Honduras[b]	2012/13	0.401***	(0.121)	-8.6*	(4.9)	14.6***	(4.9)
	Colombia	2013/14	0.356***	(0.096)	-8.1**	(3.9)	16.7***	(3.9)
	Honduras	2013/14	0.241**	(0.102)	-20.5***	(4.1)	8.1**	(4.0)

Standard errors are in parentheses.

REFERENCES

Ballard, T., J. Coates, A. Swindale and M. Deitchler. (2011). *Household Hunger Scale: Indicator Definition and Measurement Guide*. Retrieved 10 August 2016 from http://www.fantaproject.org/monitoring-and-evaluation/household-hunger-scale-hhs.

Banerjee, A., E. Duflo, N. Goldberg, D. Karlan, R. Osei, W. Pariente, J. Shapiro, B. Thuysbaert and C. Udry. (2015). *A Multifaceted Program Causes Lasting Progress for the Very Poor: Evidence from Six Countries*. *Science*, 348(6236). DOI: https://doi.org/10.1126/science.1260799

Beegle, K., J. De Weerdt, J. Friedman and J. Gibson. (2012). *Methods of Household Consumption Measurement through Surveys: Experimental Results from Tanzania*. *Journal of Development Economics*, 98(1), 3–18. DOI: https://doi.org/10.1016/j.jdeveco.2011.11.001

Borenstein, M., L.V. Hedges, J.P.T. Higgins and H. Rothstein. (2009). *Introduction to Meta-Analysis*. Chichester: Wiley. DOI: https://doi.org/10.1002/9780470743386

Caliendo, M. and S. Kopeinig. (2008). *Some Practical Guidance for the Implementation of Propensity Score Matching. Journal of Economic Surveys*, 22(1), 31–72. DOI: https://doi.org/10.1111/j.1467-6419.2007.00527.x

Cameron, A.C. and D.L. Miller. (2015). *A Practitioner's Guide to Cluster-Robust Inference. Journal of Human Resources*, 50(2), 317–72. DOI: https://doi.org/10.3368/jhr.50.2.317

Cohen, J. (1960). *A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement,* 20(1), 37–46. DOI: https://doi.org/10.1177/001316446002000104

Cohen, J. (1992). *A Power Primer. Psychological Bulletin*, 112(1), 155–9. DOI: https://doi.org/10.1037/0033-2909.112.1.155

Deaton, A. (1997). *The Analysis of Household Surveys: A Microeconometric Approach to Development Policy*. Baltimore: The Johns Hopkins University Press. DOI: https://doi.org/10.1596/0-8018-5254-4

Deaton, A. and S. Zaidi. (2002). *Guidelines for Constructing Consumption Aggregates for Welfare Analysis. Living Standards Measurement Study Working Paper 135*. Retrieved 20 May 2016 from http://documents.worldbank.org/curated/en/2002/05/1944189/guidelines-constructing-consumption-aggregates-welfare-analysis.

Deitchler, M., T. Ballard, A. Swindale and J. Coates. (2010). *Validation of a Measure of Household Hunger for Cross-Cultural Use*. Retrieved 24 August 2016 from http://www.fantaproject.org/research/validation-hhs

Duvendack, M., J.G. Hombrados, R. Palmer-Jones and H. Waddington. (2012). *Assessing 'what Works' in International Development: Meta-Analysis for Sophisticated Dummies*. *Journal of Development Effectiveness*, 4(3), 456–71. DOI: https://doi.org/10.1080/19439342.2012.710642

Filmer, D. and L.H. Pritchett. (2001). *Estimating Wealth Effects without Expenditure Data – or Tears: An Application to Educational Enrollments in States of India*. Demography, 38(1), 115–32.

Filmer, D. and K. Scott. (2012). *Assessing Asset Indices. Demography*, 49(1), 359–92. DOI: https://doi.org/10.1007/s13524-011-0077-5

Green, D.P., S.E. Ha and J.G. Bullock. (2010). *Enough Already about 'Black Box' Experiments: Studying Mediation Is More Difficult than Most Scholars Suppose. Annals of the American Academy of Political and Social Science*, 628(1), 200–8. DOI: https://doi.org/10.1177/0002716209351526

Harris, R. J., M. Bradburn, J. Deeks, R. M. Harbord, D. Altman, T. Steichen and J. Sterne. (2007). *METAN: Stata Module for Fixed and Random Effects Meta-Analysis. Version 3.04*. Retrieved 20 April 2017 from https://ideas.repec.org/c/boc/bocode/s456798.html

Hatløy, A., L.E. Torheim and A. Oshaug. (1998). *Food Variety – a Good Indicator of Nutritional Adequacy of the Diet? A Case Study from an Urban Area in Mali, West Africa. European Journal of Clinical Nutrition*, 52, 891–8. DOI: https://doi.org/10.1038/sj.ejcn.1600662

Higgins, J.P.T. and S. Green (eds.). (2011). *Cochrane Handbook for Systematic Reviews of Interventions version 5.1.0*. London: The Cochrane Collaboration.

Higgins, J.P.T., S.G. Thompson, J.J. Deeks and D.G. Altman. (2003). *Measuring Inconsistency in Meta-Analyses. BMJ*, 327(7414), 557–60. DOI: https://doi.org/10.1136/bmj.327.7414.557

Howe, L.D., J.R. Hargreaves, S. Gabrysch and S.R.A. Huttly. (2009). *Is the Wealth Index a Proxy for Consumption Expenditure? A Systematic Review. Journal of Epidemiology and Community Health*, 63(11), 871–7. DOI: https://doi.org/10.1136/jech.2009.088021

Howe, L.D., J.R. Hargreaves and S.R.A. Huttly. (2008). *Issues in the Construction of Wealth Indices for the Measurement of Socio-Economic Position in Low-Income Countries. Emerging Themes in Epidemiology*, 5, 3. DOI: https://doi.org/10.1186/1742-7622-5-3

Howe, L.D., J.R. Hargreaves, G.B. Ploubidis, B.L. De Stavola and S.R.A. Huttly. (2011). *Subjective Measures of Socio-Economic Position and the Wealth Index: A Comparative Analysis. Health Policy and Planning*, 26(3), 223–32. DOI: https://doi.org/10.1093/heapol/czq043

Hughes, K. and C. Hutchings. (2011). *Can We Obtain the Required Rigour without Randomisation? Oxfam GB's Non-Experimental Global Performance Framework. International Initiative for Impact Evaluation Working Paper 13*. Retrieved 31 January 2017, from http://www.3ieimpact.org/en/publications/working-papers/working-paper-13/

Knueppel, D., M. Demment and L. Kaiser. (2010). *Validation of the Household Food Insecurity Access Scale in Rural Tanzania. Public Health Nutrition*, 13(3), 360–7. DOI: https://doi.org/10.1017/S1368980009991121

Leuven, E. and B. Sianesi. (2003). *PSMATCH2: Stata Module to Perform Full Mahalanobis and Propensity Score Matching, Common Support Graphing, and Covariate Imbalance Testing. Version 4.0.11*. Retrieved 20 April 2017 from https://ideas.repec.org/c/boc/bocode/s432001.html

Melgar-Quinonez, H.R., A.C. Zubieta, B. MkNelly, A. Nteziyaremye, M.F.D. Gerardo and C. Dunford. (2006). *Household Food Insecurity and Food Expenditure in Bolivia, Burkina Faso, and the Philippines. Journal of Nutrition*, 136(5), 1431S – 1437S.

Oya, C., F. Schaefer, D. Skalidou, C. Mc Cosker and L. Langer. (2017). *Effects of Certification Schemes for Agricultural Production on Socio-Economic Outcomes in Low- and Middle-Income Countries: A Systematic Review. Campbell Systematic Reviews* 2017:3. Retrieved 31 May 2017 from https://campbellcollaboration.org/library/agricultural-commodity-production-certification-systems-outcomes.html

Rosenbaum, P.R. and D.B. Rubin. (1983). *The Central Role of the Propensity Score in Observational Studies for Causal Effects. Biometrika*, 70(1), 41–55. DOI: https://doi.org/10.1093/biomet/70.1.41

Rosenbaum, P.R. and D.B. Rubin. (1985). *Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score. The American Statistician*, 39(1), 33–8.

Sahn, D.E. and D. Stifel. (2003). *Exploring Alternative Measures of Welfare in the Absence of Expenditure Data. Review of Income and Wealth*, 49(4), 463–89. DOI: https://doi.org/10.1111/j.0034-6586.2003.00100.x

Sterne, J.A.C. and M. Egger. (2005). *Regression Methods to Detect Publication and Other Bias in Meta-Analysis*. In H. R. Rothstein, A. J. Sutton, & M. Borenstein, eds. *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments*. Chichester: John Wiley & Sons, pp. 99–110. DOI: https://doi.org/10.1002/0470870168.ch6

Stewart, R., L. Langer, N.R. Da Silva, E. Muchiri, H. Zaranyika, Y. Erasmus, N. Randall, S. Rafferty, M. Korth, N. Madinga et al. (2015). *The Effects of Training, Innovation and New Technology on African Smallholder Farmers' Wealth and Food Security: A Systematic Review. Campbell Systematic Reviews* 2015:16. Retrieved 11 August 2017 from https://www.campbellcollaboration.org/library/training-innovation-new-tech-african-smallholder-farmers.html

Sun, S. (2011). *Meta-Analysis of Cohen's Kappa. Health Services & Outcomes Research Methodology*, 11(3–4), 145–63. DOI: https://doi.org/10.1007/s10742-011-0077-3

Swindale, A. and P. Bilinsky. (2006). *Household Dietary Diversity Score (HDDS) for Measurement of Household Food Access: Indicator Guide*. Retrieved 24 August 2016 from http://www.fantaproject.org/monitoring-and-evaluation/household-dietary-diversity-score

Waddington, H., H. White, B. Snilstveit, J.G. Hombrados, M. Vojtkova, P. Davies, A. Bhavsar, J. Eyers, T.P. Koehlmoos, M. Petticrew, J.C. Valentine and P. Tugwell. (2012). *How to Do a Good Systematic Review of Effects in International Development: A Tool Kit. Journal of Development Effectiveness*, 4(3), 359–87. DOI: https://doi.org/10.1080/19439342.2012.711765

Walsh, M. (2016). *Using Interpretive Research to Make Quantitative Evaluation More Effective: Oxfam's Experience in Pakistan and Zimbabwe*. In S. Bell & P. Aggleton, eds. *Monitoring and Evaluation in Health and Social Development: Interpretive and Ethnographic Perspectives*. Abingdon: Routledge, pp. 219–31.

NOTES

1 This initiative is described further in Hughes and Hutchings (2011). In addition to the livelihoods Effectiveness Reviews, other Effectiveness Reviews have sought to evaluate projects' impacts on women's empowerment, resilience, citizen's voice, policy influencing, or humanitarian response. Reports describing the full results for each of the Effectiveness Reviews can be found at www.oxfam.org.uk/effectiveness.

2 Field work was carried out for one additional Effectiveness Review, in Tanzania in 2015/16, but this report was not finalized because of data quality concerns. The data from this Effectiveness Review is therefore not included in the meta-analysis.

3 Nineteen of the evaluations included in this meta-analysis are of projects that were selected for livelihoods-specific Effectiveness Reviews. The five additional evaluations are of projects selected for evaluation under the women's empowerment or resilience themes, but where data on household consumption was also collected and these results have been aggregated with the other 19. It should be noted that the results of the 23 evaluations are not fully representative of Oxfam's livelihoods support projects. In a small number of cases, projects that were randomly selected were not evaluated because they were considered not yet to be mature enough to show significant impact, or had had an impact evaluation conducted in the recent past (albeit without using a comparison group). Among the 23 projects that were evaluated, it was necessary in many cases to select specific project components or specific geographic areas to be included in the evaluation. These decisions were generally made on the grounds of evaluability (for example, based on locations where suitable comparison groups could be identified) and on the potential for maximizing learning (for example, learning around project components that were being considered for scale-up), rather than with the aim of maximizing representativeness.

4 The lack of publication bias is confirmed by applying an Egger test (Sterne and Egger, 2005) to the main outcomes considered in this meta-analysis. That is, the standard normal deviates of the effect estimates were regressed on the inverse of their standard errors across the 23 datasets, and the coefficient of the intercept was found not to be statistically significantly different from zero ($p > 0.1$).

5 This paper refers to 'intervention' and 'comparison' groups rather than 'treatment' and 'control' groups. The term 'comparison' seems more appropriate than 'control' in the context of quasi-experimental evaluations in which the comparison group is identified *ex-post*.

6 The evaluation carried out in Zimbabwe in 2011/12 has subsequently been criticized on the basis that the comparison respondents had benefited from agricultural inputs and employment opportunities provided by the project (Walsh, 2016). In principle, the provision of agricultural inputs to the comparison households should not have had an effect on the outcomes estimated in that evaluation because they were provided during the farming season in which the evaluation was conducted; the crops for which they were used had not yet been harvested at the time of the survey. However, the employment opportunities may well have already had a positive effect on income among the comparison group: if so, this would imply that the evaluation underestimated the impact of the project on income and welfare.

7 This approach to recall of food consumption data is approximately the same as that found to be most effective in an experiment in Tanzania by Beegle *et al.* (2012).

8 Following the guidance of Deaton and Zaidi (2002: pp.50–51), the number of adult equivalent household members is calculated using the formula $(A + \frac{1}{3}K)0.9$, where A stands for the number of adults in the household, and K the number of children. This assumes that children on average have consumption needs on average of a third of those of adults, and that there are modest economies of scale within the household. The sensitivity of our results to these assumptions is tested in Section 5.1.

9 That is, principal component analysis is used, with the first principal component being assumed to represent household wealth. Evaluations differed on whether information on asset ownership was collected in binary form (a simple yes/no response to whether the household owned each type of asset) or whether the number of each asset type owned was recorded. Where available, data on the quantity of each asset type was used in deriving the wealth index. The details of the procedure used to calculate the wealth index varied over time: for the purposes of this meta-analysis, wealth indices from evaluations carried out in earlier years have been re-estimated so as to be consistent with the approach applied in later years. Histograms for the wealth index for each evaluation are shown in Appendix 4. It can be seen that the wealth indices have an approximately normal distribution in most of the evaluation datasets, though there is some evidence of 'clumping' of observations at specific values in some datasets.

10 Meta-analysis is carried out using the *metan* module in Stata (Harris *et al.*, 2007).

11 An exception is the analysis for the outcome shown in Figure 9, which is based on a binary variable (whether each household made any sales of agricultural products). This variable was retained in its original form in the the analysis, so the project effects shown in Figure 8 are expressed in terms of a percentage-point difference between the intervention and comparison groups.

12 For analyses involving the estimation of project effects, standardization is carried out by dividing by the estimated standard deviation from the pooled sample of intervention and comparison observations after matching. For the analysis in Section 5, in which the relationship between two variables is tested, standardization is carried out by subtracting the sample mean and dividing by the estimated standard deviation across the entire sample.

13 The $I2$ statistic is defined as $(Q - df)/Q$, where Q is Cochran's heterogeneity statistic, the weighted sum of the squared deviations of each study's estimate from the meta-analysis estimate, and df represents

the degrees of freedom. The interpretation of this statistic is discussed in Higgins *et al.* (2003) and Higgins and Green (2011).

14 An alternative approach would be to weight each evaluation by the size of the sample frame in the intervention group for that evaluation, so that the overall effect represents the average effect across all those targeted as beneficiaries by the 23 projects. Unfortunately, there does not seem to be any guidance in the literature on how to combine a random-effects model with externally defined weights.

15 This was tested formally by carrying out a meta-regression of effect size on dummy variables representing three of the four regions shown in Figure 1. The *F*-test for joint significance of the dummy variables was not passed.

16 A meta-regression for effect size on a dummy variable coded as 1 for projects in middle-income countries (according to the World Bank classification) and 0 for those in lower-income countries produces a coefficient of –0.01, with a 95% confidence interval of (–0.19, 0.17).

17 A meta-regression for effect size on the scale of the project (defined by the number of households in the sample frame for the intervention group for the evaluation) produces a coefficient of –0.000007 with a 95% confidence interval of (–0.000031, 0.000045).

18 Data on the expenditure on the project interventions evaluated was not available in a consistent format. The data available in many cases included expenditure on components of the project (such as advocacy work) that were not included in the evaluation, or on activities in geographic areas that were not covered by the evaluation. An additional limitation to these data is that in several cases the evaluations covered the cumulative impacts of a series of two or more projects, but expenditure data was available only for the most recent project in the series.

Despite these limitations, the total expenditure figure for each project was used to create four estimates of expenditure on the activities evaluated, thought to be an underestimate and an overestimate of total expenditure on the activities, and an underestimate and overestimate of expenditure per household. Total project expenditure had a range of approximately £110,000 to £2,000,000. The median was £230,000 using the low estimate or £350,000 using the high estimate. Expenditure per participant household had a range of approximately £50 to £9,700, with a median of approximately £750 (these figures do not vary significantly between the lower and higher estimates). When included in meta-regression models, none of the four expenditure estimates showed any significant relationship with the size of the project effect.

19 A meta-regression for effect size on approximate project duration (measured in years) produces a coefficient of 0.018 with a 95% confidence interval of (–0.009, 0.044).

20 A further systematic review of agricultural training and technology interventions found a (non-significant) effect from agricultural training interventions of 0.12 standard deviations, and an effect from the promotion of new agricultural technologies of 0.26 standard deviations (Stewart *et al.*, 2015). However, the outcomes considered here are measured mostly by the value of crops produced (and in most cases only for one specific crop type), rather than by overall (net) household income. These measures correspond to what are considered as 'intermediate outcomes', discussed in Section 4.

21 Regressing the project effect on household consumption on the project effect on change in the wealth index (or vice versa) across across the 23 datasets produces a model with an $R2$ coefficient of 0.23.

22 Like some of the Oxfam projects, the graduation schemes discussed by Banerjee *et al.* (2015) involved distributing assets to participant households.

23 The meta-analysis described in this section is based on linear regression models of the form

$$Yi = \beta 0 + \beta 1 Fi \times Di + \beta 2 Fi + \beta 3 Di + \varepsilon i$$

where i is a household identifier, Yi is the outcome measure (household consumption per adult equivalent), Fi is a dichotomous variable that takes the value of 1 for female-headed households and 0 for male-headed households, and Di is a dichotomous variable that takes the value of 1 for households included in the intervention group for each evaluation, and 0 for households in the comparison group. Observations in each of the regression models are weighted by the propensity-score model used in the analysis shown in Figures 1 and 2. The coefficient $\beta 3$ is estimated separately for each evaluation; these estimates are then used as input into a meta-analysis model, using random effects with inverse variance weighting.

24 In most of the evaluations, little guidance was provided to respondents in identifying the head of household, other than that the head should be a current member of the household (and, in particular, could not be a migrant worker who spends most of her or his time living outside the household). It is possible that, if the project interventions had had a significant impact on intra-household relations (for example, through empowering women), then this may have affected which individual was identified as the head of household at the time of the survey. If so, this could result in bias between the intervention and comparison groups in the gender of the head of household. It is thought that any such bias would be small, though it is not possible to check this with the data available.

In the evaluation conducted in Ethiopia an individual head of household was not identified at the time of the survey. This evaluation is therefore excluded from the analysis in this section.

25 This analysis was repeated after including data from all Effectiveness Reviews that were carried out between 2011 and 2015 under the 'resilience' theme, as well as those carried out under the theme of 'livelihoods' – that is, 37 evaluations in total. The overall difference in project impact between female-headed and male-headed households was smaller than that reported here, at 0.10 standard deviations, but it is still statistically significant at the five percent level.

26 In our data, female-headed households are found to be significantly worse off than male-headed households (by 0.18 standard deviations), in terms of the index of pre-project wealth indicators. A similar

(in fact, larger) difference is seen between female-headed and male-headed households in terms of their wealth indicators at the time of the survey (with the analysis restricted to the comparison group so as to avoid confounding with the effects of the projects being evaluated), but not in terms of household consumption.

27 The allocation of survey respondents to pre-project wealth quintiles is made across the whole sample, not only across the project participants. In cases in which the project participants differ systematically from the comparison respondents in terms of their wealth indicators, the proportion of project participants in each quintile can be considerably greater or less than 20%.

The meta-analysis described in this section is based on linear regression models of the form

$$Yij = \beta 0j + \beta 1jWij \times Di + \beta 2jWi1 + \beta 3jWi2 + \beta 4jWi3 + \beta 5jWi4 + \beta 6jDij + \varepsilon i \ , \ j = 1, \ldots, 5$$

where *i* is a household identifier, *j* is an identifier of the quintile of the baseline wealth index, *Yij* is the outcome measure (household consumption per adult equivalent), W*ij* is a dichotomous variable that takes the value of 1 for households estimated to be in the *j*th quintile of the baseline wealth index and 0 for households not estimated to be in that quintile, and *Di* is a dichotomous variable that takes the value of 1 for households included in the intervention group for each evaluation, and 0 for households in the comparison group. Observations in each of the regression models are weighted by the propensity scores used for the analysis shown in Figures 1 and 2. For each baseline wealth quintile *j*, the coefficient $\beta 1j$ is estimated separately for each evaluation; these estimates are then used as input into a meta-analysis model, using random effects with inverse variance weighting.

28 For the majority of the evaluations listed in Figure 7, survey respondents were also asked to recall the range of crop types they were producing from a notional pre-project period. This recalled data can be used to provide pseudo difference-in-difference estimates of the effect of projects on crop diversity. Carrying out meta-analysis on those difference-in-difference estimates suggests that the projects overall increased crop diversity by 0.14 standard deviations (with a 95% confidence interval of (0.03, 0.24)), larger than the 0.06 standard deviations reported in Figure 7. However, the most noticeable change in the difference-in-difference estimates for the specific projects is that the apparent impacts of the projects in Haiti, Armenia (in 2015/16), and Zimbabwe (in 2013/14) are eliminated or considerably reduced in size. Of course, it is not known how accurate the pre-project recall data are, and consequently whether the single-difference or difference-in-difference estimates are a better reflection of reality.

29 Unlike the other indicators for which meta-analysis is carried out in this report, the outcome measure here is a binary variable, and has not been standardized before analysis. The analysis is therefore in terms of 'risk difference', as suggested by Borenstein *et al.* (2009) and Higgins and Green (2011).

30 The use of Cohen's kappa in this respect follows Howe *et al.* (2008; 2011). A weighted kappa statistic is used, such that the weight decreases linearly as the number of quintiles by which the two outcome measures disagree increases. Meta-analysis of the kappa statistics derived from each dataset is carried out according to the procedure described by Sun (2011).

31 The measure being considered here is the total number of individual food items for which the respondent reported there having been some consumption during the seven days prior to the survey. This measure is distinct from the Household Dietary Diversity Indicator (Swindale and Bilinsky 2006), which instead measures the frequency of consumption of various food *groups*.

32 The exception is the evaluation from Somalia, which included only three such indicators. The evaluation in Sri Lanka included an alternative measure of food security – a count of the number of months the household had experienced food shortages during the previous year. Because this measure is of a different nature to that used in the other evaluations, the results from Sri Lanka are not included in this meta-analysis.

33 Knueppel *et al.* (2010) and Deitchler *et al.* (2010) find correlations between similar food security scales and indices of wealth indicators, in data from Tanzania and Mozambique respectively. As shown in Table 3, the kappa statistic for agreement between the food security measure and the wealth index is only 0.13.

34 A meta-regression for the kappa statistic for agreement between the wealth index and consumption per adult equivalent, on a dummy variable coded as 1 for projects in middle-income countries (according to the World Bank classification) and 0 for those in lower-income countries, produces a coefficient of –0.01, with a 95% confidence interval of (–0.08, 0.06).

35 The number of indicators making up the wealth indices in the Oxfam evaluations ranges from 12 to 74. Regressing the kappa statistic for agreement between the wealth index and household consumption per adult equivalent on the number of wealth indicators produces a coefficient of –0.001, with a 95% confidence interval of (–0.003, 0.002).

36 Other evaluations included similar questions, but relating to change in income experienced from particular sources, such as from sales of the particular agricultural products on which the project was focusing, or from agricultural sales overall. Unfortunately, there is little consistency between the evaluations in the questions that were asked, so the results are not suitable for meta-analysis.

37 All respondents were informed at the start of the survey that no support would come to them or their households as a result of their responses to the survey questions. However, interviewers in several of the evaluations expressed concerns that respondents may have had that impression anyway.

38 Appendix 6 shows that, in 13 of the 23 evaluations (including five of the six evaluations in Latin America and the Caribbean, and six of the nine carried out in Africa), the intervention group was found to be significantly better off than the comparison group in terms of the index of pre-project wealth indicators. Only in three cases (in Indonesia, Pakistan and Georgia) was the comparison group significantly better off than the intervention group. In cases in which the project participants were self-selected but the comparison groups were selected at random from the population, having project participants that are

better off than the comparison group is suggesting that the project did not reach the poorest households in the implementation area.

39 The second stage, of testing matching variables' power to predict the outcome variable, was only applied in cases in which sufficiently accurate matching models could not be constructed using the full set of variables derived from the first stage. The threshold for significance for excluding variables from the matching models in these stepwise processes was defined for each evaluation, but was between 0.2 and 0.3 in all cases.

40 In several of the evaluations, PSM was carried out separately in two geographical regions or among two subgroups. In these cases, the results of the two regions or subgroups were aggregated using a fixed-effects meta-analysis model, weighted by the sampling weights.

41 In most of the evaluations, alternative matching procedures and regression models were also applied in order to test the robustness of the results obtained from the kernel matching procedure. The matching models obtained through the kernel procedure were generally found to have lower mean bias in the matching variables than those obtained through other matching procedures, so only the kernel models are included in this meta-analysis.

Research reports

This research report was written to share research results, to contribute to public debate, and to invite feedback on development and humanitarian policy and practice. It does not necessarily reflect the policy positions of the publishing organizations. The views expressed are those of the author and not necessarily those of the publishers.

For more information, or to comment on this report, email policyandpractice@oxfam.org.uk

The information in this publication is correct at the time of going to press.

Published by Oxfam GB for Oxfam International under ISBN 978-1-78748-109-1 in November 2017
Oxfam GB, Oxfam House, John Smith Drive, Cowley, Oxford, OX4 2JY, UK.
DOI: 10.21201/2017.1091

OXFAM

Oxfam is an international confederation of 20 organizations networked together in more than 90 countries, as part of a global movement for change, to build a future free from the injustice of poverty. Please write to any of the agencies for further information, or visit www.oxfam.org.

Oxfam America (www.oxfamamerica.org)
Oxfam Australia (www.oxfam.org.au)
Oxfam-in-Belgium (www.oxfamsol.be)
Oxfam Brasil (www.oxfam.org.br)
Oxfam Canada (www.oxfam.ca)
Oxfam France (www.oxfamfrance.org)
Oxfam Germany (www.oxfam.de)
Oxfam GB (www.oxfam.org.uk)
Oxfam Hong Kong (www.oxfam.org.hk)
Oxfam IBIS (Denmark) (http://oxfamibis.dk/)
Oxfam India (www.oxfamindia.org)
Oxfam Intermón (Spain) (www.oxfamintermon.org)
Oxfam Ireland (www.oxfamireland.org)
Oxfam Italy (www.oxfamitalia.org)
Oxfam Japan (www.oxfam.jp)
Oxfam Mexico (www.oxfammexico.org)
Oxfam New Zealand (www.oxfam.org.nz)
Oxfam Novib (Netherlands) (www.oxfamnovib.nl)
Oxfam Québec (www.oxfam.qc.ca)
Oxfam South Africa (www.oxfam.org.za)

OXFAM