Simple Poverty Scorecard® Poverty-Assessment Tool Ethiopia

Mark Schreiner and Shiyuan Chen

10 April 2009

This document is at `SimplePovertyScorecard.com`.

Abstract

The Simple Poverty Scorecard-brand poverty-assessment tool uses 11 low-cost indicators from Ethiopia's 2004/5 Household Income, Consumption and Expenditure Survey (HICE) and 2004 Welfare Monitoring Survey (WMS) to estimate the likelihood that a household has expenditure below a given poverty line. Field workers can collect responses in about ten minutes. The scorecard's accuracy is reported for a range of poverty lines. The scorecard is a practical way for pro-poor programs in Ethiopia to measure poverty rates, to track changes in poverty rates over time, and to segment clients for targeted services.

Acknowledgements

This paper was funded by Terrafina. Data come from Ethiopia's Central Statistical Agency. Thanks go to Malika Anand, Gabrielle Athmer, Ellie Bosch, Dagnew Gessese, Harm Haverkort, Getachew Mekonnin, Mariel Mensink, Nadia Ouriemchi, Don Sillers, Cor Wattel, and Teshome Yohannes.

Simple Poverty Scorecard® Poverty-Assessment Tool

			Name	Identifier
Interview ID:				
Interview date:		Participant:		
Country:	ETH	Field agent:		
Scorecard:	001	Service point:		
Sampling wgt.:		Number of household members:		

Indicator	Value	Points	Score
1. How many people are in the household?	A. Six or more	0	
	B. Five	10	
	C. Four	20	
	D. Three	28	
	E. Two or one	45	
2. Do all children ages 6 to 12 attend school?	A. No	0	
	B. Yes	1	
	C. No children ages 6 to 12	3	
3. Excluding kitchen and toilets, how many rooms does the dwelling unit have?	A. One	0	
	B. Two	1	
	C. Three or more	5	
4. What is the main construction material of the walls of the dwelling unit?	A. Wood and grass, mud and stone, or other	0	
	B. Wood and mud, reeds and bamboo, cement and stone, hollow blocks. or bricks	5	
5. What type of toilet facility does the household use?	A. Pit latrine (shared), field/forest, container (household utensils), or other	0	
	B. Pit latrine (private)	4	
	C. Flush toilet (private or shared)	9	
6. What is the main source of cooking fuel?	A. Mainly firewood (purchase or collected), animal dung, or other	0	
	B. Crop residue	3	
	C. Charcoal, kerosene, butane gas, electricity, or does not use fuel	5	
7. Does the household currently own any mattresses and/or beds?	A. No	0	
	B. Yes	5	
8. Does the household currently own any radios?	A. No	0	
	B. Yes	6	
9. Does the household currently own any watches or clocks?	A. No	0	
	B. Yes	5	
10. Does the household currently own any cattle, sheep, or goats?	A. No	0	
	B. Yes	10	
11. Does the household currently own any jewelry (gold/silver)?	A. No	0	
	B. Yes	2	

SimplePovertyScorecard.com **Score:**

Simple Poverty Scorecard® Poverty-Assessment Tool Ethiopia

1. Introduction

The Simple Poverty Scorecard poverty-assessment tool is a low-cost way for pro-poor programs in Ethiopia to estimate the likelihood that a household has per-capita expenditure below a given poverty line, to monitor groups' poverty rates at a point in time, to track changes in groups' poverty rates between two points in time, and to target services to households.

The direct approach to poverty measurement via surveys is difficult and costly. In Ethiopia, enumerators visited households twice a week for two months (16 visits in all), asking each time about a lengthy list of consumption items.

In contrast, the indirect approach via scoring is simple, quick, and inexpensive. In a single visit, it uses 11 verifiable indicators (such as "What is the main source of cooking fuel?" or "Does the household currently own any radios?") to get a score that is highly correlated with poverty status as measured by the exhaustive survey.

The scorecard here differs from "proxy means tests" (Coady, Grosh, and Hoddinott, 2002) in that it is tailored to the capabilities and purposes not of national governments but rather of local, pro-poor organizations. The feasible poverty-measurement options for these organizations are typically subjective and relative (such as participatory wealth ranking by skilled field workers) or blunt (such as rules based

on land-ownership or housing quality). Results from these approaches are not comparable across organizations nor across countries, they may be costly, and their accuracy is unknown.

Suppose an organization wants to know what share of its participants are below a poverty line, say, USD1.25/day at 2005 purchase-power parity for the Millennium Development Goals, or the line defined the poorest half of those below the national poverty line as required of USAID microenterprise partners. Or suppose an organization—such as the Microcredit Summit Campaign—wants to measure movement across a poverty line. For these tasks, organizations need expenditure-based, objective tools with known accuracy. While expenditure surveys are costly even for governments, scorecards can be inexpensive enough to be used by small, local organizations for monitoring, management, and targeting.

The statistical approach here aims to be understood by non-specialists. After all, if managers are to adopt the scorecard on their own and apply it to inform their decisions, they must first trust that it works. Transparency and simplicity build trust. Getting "buy-in" matters; proxy means tests and regressions on the "determinants of poverty" have been around for three decades, but they are rarely used to inform decisions, not because they do not work, but because they are presented (when they are presented at all) as tables of regression coefficients incomprehensible to lay people (with cryptic indicator names such as "`HHSIZE_2`", negative values, and many decimal places).

Thanks to the predictive-modeling phenomenon known as the "flat max", simple scorecards can be about accurate as complex ones.

The technical approach here is also innovative in how it associates scores with poverty likelihoods, in the extent of its accuracy tests, and in how it derives formulas for standard errors. Although these techniques are simple and/or standard, they have rarely or never been applied to proxy means tests.

The scorecard (Figure 1) is based on the 21,297 households covered by both the 2004/5 Household Income, Consumption and Expenditure Survey (HICE) and 2004 Welfare Monitoring Survey (WMS) conducted by Ethiopia's Central Statistical Agency (CSA). Indicators are selected to be:

- Inexpensive to collect, easy to answer quickly, and simple to verify
- Strongly correlated with poverty
- Liable to change over time as poverty status changes

All points in the scorecard are non-negative integers, and total scores range from 0 (most likely below a poverty line) to 100 (least likely below a poverty line). Non-specialists can collect data and tally scores on paper in the field in five to ten minutes.

The scorecard can be used to estimate three basic quantities. First, it can estimate a household's "poverty likelihood", that is, the probability that the household has per-capita expenditure below a given poverty line.

Second, the scorecard can estimate the poverty rate of a group of households at a point in time. This is simply the average poverty likelihood among the households in the group.

Third, the scorecard can estimate changes in the poverty rate for a group of households between two points in time. This estimate is defined as the change in the average poverty likelihood of the households in the group over time.

The scorecard can also be used for targeting. To help managers choose a targeting cut-off, this paper reports several measures of targeting accuracy for a range of possible cut-offs.

This paper presents a single scorecard (Figure 1) whose indicators and points are derived from household expenditure data and the USD1.25/day (2005 PPP) poverty line. Scores from this scorecard are calibrated to poverty likelihoods for four poverty lines.

The scorecard is constructed and calibrated using a sub-sample of the data from the 2004/5 HICE and 2004 WMS. Its accuracy is validated on a different sub-sample. While all three scoring estimators are unbiased when applied to the population they were derived for (that is, they match the true value on average in repeated samples from the same population from which the scorecard was built), they are—like all predictive models—biased to some extent when applied to a different population.[1]

Thus, while the indirect scoring approach is less costly than the direct survey approach, it is also biased. (The survey approach is unbiased by assumption.) There is bias because scoring must assume that the relationship between indicators and poverty

[1] For example, a nationally representative sample at a different point in time or a non-representative sub-group (Tarozzi and Deaton, 2007).

will be the same in the future as in the data used to build the scorecard.[2] Of course, this assumption—ubiquitous and inevitable in predictive modeling—holds only partly.

When applied to the validation sample for Ethiopia, the absolute difference between scorecard estimates of groups' poverty rates and the true rates is 1.4 percentage points for the USD1.25/day 2005 PPP line and 1.3 percentage points on average across all four lines. These differences are due to sampling variation and not bias; the average difference would be zero if the whole 2004/5 HICE and 2004 WMS were to be repeatedly redrawn and divided into sub-samples before repeating the entire scorecard-building process.

For sample sizes of n = 16,384, the 90-percent confidence intervals for these estimates are +/–0.6 percentage points or less. For n = 1,024, the 90-percent intervals are +/–2.5 percentage points or less.

Section 2 below describes data and poverty lines. Section 3 places the new scorecard here in the context of existing exercises for Ethiopia. Sections 4 and 5 describe scorecard construction and offer practical guidelines for use. Sections 6 and 7 detail the estimation of households' poverty likelihoods and of groups' poverty rates at a point in time. Section 8 discusses estimating changes in poverty rates. Section 9 covers targeting. The final section is a summary.

[2] Bias may also result from changes in the quality of data collection, from imperfect adjustment of poverty lines across time or geographic regions, or from sampling variation across expenditure surveys.

2. Data and poverty lines

This section discusses the data used to construct and test the scorecard. It also presents the poverty lines to which scores are calibrated.

2.1 Data

The scorecard is based on data from the 21,297 households 2004/5 HICE and 2004 WMS. Households are randomly divided into three sub-samples (Figure 2):

- *Construction* for selecting indicators and points
- *Calibration* for associating scores with poverty likelihoods
- *Validation* for testing accuracy on data not used in construction or calibration

2.2 Poverty rates and poverty lines

2.2.1 Rates

As a general definition, the poverty rate is the share of people in a given group who live in households whose total household expenditure divided by the number of household members is below a given poverty line.

Beyond this general definition, there two special cases, household-level poverty rates and person-level poverty rates. With household-level rates, each household is counted as if it had only one person, regardless of true household size, so all households are counted equally. With person-level rates (the "head-count index"), each household is weighted by the number of people in it, so larger households count more.

Consider, for example, a group of two households, the first with one member and the second with two members. Suppose further that the first household has per-capita expenditure above a poverty line (it is "non-poor") and that the second household has per-capita expenditure below a poverty line (it is "poor"). The household-level rate counts both households as if they had only one person and so gives a poverty rate of 1 ÷ (1 + 1) = 50 percent. In contrast, the person-level rate weights each household by the number of people in it and so gives a poverty rate of 2 ÷ (1 + 2) = 67 percent.

Whether the household-level rate or the person-level rate is relevant depends on the situation. If an organization's "participants" include all the people in a household, then the person-level rate is relevant. Governments, for example, are concerned with the well-being of people, regardless of how those people are arranged in households, so they typically report person-level poverty rates.

If an organization serves one person per household, however, then the household-level rate is relevant. For example, if a microfinance organization serves only one person in a household, then it might prefer to report household-level poverty rates.

Based on the 2004/5 HICE and 2004 WMS, this paper reports (Figure 3) household-level poverty rates and person-level poverty rates by region for Ethiopia. The scorecard here is constructed using household-level rates, scores are calibrated to household-level poverty likelihoods, and accuracy is measured for household-level rates. This use of household-level rates reflects the belief that they are the relevant measure for most pro-poor organizations.

Still, organizations can estimate person-level poverty rates by taking a household-size-weighted average of the household-level poverty likelihoods. It is also possible to construct a scorecard based on person-level rates, calibrate scores to person-level poverty likelihoods, and measure accuracy for person-level rates, but it has not been done here.

2.2.2 Poverty lines

Ethiopia has not established any official poverty lines. Based on the 1995/6 HICE and information from Ethiopia's CSA, Dercon (1997) calculates a food poverty line and a "total" poverty line of 1.77 Birr and 2.95 Birr per adult equivalent per day at 1995/6 prices. Unfortunately, there are no series of regional and temporal price indices adequate and available that would permit the conversion of Dercon's lines to 2004/5 prices. Other major poverty documents (Woldehanna, Hoddinott, and Dercon, 2008; Ministry of Finance and Economic Development, 2006 and 2002) apparently use Dercon's (1997) lines, converting expenditure to units in 1995/6 prices, but without providing enough information to replicate that conversion. After almost a year of inquiries with CSA and other Ethiopian poverty specialists, it was decided that, due to lack of other options, this document would use international poverty lines in dollars at 2005 purchase-power parity, unadjusted for regional differences in the cost-of-living.

The scorecard is constructed based on the USD1.25/day 2005 PPP line. Because local pro-poor organizations may want to use different or various poverty lines, this paper calibrates scores from its single scorecard to poverty likelihoods for four lines:

- USD1.00/day 2005 PPP
- USD1.25/day 2005 PPP
- USD1.75/day 2005 PPP
- USD2.50/day 2005 PPP

The USD1.25/day 2005 PPP line is derived from:

- 2005 PPP exchange rate for "individual consumption expenditure by households" (International Comparison Project, 2008): Birr2.75 per USD1.00
- Consumer Price Indices (CPI) for July 2004 (76.3) and January 2005 to December 2005 (78.5, 78.9, 80.2, 81.4, 81.6, 84.2, 85.1, 85.6, 86.9, 86.2, 86.1, and 85.3).[3] The average CPI for 2005 is 83.33.

The USD1.25/day 2005 PPP line for Ethiopia on average in July 2005 is:[4]

$$(\text{2005 PPP exchange rate}) \cdot \text{USD1.25} \cdot \left(\frac{\text{CPI}_{\text{July 2005}}}{\text{CPI}_{\text{Ave. 2005}}} \right) =$$

$$\left(\frac{\text{Birr2.75}}{\text{USD1.00}} \right) \cdot \text{USD1.25} \cdot \left(\frac{85.1}{83.33} \right) = \text{Birr3.51}.$$

The USD1.00/day, USD1.75/day and USD2.50/day 2005 PPP lines are multiples of the USD1.25/day 2005 PPP line.

The 2004/5 HICE took place in two rounds, from July 4 to August 3, 2004, and then from February 4 to March 5 2005. The CPIs for July 2004, February 2005, and July 2005 are used to convert expenditure and poverty lines to July 2005 prices.

[3] The CPIs are from Loening, Durevall, and Birru (2008).
[4] Sillers (2006) provides this formula.

3. Context of poverty-assessment tools for Ethiopia

This section reviews existing scorecards for Ethiopia. The new scorecard is different because it estimates households' poverty likelihoods for several expenditure-based poverty lines, because it tests accuracy on data not used in construction, because it uses more recent data (and nationally representative data), and because it reports accuracy and sample-size formulas for a range of scoring purposes.

3.1 Studies of "Determinants of Poverty"

Three studies (Bigsten *et al.*, 2003; Tafesse, 2003; and Hagos and Holden, 2003) seek "poverty determinants" to guide policy. They use regression to relate poverty with indicators from old, small, non-nationally representative surveys.

None of the three papers mention that while their indicators determine poverty, poverty may also determine their indicators. This reduces their value as a guide to policy. For example, Hagos and Holden, upon finding that receipt of food aid is linked with greater poverty, surmise that food aid may create disincentives for households to try to leave poverty (p. 27). They do not discuss the possibility that poorer households receive food aid because food aid it purposely targeted to poorer households.

Furthermore, the scorecards cannot be easily used by non-specialists in local organizations. Some indicators are complex, and a ready-to-use scorecard is not presented. For example, Bigsten *et al.* and Tafesse paste regression results from their

statistical software straight into their papers; points have seven decimal places, may be negative, and the wording of the indicators is not reported.

In general, these papers' "policy recommendations" are banal. For example, Tafesse says that his analysis shows that "fundamental change is needed to improve productivity in the agricultural sector", and he recommends "revisiting policy to help arrest declining trend in agricultural value added per capita and create dynamism in the sector" and introducing "measures conducive to a market environment" (p. 6). No one disagrees that these are worthy goals, but the real question is how to achieve them.

In sum, these three papers are examples of a dead-end genre in poverty analysis: run regressions, rediscover well-known correlations between policy and poverty, and—without providing any new motivations or tools—exhort governments to do better job at what governments already know they should be doing. Of course, governments struggle with poverty alleviation not because they do not know, for example, that agricultural productivity is important, but rather because they face technical, financial, and—most important—organizational/institutional constraints.

3.2 Gwatkin *et al.*

Gwatkin *et al.* (2007) apply to Ethiopia an approach used by USAID in 56 countries with Demographic and Health Surveys (Rutstein and Johnson, 2004). They use Principal Components Analysis to make a "wealth index" from simple, low-cost indicators available for the 13,721 households in Ethiopia's 2005 DHS. The index is like

the scorecard here except that, because it is based on a relative definition of poverty, its accuracy is unknown, and it can only be assumed to be a proxy for long-term wealth/economic status.[5] Important examples of the PCA-index approach are Stifel and Christiaensen (2007), Zeller *et al.* (2006), Sahn and Stifle (2003 and 2000), Devereux and Sharp (2003, see below), and Filmer and Pritchett (2001).

The 23 indicators in Gwatkin *et al.* are similar in their simplicity and verifiability to those in the scorecard here:

- Characteristics of the residence:
  - — Presence of electricity
  - — Source of drinking water
  - — Type of fuel for cooking
  - — Type of toilet arrangement
  - — Type of floor
  - — Type of roof
- Number of people per sleeping room
- Whether any household members work agricultural land
- Ownership of agricultural assets:
  - — Crop land
  - — Cash crops
  - — Cattle or camels
  - — Horses, mules, or donkeys
  - — Sheep or goats
- Ownership of consumer durables:
  - — Radio
  - — Television
  - — Telephone
  - — House

[5] Still, because their indicators are similar and because the "flat max" is important, carefully built PCA indices and expenditure-based poverty-assessment tools probably pick up the same underlying construct (such as "permanent income", see Bollen, Glanville, and Stecklov, 2007), and they probably rank households much the same. Tests of how well PCA indices predict expenditure include Filmer and Scott (2008), Lindelow (2006), Wagstaff and Watanabe (2003), and Montgomery *et al.* (2000).

— Bicycle
— Motorcycle or scooter
— Car or truck
— Electric griddle for making *injera*
— Kerosene lamps or pressure lamps
— Beds or tables

Gwatkin *et al.* has three basic goals for the PCA-based wealth index:

- Segment people by quintiles in order to see how health, population, and nutrition vary with socio-economic status
- Monitor (via exit surveys) how well health-service points reach the poor
- Measure coverage of services via small-scale local surveys

These last two goals are like the monitoring and targeting goals here, and the first goal of ranking households by quintiles is akin to targeting. As here, Gwatkin *et al.* present the index in a format that is ready take to the field, although their index is more difficult to use because it has two pages, all the points have 5 decimal places, and some points are negative.

The central contrast between the scorecard here and the PCA index is the use/non-use of an absolute, expenditure-based poverty line. Thus, while both approaches can rank households, only the scorecard can estimate quantitative, expenditure-based poverty status. Furthermore, relative accuracy (that is, ability to rank or target) is tested here more completely here than in Gwatkin *et al.* (where it is not explicitly tested at all); generally, discussion of the accuracy of PCA indices rests on how well they correlate with health, education, or self-assessed poverty.

3.3 Devereux and Sharp

Devereux and Sharp (2003) suggest that the measured decrease in poverty in Ethiopia since the early 1990s (attributed by Dercon—2002 and 2000—to improved governance and economic liberalization) might be an artifact of sampling or non-universal. To this end, they survey 2,127 households in the Wollo region in 2001/2, collecting "objective indicators of basic needs and livelihood resources, plus one more holistic indicator of household (in)dependence based on self-assessment" (p. 16). They defined as "destitute" those who say they are "unable to meet household needs by own efforts: dependent on support from community or government (could not survive without it)" (p. 17).

Given this, Devereux and Sharp build a 15-indicator PCA-based index and compare its scores to self-assessed destitution. They find that 95 percent of the destitute fall in the bottom two quintiles of the PCA index, and they define these cases as "poor".

Finally, Devereux and Sharp ask respondents to recall indicators in the PCA index and to self-assess destitution as of one, two, and ten years ago. They then measure change over time, concluding that poverty by their definition in Wollo more than doubled since 1990.

How does Devereux and Sharp's PCA index compare to the scorecard here? Both can be used for targeting, measuring poverty rates, and measuring change in poverty rates. Their definition of poverty is not linked to expenditure but rather to self-assessed (in)dependence, which is a valid benchmark, but qualitative and rarely used.

The main difference is that Devereux and Sharp's index is more difficult for non-specialists to implement in a local, pro-poor organization. In particular, their 15 indicators are generally more complex and costly to verify:

- Agricultural assets:
  - Whether any livestock are owned
  - Whether any plough oxen are owned
  - Whether less than half a hectare of land is owned
  - Whether less than half a hectare of land is cultivated
- Labor capacity:
  - Whether there are less than two adult-equivalent potential workers
  - Whether there are any adult male potential workers
  - Whether there is access to non-household labor
- Basic needs:
  - Whether the roof and walls are of poor quality
  - Whether "basic items" are present in the home
  - Whether clothes were bought less than three times in the past three years
  - Whether the household ate one or no times in a day in the worst month in the past year
  - Whether the household had food shortages in three or more months in the past year
- Social capital:
  - Whether the household has social support networks that offer help
  - Whether the household participates in any social institutions
- Whether the household receives formal or informal credit in cash or any cash gifts or remittances

These indicators involve subjective judgments (What is "access" to non-household labor? What is a "poor quality" roof or wall?), non-verifiable reports of past events (food consumed and clothes purchased), calculations (adult-equivalent workers), and sensitive issues (receipt of cash).

While Devereux and Sharp's PCA index is not as simple or inexpensive as the scorecard here, their paper nevertheless breaks new ground in poverty measurement,

effectively combining quantitative methods (PCA index) with qualitative methods (self-assessment of (in)dependence) as well as using recall to measure changes in poverty.

3.4 Vyas and Kumaranayake

Vyas and Kumaranayake (2006) is billed as a "how-to" primer on PCA indices. As a running example, they use urban and rural indices from the 14,072 households in Ethiopia's 2000 DHS. Vyas and Kumaranayake's indicators resemble those here and in Gwatkin *et al.* (2007) in that they are few, simple, and verifiable:

- Characteristics of the residence:
  - Presence of electricity
  - Number of rooms for sleeping
  - Source of water
  - Type of toilet facility
  - Type of floor
- Asset ownership:
  - Radio
  - Television
  - Refrigerator
  - Telephone
  - Bicycle
  - Car

As usual, Vyas and Kumaranayake can only assume that the indices represent economic status. Indeed, they do not relate their index to anything, not a quantitative measure of health as in Gwatkin *et al.* (2003) nor a qualitative measure of destitution as in Devereux and Sharp (2003). They do not present ready-to-use indices.

3.5 IRIS Center

USAID commissioned IRIS Center ("IRIS", 2008) to build a poverty-assessment tool for its Ethiopian microenterprise partners to use for reporting on their participants' poverty rates. IRIS uses the same data as in this paper with the USD1.25/day and USD2.50 2005 PPP poverty lines.

After comparing several statistical approaches, IRIS settles on quantile regression (Koenker and Hallock, 2001). All their indicators are simple, inexpensive, and verifiable:[6]

- Household demographics:
  - Number of members
  - Sex of the head
  - Age of the head
  - Marital status of the head
  - Number of household members ages 16 or older who can read and write
- Characteristics of the residence
  - Number of rooms
  - Main construction material of roof
  - Main source of lighting
  - Main cooking fuel
  - Main source of drinking water in the rainy season
- Ownership of agricultural assets:
  - Number of cattle
  - Number of axes/*gejera*
- Ownership of consumer durables:
  - Number of blankets/*gabi*
  - Radio
  - Television
  - Video deck

[6] IRIS does not report the actual scorecard, only the questionnaire used to collect data, so their actual indicators may differ slightly from those listed here.

IRIS' preferred measure of accuracy is the "Balanced Poverty Accuracy Criterion" (BPAC), and USAID has adopted BPAC as the criterion for certifying poverty-assessment tools (IRIS Center, 2005). BPAC depends on the difference between the estimated poverty rate and its true value and on *inclusion*, that is, the correct classification of households as "below poverty line" when their per capita expenditure is truly below the line. A higher BPAC is preferred. The formula is:

(Inclusion – |Undercoverage – Leakage|) x [100 ÷ (Inclusion + Undercoverage)].

In personal communication, Anthony Leegwater reports that BPAC for the IRIS scorecard with the $1.25/day 2005 PPP line (poverty rate of 20.0 percent at the household level) is 47.6 when tested "in-sample", that is, with the same data used to construct the scorecard. When tested "out-of-sample" (that is, with data not used to construct the scorecard), Leegwater reports a 95-percent confidence interval from 38.8 to 48.5. For comparison, the highest out-of-sample BPAC for the scorecard here with the $1.25/day 2005 PPP line (corresponding to a household-level poverty rate of 27.2 percent) is 32.4 (if scores are not grouped into 20 ranges, the highest is 50.3).

The BPACs are not strictly comparable because the USD1.25/day 2005 PPP poverty lines are different. One possible reason for this is that IRIS does not document how they convert monetary units to 2005. Also, IRIS uses the PPP factor of USD1.00 = Birr 2.30 for "Actual Individual Consumption", while the paper here uses USD1.00 = Birr 2.75 for "Individual Consumption Expenditure by Households" (International Comparison Project, 2008). In personal communication, Shaohua Chen reports that the

World Bank uses "Individual Consumption Expenditure by Households" for poverty measurement based on national household expenditure surveys (for example, Chen, Ravaillon, and Sangraula, 2008).

Overall, the main difference between the scorecard here and that of IRIS is transparency. For example, IRIS does not report the Birr value of its poverty line nor standard errors of any kind.

3.6 Dekker

Dekker (2006) uses the 1,400 households in the 1994/5 Ethiopian Rural Household Survey to construct PCA-based asset indices for 19 villages across Ethiopia to check whether scorecards vary by locality. This question is similar to that in Elbers, Lanjouw, and Leite (2008), Tarozzi (2008), Tarozzi and Deaton (2007), and Demombynes, Elbers, and Lanjouw (2007). Like these authors, Dekker concludes that an all-Ethiopia scorecard differs from village-specific scorecards. With about 70 households per village, this could result from sampling variation, but this is not tested.

The indicators in the various scorecards are simple and inexpensive to verify:

- Characteristics of the residence:
  - Rooms per person
  - Wall material
  - Roof material
  - Floor durability
  - Presence of electricity
  - Source of water
  - Type of toilet arrangement
  - Type of cooking fuel

- Ownership of consumer durables:
  - — Watch
  - — Flashlight
  - — Radio
  - — Television
  - — Table or chair
  - — Sofa
  - — Modern bed
  - — Refrigerator
  - — Mill
  - — Cupboard
  - — Pouch
  - — Weaving equipment
  - — Leather mat
  - — Cart
  - — Bicycle
  - — Motorbike/scooter
  - — Landline telephone
  - — Mobile telephone
- Ownership of oxen

Dekker conducts two tests of how well the all-Ethiopia index ranks households in terms of the weeks in a typical year in which a household eats substantially less:

- Mean weeks of low food consumption by PCA-index quintile
- Size of coefficient on the PCA index in a Poisson regression on weeks of low food consumption

Compared to a measure of current consumption, the PCA index does a better job of ranking households by food security.

Dekker differs from the scorecard here in that it is based on older data, it does not attempt to estimate expenditure-based poverty status, it does not report formal measures of accuracy nor out-of-sample tests, and it does not discuss targeting, estimating poverty rates, nor estimating changes in poverty rates.

4. Scorecard construction

About 80 potential indicators are initially prepared in the areas of:

- Family composition (such as household size)
- Education (such as school attendance of children)
- Housing (such as the main source of cooking fuel)
- Ownership of durable goods (such as radios and clocks)
- Ownership of agricultural assets (such as cattle, sheep, and goats)

Each indicator is first screened with the entropy-based "uncertainty coefficient" (Goodman and Kruskal, 1979) that measures how well the indicator predicts poverty on its own. Figure 4 lists the best indicators, ranked by uncertainty coefficient. Responses for each indicator are ordered starting with those most strongly associated with poverty.

The scorecard also aims to measure *changes* in poverty through time. This means that, when selecting indicators and holding other considerations constant, preference is given to more sensitive indicators. For example, ownership of a radio is probably more likely to change in response to changes in poverty than is the highest grade completed by the female head/spouse.

The scorecard itself is built using the USD1.25/day 2005 PPP line and Logit regression on the construction sub-sample (Figure 2). Indicator selection uses both judgment and statistics (forward stepwise, based on "c"). The first step is to use Logit to build one scorecard for each candidate indicator. Each scorecard's accuracy is taken as "c", a measure of ability to rank by poverty status (SAS Institute Inc., 2004).

One of these one-indicator scorecards is then selected based on several factors (Schreiner *et al.*, 2004; Zeller, 2004), including improvement in accuracy, likelihood of acceptance by users (determined by simplicity, cost of collection, and "face validity" in terms of experience, theory, and common sense), sensitivity to changes in poverty status, variety among indicators, and verifiability.

A series of two-indicator scorecards are then built, each based on the one-indicator scorecard selected from the first step, with a second candidate indicator added. The best two-indicator scorecard is then selected, again based on "c" and judgment. These steps are repeated until the scorecard has 11 indicators.

The final step is to transform the Logit coefficients into non-negative integers such that total scores range from 0 (most likely below a poverty line) to 100 (least likely below a poverty line).

This algorithm is the Logit analogue to the familiar R^2-based stepwise with least-squares regression. It differs from naïve stepwise in that the criteria for selecting indicators include not only statistical accuracy but also judgment and non-statistical factors. The use of non-statistical criteria can improve robustness through time and, more important, helps ensure that indicators are simple and make sense to users.

The single scorecard here applies to all of Ethiopia. Evidence from India and Mexico (Schreiner, 2006a and 2005a), Sri Lanka (Narayan and Yoshida, 2005), and Jamaica (Grosh and Baker, 1995) suggests that segmenting scorecards by urban/rural does not improve accuracy much.

5. Practical guidelines for scorecard use

The main challenge of scorecard design is not to maximize accuracy but rather to improve the chances that scoring is actually used (Schreiner, 2005b). When scoring projects fail, the reason is not usually technical inaccuracy but rather the failure of an organization to decide to do what is needed to integrate scoring in its processes and to learn to use it properly (Schreiner, 2002). After all, most reasonable scorecards predict tolerably well, thanks to the empirical phenomenon known as the "flat max" (Hand, 2006; Baesens *et al.*, 2003; Lovie and Lovie, 1986; Kolesar and Showers, 1985; Stillwell, Barron, and Edwards, 1983; Dawes, 1979; Wainer, 1976; Myers and Forgy, 1963). The bottleneck is less technical and more human, not statistics but organizational change management. Accuracy is easier to achieve than adoption.

The scorecard here is designed to encourage understanding and trust so that users will adopt it and use it properly. Of course, accuracy matters, but it is balanced against simplicity, ease-of-use, and "face validity". Programs are more likely to collect data, compute scores, and pay attention to the results if, in their view, scoring does not make a lot of "extra" work and if the whole process generally seems to make sense.

To this end, the scorecard here fits on one page (Figure 1). The construction process, indicators, and points are simple and transparent. "Extra" work is minimized; non-specialists can compute scores by hand in the field because the scorecard has:

- Only 11 indicators
- Only categorical indicators
- Simple weights (non-negative integers, no arithmetic beyond addition)

A field worker using the paper scorecard would:

- Record participant identifiers
- Read each question from the scorecard
- Circle the response and its points
- Write the points in the far-right column
- Add up the points to get the total score
- Implement targeting policy (if any)
- Deliver the paper scorecard to a central office for filing or data entry

Of course, field workers must be trained. Quality outputs depend on quality inputs. If organizations or field workers gather their own data and have an incentive to exaggerate poverty rates (for example, if funders reward them for higher poverty rates), then it is wise to do on-going quality control via data review and random audits (Matul and Kline, 2003).[7] IRIS Center (2007a) and Toohig (2007) are useful nuts-and-bolts guides for budgeting, training field workers and supervisors, logistics, sampling, interviewing, piloting, recording data, and controlling quality.

In particular, while collecting scorecard indicators is relatively easier than alternatives, it is still absolutely difficult. Training and explicit definitions of the terms in the scorecard is essential. For the example of Nigeria, Onwujekwe, Hanson, and Fox-Rushby (2006) find distressingly low inter-rater and test-retest correlations for indicators as seemingly simple and obvious as whether the household owns an automobile. In Mexico, however, Martinelli and Parker (2007) find that errors by

[7] If an organization does not want field workers to know the points associated with indicators, then they can use the version of Figure 1 without points and apply the points later in a spreadsheet or database at the central office.

interviewers and lies by respondents had negligible effects on targeting accuracy. For now, it is unknown whether these results are universal or country-specific.

In terms of sampling design, an organization must make choices about:

- Who will do the scoring
- How scores will be recorded
- What participants will be scored
- How many participants will be scored
- How frequently participants will be scored
- Whether scoring will be applied at more than one point in time
- Whether the same participants will be scored at more than one point in time

The non-specialists who apply the scorecard with participants in the field can be:

- Employees of the organization
- Third-party contractors

Responses, scores, and poverty likelihoods can be recorded:

- On paper in the field and then filed at an office
- On paper in the field and then keyed into a database or spreadsheet at an office
- On portable electronic devices in the field and downloaded to a database

The subjects to be scored can be:

- All participants (or all new participants)
- A representative sample of all participants (or of all new participants)
- All participants (or all new participants) in a representative sample of branches
- A representative sample of all participants (or of all new participants) in a representative sample of branches

If not determined by other factors, the number of participants to be scored can be derived from sample-size formulas (presented later) for a desired confidence level and a desired confidence interval.

Frequency of application can be:

- At in-take of new clients only (precluding measuring change in poverty rates)
- As a once-off project for current participants (precluding measuring change)
- Once a year (or at some other fixed time interval, allowing measuring change)
- Each time a field worker visits a participant at home (allowing measuring change)

When the scorecard is applied more than once in order to measure change in poverty rates, it can be applied:

- With a different set of participants from a given population
- With the same set of participants

An example set of choices were made by BRAC and ASA, two microlenders in Bangladesh (each with 7 million participants) who are applying a the Simple Poverty Scorecard tool for Bangaldesh (Schreiner, 2006b). Their design is that loan officers in a random sample of branches will score all participants each time they visit a homestead (about once a year) as part of their standard due diligence prior to loan disbursement. Responses are recorded on paper in the field before being sent to a central office to be entered into a database. ASA's and BRAC's sampling plans cover 50,000–100,000 participants each.

6. Estimates of household poverty likelihoods

The sum of scorecard points for a household is called the *score*. For Ethiopia, scores range from 0 (most likely below a poverty line) to 100 (least likely below a poverty line). While higher scores indicate less likelihood of being below a poverty line, the scores themselves have only relative units. For example, doubling the score does not double the likelihood of being above a poverty line.

To get absolute units, scores must be converted to *poverty likelihoods*, that is, probabilities of being below a poverty line. This is done via simple look-up tables. For the example of the USD1.25 2005 PPP line, scores of 10–14 have a poverty likelihood of 63.6 percent, and scores of 40–44 have a poverty likelihood of 18.4 percent (Figure 5).

The poverty likelihood associated with a score varies by poverty line. For example, scores of 40–44 are associated with a poverty likelihood of 18.4 percent for the USD1.25 2005 PPP line but 8.3 percent for the USD1.00 2005 PPP line.[8]

6.1 Calibrating scores with poverty likelihoods

A given score is non-parametrically associated ("calibrated") with a poverty likelihood by defining the poverty likelihood as the share of households in the calibration sub-sample who have the score and who are below a given poverty line.

[8] Starting with Figure 5, most figures have four versions, one for each poverty line. To keep them straight, they are grouped by poverty line. Single tables that pertain to all poverty lines are placed with the tables for the USD1.00/day 2005 PPP line.

For the example of the USD1.25 2005 PPP line (Figure 6), there are 10,580 (normalized) households in the calibration sub-sample with a score of 20–24, of whom 5,045 (normalized) are below the poverty line. The estimated poverty likelihood associated with a score of 20–24 is then 47.7 percent, because 5,045 ÷ 10,580 = 47.7 percent.

To illustrate with the USD1.25/day 2005 PPP line and a score of 40–44, there are 9,216 (normalized) households in the calibration sample, of whom 1,691 (normalized) are below the line (Figure 6). Thus, the poverty likelihood for this score is 1,691 ÷ 9,216 = 18.4 percent.

The same method is used to calibrate scores with estimated poverty likelihoods for the other poverty lines.

Figure 7 shows, for all scores, the likelihood that expenditure falls in a range demarcated by two adjacent poverty lines. For example, the daily expenditure of someone with a score of 35–39 falls in the following ranges with probability:

- 8.0 percent below the USD1.00/day 2005 PPP line
- 10.5 percent between the USD1.00/day and USD1.25/day 2005 PPP lines
- 39.1 percent between the USD1.25/day and USD1.75/day 2005 PPP lines
- 27.7 percent between the USD1.75/day and USD2.50/day 2005 PPP lines
- 14.7 percent above the USD2.50/day 2005 PPP line

Even though the scorecard is constructed partly based on judgment, the calibration process produces poverty likelihoods that are objective, that is, derived from survey data on expenditure and quantitative poverty lines. The poverty likelihoods would be objective even if indicators and/or points were selected without any data at

all. In fact, objective scorecards of proven accuracy are often based only on judgment (Fuller, 2006; Caire, 2004; Schreiner *et al.*, 2004). Of course, the scorecard here is constructed with both data and judgment. The fact that this paper acknowledges that some choices in scorecard construction—as in any statistical analysis—are informed by judgment in no way impugns the objectivity of the poverty likelihoods, as this depends on using data in score calibration, not on using data (and nothing else) in scorecard construction.

Although the points in Ethiopia's scorecard are transformed coefficients from a Logit regression, scores are not converted to poverty likelihoods via the Logit formula of 2.718281828^{score} x $(1+ 2.718281828^{score})^{-1}$. This is because the Logit formula is esoteric and difficult to compute by hand. Non-specialists find it more intuitive to define the poverty likelihood as the share of households with a given score in the calibration sample who are below a poverty line. In the field, converting scores to poverty likelihoods requires no arithmetic at all, just a look-up table. This non-parametric calibration can also improve accuracy, especially with large calibration samples.

6.2 Accuracy of estimates of poverty likelihoods

As long as the relationship between indicators and poverty does not change and the scorecard is applied to households from the same population from which it was constructed, this calibration process produces unbiased estimates of poverty likelihoods. *Unbiased* means that in repeated samples from the same population, the average

estimate matches the true poverty likelihood. The scorecard also produces unbiased estimates of poverty rates at a point in time and of changes in poverty rates between two points in time.[9]

Of course, the relationship between indicators and poverty changes with time and across sub-groups within Ethiopia's population, so the scorecard applied after March 2005 (as it must be in practice) and/or to non-nationally representative groups will generally be biased.

How accurate are estimates of poverty likelihoods? To measure, the scorecard is applied to 1,000 bootstrap samples of size $n = 16{,}384$ from the validation sub-sample (Figure 2). Bootstrapping entails (Efron and Tibshirani, 1993):

- Score each household in the validation sample
- Draw a new bootstrap sample *with replacement* from the validation sample
- For each score, compute the true poverty likelihood in the bootstrap sample, that is, the share of households with the score and expenditure below a poverty line
- For each score, record the difference between the estimated poverty likelihood (Figure 5) and the true poverty likelihood in the bootstrap sample
- Repeat the previous three steps 1,000 times
- For each score, report the average difference between estimated and true poverty likelihoods across the 1,000 bootstrap samples
- For each score, report the two-sided interval containing the central 900, 950, or 990 differences between estimated and true poverty likelihoods

For each score range, Figure 8 shows the average difference between estimated and true poverty likelihoods as well as confidence intervals for the differences.

[9] This follows because these estimates of groups' poverty rates are linear functions of the unbiased estimates of individual households' poverty likelihoods.

For the USD1.25/day 2005 PPP line, the average poverty likelihood across bootstrap samples for scores of 20–24 in the validation sample is too high by 0.7 percentage points (Figure 8). For scores of 30–34, the estimate is too low by 4.4 percentage points.[10]

For the validation sample, the 90-percent confidence interval for the differences for scores of 20–24 is +/–2.2 percentage points (Figure 8). This means that in 900 of 1,000 bootstraps, the difference between the estimate and the true value is between –1.5 and +2.9 percentage points (because 0.7 – 2.2 = –1.5, and 0.7 + 2.2 = +2.9). In 950 of 1,000 bootstraps (95 percent), the difference is 0.7 +/–2.5 percentage points, and in 990 of 1,000 bootstraps (99 percent), the difference is 0.7 +/–3.2 percentage points.

For almost all score ranges, Figure 8 shows differences—sometimes large ones—between estimated poverty likelihoods and true values. This is because the validation sub-sample is a single sample that—thanks to sampling variation—differs in distribution from the construction/calibration sub-samples and from Ethiopia's population. For targeting, however, what matters is less the difference in all score ranges and more the difference in score ranges just above and below the targeting cut-off. This mitigates the effects of bias and sampling variation on targeting (Friedman, 1997). Section 9 below looks at targeting accuracy in detail.

[10] There are differences, in spite of the estimator's unbiasedness, because the scorecard comes from a single sample. The average difference by score would be zero if samples were repeatedly drawn from the same population and split into sub-samples before repeating the entire scorecard-building process.

Of course, if estimates of groups' poverty rates are to be usefully accurate, then errors for individual households must largely cancel out. As discussed later, this is generally the case.

By construction, the scorecard here is unbiased. It may still, however, be *overfit* when applied after March 2005. That is, it may fit the 2004/5 HICE and 2004 WMS data so closely that it captures not only some timeless patterns but also some random patterns that, due to sampling variation, show up only in the 2004/5 HICE and 2004 WMS. Or the scorecard may be overfit in the sense that it becomes biased as the relationships between indicators and poverty change or when it is applied to non-nationally representative samples.

Overfitting can be mitigated by simplifying the scorecard and by not relying only on data but rather also considering experience, judgment, and theory. Of course, the scorecard here does this. Bootstrapping can also mitigate overfitting by reducing (but not eliminating) dependence on a single sampling instance. Combining scorecards can also help, at the cost of greater complexity.

Most errors in individual households' likelihoods, however, cancel out in the estimates of groups' poverty rates (see later sections). Furthermore, much of the differences may come from non-scorecard sources such as changes in the relationship between indicators and poverty, sampling variation, changes in poverty lines, inconsistencies in data quality across time, and inconsistencies/imperfections in cost-of-living adjustments. These factors can be addressed only by improving data quantity

and quality (which is beyond the scope of the scorecard) or by reducing overfitting (which likely has limited returns, given the scorecard's parsimony).

7. Estimates of a group's poverty rate at a point in time

A group's estimated poverty rate at a point in time is the average of the estimated poverty likelihoods of the individual households in the group.

To illustrate, suppose a program samples three households on Jan. 1, 2009 and that they have scores of 20, 30, and 40, corresponding to poverty likelihoods of 47.7, 28.4, and 18.4 percent (USD1.25/day 2005 PPP line, Figure 5). The group's estimated poverty rate is the households' average poverty likelihood of $(47.7 + 28.4 + 18.4) \div 3 =$ 31.5 percent.[11]

7.1 Accuracy of estimated poverty rates at a point in time

How accurate is this estimate? For a range of sample sizes, Figure 10 reports average differences between estimated and true poverty rates as well as precision (confidence intervals for the differences) for the scorecard applied to 1,000 bootstrap samples from the validation sample. For the USD1.25/day 2005 PPP line, the scorecard is generally too high by about 1.4 percentage points; it estimates a poverty rate of 28.6 percent for the validation sample, but the true value is 27.2 percent (Figure 2). For all poverty lines, absolute differences for the validation sample are 2.1 percentage points or

[11] The group's poverty rate is *not* the poverty likelihood associated with the average score. Here, the average score is $(20 + 30 + 40) \div 3 = 30$, and the poverty likelihood associated with the average score is 28.4 percent. This is not the 31.5 percent found as the average of the three poverty likelihoods associated with each of the three scores.

less, with an average of about 1.3 percentage points (Figure 9, summarizing Figure 10 across poverty lines).

As before, these differences are due to sampling variation in the validation sample and in the random division of the 2004/5 HICE and 2004 WMS into three sub-samples.

In terms of precision, the 90-percent confidence interval for a group's estimated poverty rate at a point in time and n = 16,384 is 0.6 percentage points or less (Figure 9). This means that in 900 of 1,000 bootstraps of this size, the difference between the estimate and the true value is within 0.6 percentage points of the average difference. In the specific case of the USD1.25/day 2005 PPP line and the validation sample, 90 percent of all samples of n = 16,384 produce estimates that differ from the true value in the range of 1.4 – 0.6 = 0.8 to 1.4 + 0.6 = 2.0 percentage points, as 1.4 is the average difference and +/–0.6 is its 90-percent confidence interval.

7.2 Formula for standard errors for estimates of poverty rates

How precise are the point-in-time estimates? Because they are averages of binary (0/1, or poor/non-poor) variables, the estimates have a Normal distribution and can be characterized by their average difference vis-à-vis true values together with the standard error of the average difference.

To derive a formula for the standard errors of estimated poverty rates at a point in time from indirect measurement via scorecards (Schreiner, 2008a), note that the

textbook formula (Cochran, 1977) that relates confidence intervals with standard errors in the case of direct measurement of poverty status is $c = +/- z \cdot \sigma$, where:

c is the confidence interval as a proportion
(for example, 0.2 for +/–2 percentage points),

z is from the Normal distribution and is $\begin{cases} 1.64 \text{ for confidence levels of 90 percent} \\ 1.96 \text{ for confidence levels of 95 percent} \\ 2.58 \text{ for confidence levels of 99 percent} \end{cases}$,

σ is the standard error of the estimated poverty rate, that is, $\sqrt{\frac{p \cdot (1-p)}{n}}$,

p is the proportion of households below the poverty line in the sample, and

n is the sample size.

For example, this implies that for a sample n of 16,384 with 90-percent confidence ($z = 1.64$) and a poverty rate p of 27.2 percent (the true poverty rate in the validation sample for $1.25/day 2005 PPP in Figure 2), the confidence interval c is $+/- z \cdot \sqrt{\frac{p \cdot (1-p)}{n}} = +/- 1.64 \cdot \sqrt{\frac{0.272 \cdot (1-0.272)}{16{,}384}} = 0.0057$, or 0.57 percentage points.

Scorecards, however, do not measure poverty directly, so this formula is not immediately applicable. To derive a formula for the Ethiopia scorecard, consider Figure 10, which reports empirical confidence intervals c for the differences for the scorecard applied to 1,000 bootstrap samples of various sample sizes from the validation sample.

For $n = 16{,}384$ and the \$1.25/day 2005 PPP line, the 90-percent confidence interval is 0.55 percentage points.[12]

Thus, the 90-percent confidence interval with $n = 16{,}384$ is 0.55 percentage points for Ethiopia's scorecard and 0.57 percentage points for direct measurement. The ratio of the two intervals is $0.55 \div 0.57 = 0.96$.

Now consider the same case, but with $n = 8{,}192$. The confidence interval under direct measurement is $+/-1.64 \cdot \sqrt{\frac{0.272 \cdot (1-0.272)}{8{,}192}} = 0.0081$, or about 0.81 percentage points. The empirical confidence interval with the Ethiopia scorecard (Figure 10) is 0.00815, or about 0.82 percentage points. Thus for $n = 8{,}192$, the ratio of the two intervals is $0.82 \div 0.81 = 1.01$.

This ratio of 1.01 for $n = 8{,}182$ is not far from the ratio of 0.96 for $n = 16{,}384$. Across all sample sizes of 256 or more in Figure 10, the average ratio turns out to be 1.00, implying that confidence intervals for indirect estimates of poverty rates via the Ethiopia scorecard and this poverty line are about the same as confidence intervals for direct estimates via the 2004/5 HICE. This 1.00 appears in Figure 9 as the "α factor" because if $\alpha = 1.00$, then the formula relating confidence intervals c and standard errors σ for the Ethiopia scorecard is $c = +/- z \cdot \alpha \cdot \sigma$. That is, the formula for the standard error σ for point-in-time estimates of poverty rates via scoring is $\alpha \cdot \sqrt{\frac{p \cdot (1-p)}{n}}$.

[12] Due to rounding, Figure 10 displays 0.6, not 0.55.

In general, α could be more or less than 1.00. When α is less than 1.00, it means that the scorecard is more precise than direct measurement. This occurs for two of four poverty lines in Figure 9.

The formula relating confidence intervals to standard errors for the scorecard can be rearranged to give a formula for determining sample size before measurement.[13] If $\hat{p}$ is the expected poverty rate before measurement, then the formula for sample size n based on the desired confidence level that corresponds to z and the desired confidence interval $+/-c$ is $n = \left(\frac{\alpha \cdot z}{c}\right)^2 \cdot \hat{p} \cdot (1 - \hat{p})$.

To illustrate how to use this, suppose $c = 0.04645$ and $z = 1.64$ (90-percent confidence). Then the formula gives $n = \left(\frac{1.00 \cdot 1.64}{0.04645}\right)^2 \cdot 0.272 \cdot (1 - 0.272) = 247$, close to the sample size of 256 observed for these parameters in Figure 10.

Of course, the α factors in Figure 9 are specific to Ethiopia, its poverty lines, its poverty rates, and this scorecard. The derivation of the formulas, however, is valid for any scorecard following the approach in this paper.

In practice after March 2005, an organization would select a poverty line (say, the USD1.25/day 2005 PPP line), select a desired confidence level (say, 90 percent, or z

[13] IRIS Center (2007a and 2007b) says that a sample size of $n = 300$ is sufficient for USAID reporting. If a scorecard is as precise as direct measurement, if the expected (before measurement) poverty rate is 50 percent, and if the confidence level is 90 percent, then n = 300 implies a confidence interval of +/–2.2 percentage points. In fact, USAID has not specified confidence levels or intervals. Furthermore, the expected poverty rate may not be 50 percent, and the scorecard could be more or less precise than direct measurement.

= 1.64), select a desired confidence interval (say, +/–2.0 percentage points, or c = 0.02), make an assumption about $\hat{p}$ (perhaps based on a previous measurement such as the 27.2 percent national average for the 2004/5 HICE in Figure 2), look up α (here, 1.00), assume that the scorecard will still work in the future and/or for non-nationally representative sub-groups,[14] and then compute the required sample size. In this illustration, $n = \left(\frac{1.00 \cdot 1.64}{0.02}\right)^2 \cdot 0.272 \cdot (1 - 0.272) = 1{,}332$.

[14] This paper reports accuracy for the scorecard applied to the validation sample, but it cannot test accuracy for later years or for other groups. Still, performance after March 2005 will probably resemble that in the 2004/5 HICE and 2004 WMS, with some deterioration as time passes.

8. Estimates of changes in group poverty rates over time

The change in a group's poverty rate between two points in time is estimated as the change in the average poverty likelihood of the households in the group. With data for 2004/5 HICE and 2004 WMS only, this paper cannot estimate changes over time, nor can it present sample-size formula. Nevertheless, the relevant concepts are presented here because, in practice, pro-poor organizations can apply the scorecard to collect their own data and measure change through time.

8.1 Warning: Change is not impact

Scoring can estimate change. Of course, poverty could get better or worse, and scoring does not indicate what caused change. This point is often forgotten or confused, so it bears repeating: the scorecard simply estimates change, and it does not, in and of itself, indicate the reason for the change. In particular, estimating the impact of program participation requires knowing what would have happened to participants if they had not been participants (Moffitt, 1991). Knowing this requires either strong assumptions or a control group that resembles participants in all ways except participation. To belabor the point, the scorecard can help estimate program impact only if there is some way to know what would have happened in the absence of the program. And that information must come from somewhere beyond the scorecard. Even

measuring simple change usually requires assuming that the population is constant over time and that program drop-outs do not differ from non-drop-outs.

8.2 Calculating estimated changes in poverty rates over time

Consider the illustration begun in the previous section. On Jan. 1, 2009, a program samples three households who score 20, 30, and 40 and so have poverty likelihoods of 47.7, 28.4, and 18.4 percent (USD1.25/day 2005 PPP line, Figure 5). The group's baseline estimated poverty rate is the households' average poverty likelihood of $(47.7 + 28.4 + 18.4) \div 3 = 31.5$ percent.

After baseline, two sampling approaches are possible for the follow-up round:

- Score a new, independent sample, measuring change by cohort across samples
- Score the same sample at follow-up as at baseline

By way of illustration, suppose that a year later on Jan. 1, 2010, the program samples three additional households who are in the same cohort as the three households originally sampled (or suppose that the program scores the same three original households a second time) and finds that their scores are 25, 35, and 45 (poverty likelihoods of 38.5, 18.5, and 18.6 percent, USD1.25/day 2005 PPP, Figure 5). Their average poverty likelihood at follow-up is now $(38.5 + 18.5 + 18.6) \div 3 = 25.2$ percent, an improvement of $31.5 - 25.2 = 6.3$ percentage points.

This suggests that about one of sixteen participants crossed the poverty line in 2009.[15] Among those who started below the line, one in five ($6.3 \div 31.5 = 20.0$ percent) on net ended up above the line.[16]

8.3 Accuracy for estimated change in two independent samples

With only the 2004/5 HICE and 2004 WMS, it is not possible to measure the accuracy of scorecard estimates of changes in groups' poverty rates over time. In practice, of course, local pro-poor organizations can still apply Ethiopia's scorecard to estimate change. The rest of this section suggests approximate formulas for standard errors and sample sizes that may be used until there is additional data.

For two equal-sized independent samples, the same logic as above can be used to derive a formula relating the confidence interval c with the standard error σ of a scorecard's estimate of the change in poverty rates over time:

$$c = +/- z \cdot \sigma = +/- z \cdot \alpha \cdot \sqrt{\frac{2 \cdot p \cdot (1-p)}{n}} .$$

z, c, and p are defined as above, n is the sample size at both baseline and follow-up,[17] and α is the average (across a range of bootstrapped sample sizes) of the ratio of

[15] This is a net figure; some people start above the line and end below it, and vice versa.
[16] The scorecard does not reveal the reasons for this change.
[17] This means that, for a given precision and with direct measurement, estimating the change in a poverty rate between two points in time requires four times as many measurements (not twice as many) as does estimating a poverty rate at a point in time.

the observed confidence interval from a scorecard and the theoretical confidence interval under direct measurement.

As before, the formula for standard errors can be rearranged to give a formula for sample sizes before indirect measurement via a scorecard, where $\hat{p}$ is based on previous measurements and is assumed equal at both baseline and follow-up:

$$n = 2 \cdot \left(\frac{\alpha \cdot z}{c} \right)^2 \cdot \hat{p} \cdot (1 - \hat{p}) .$$

For the only countries for which this α has been measured (Peru, the Philippines, and India, see Schreiner, 2009a, 2009b, and 2008b), the average α across poverty lines is 0.77, 0.77, and 1.40, so 1.00 may be a reasonably conservative figure for Ethiopia.

To illustrate the use of the formula above to determine sample size for estimating changes in poverty rates across two independent samples, suppose the desired confidence level is 90 percent ($z = 1.64$), the desired confidence interval is 2 percentage points ($c = 0.02$), the poverty line is the USD1.25/day 2005 PPP, $\alpha = 1.00$, and $\hat{p} = 0.272$ (from Figure 2). Then the baseline sample size is $n = 2 \cdot \left(\frac{1.00 \cdot 1.64}{0.02} \right)^2 \cdot 0.272 \cdot (1 - 0.272) = 2{,}663$, and the follow-up sample size is also 2,663.

8.4 Accuracy for estimated change for one sample, scored twice

Analogous to previous derivations, the general formula relating the confidence interval c to the standard error σ when using a scorecard to estimate change for a single group of households, all of whom are scored at two points in time, is:[18]

$$c = +/- z \cdot \sigma = +/- z \cdot \alpha \cdot \sqrt{\frac{p_{12} \cdot (1 - p_{12}) + p_{21} \cdot (1 - p_{21}) + 2 \cdot p_{12} \cdot p_{21}}{n}},$$

where z, c, and α are defined as usual, p_{12} is the share of all sampled households that move from below the poverty line to above it, and p_{21} is the share of all sampled households that move from above the line to below it.

The formula for standard errors can be rearranged to give a formula for sample size before measurement. This requires an estimate (based on information available before measurement) of the expected shares of all households who cross the poverty line $\hat{p}_{12}$ and $\hat{p}_{21}$. Before measurement, it is reasonable to assume that the change in the poverty rate will be zero, which implies $\hat{p}_{12} = \hat{p}_{21} = \hat{p}_*$, giving:

$$n = 2 \cdot \left(\frac{\alpha \cdot z}{c} \right)^2 \cdot \hat{p}_* .$$

[18] See McNemar (1947) and Johnson (2007). John Pezzullo helped find this formula.

Because $\hat{p}_*$ could be anything between 0–1, more information is needed before applying this formula. Suppose that the observed relationship between $\hat{p}_*$, the number of years y between baseline and follow-up, and $p_{\text{baseline}} \cdot (1 - p_{\text{baseline}})$ is—as in Peru—close to:

$$\hat{p}_* = -0.02 + 0.016 \cdot y + 0.47 \cdot [p_{\text{baseline}} \cdot (1 - p_{\text{baseline}})] .$$

Given this, a sample-size formula for a group of households to whom the Ethiopia scorecard is applied twice (once after March 2005 and then again later) is:

$$n = 2 \cdot \left(\frac{\alpha \cdot z}{c} \right)^2 \cdot \{ -0.02 + 0.016 \cdot y + 0.47 \cdot [p_{\text{baseline}} \cdot (1 - p_{\text{baseline}})] \}.$$

In Peru (the only other country for which there is an estimate, Schreiner 2009a), the average α across years and poverty lines is about 1.3.

To illustrate the use of this formula, suppose the desired confidence level is 90 percent ($z = 1.64$), the desired confidence interval is 2.0 percentage points ($c = 0.02$), the poverty line is the USD1.25/day 2005 PPP line, and the sample will first be scored in 2009 and then again in 2012 ($y = 3$). The before-baseline poverty rate is 27.2 percent ($p_{2005} = 0.272$, Figure 2), and suppose $\alpha = 1.3$. Then the baseline sample size is $n = 2 \cdot \left(\frac{1.3 \cdot 1.64}{0.02} \right)^2 \cdot \{ -0.02 + 0.016 \cdot 3 + 0.47 \cdot [0.272 \cdot (1 - 0.272)] \} = 2{,}751$. The same group of 2,751 households is scored at follow-up as well.

9. Targeting

When a program uses the scorecard for targeting, households with scores at or below a cut-off are labeled *targeted* and treated—for program purposes—as if they are below a given poverty line. Households with scores above a cut-off are labeled *non-targeted* and treated—for program purposes—as if they are above a given poverty line.

There is a distinction between *targeting status* (scoring at or below a targeting cut-off) and *poverty status* (expenditure below a poverty line). Poverty status is a fact that depends on whether expenditure is below a poverty line as directly measured by a survey. In contrast, targeting status is a program's policy choice that depends on a cut-off and on an indirect estimate from a scorecard.

Targeting is successful when households truly below a poverty line are targeted (*inclusion*) and when households truly above a poverty line are not targeted (*exclusion*). Of course, no scorecard is perfect, and targeting is unsuccessful when households truly below a poverty line are not targeted (*undercoverage*) or when households truly above a poverty line are targeted (*leakage*). Figure 11 depicts these four possible targeting outcomes. Targeting accuracy varies by cut-off; a higher cut-off has better inclusion (but greater leakage), while a lower cut-off has better exclusion (but higher undercoverage).

A program should weigh these trade-offs when setting a cut-off. A formal way to do this is to assign net benefits—based on a program's values and mission—to each of

the four possible targeting outcomes and then to choose the cut-off that maximizes total net benefits (Adams and Hand, 2000; Hoadley and Oliver, 1998).

Figure 12 shows the distribution of households by targeting outcome for the Ethiopia scorecard applied to the validation sample. For an example cut-off of 15–19, outcomes for the USD1.25/day 2005 PPP line applied to the validation sample are:

- Inclusion: 7.4 percent are below the line and correctly targeted
- Undercoverage: 19.9 percent are below the line and mistakenly not targeted
- Leakage: 5.4 percent are above the line and mistakenly targeted
- Exclusion: 67.4 percent are above the line and correctly not targeted

Increasing the cut-off to 20–24 improves inclusion and undercoverage but worsens leakage and exclusion:

- Inclusion: 12.5 percent are below the line and correctly targeted
- Undercoverage: 14.7 percent are below the line and mistakenly not targeted
- Leakage: 10.8 percent are above the line and mistakenly targeted
- Exclusion: 62.0 percent are above the line and correctly not targeted

Which cut-off is preferred depends on total net benefit. If each targeting outcome has a per-household benefit or cost, then the total net benefit for a given cut-off is:

Benefit per household correctly included	x	Households correctly included	–
Cost per household mistakenly not covered	x	Households mistakenly not covered	–
Cost per household mistakenly leaked	x	Households mistakenly leaked	+
Benefit per household correctly excluded	x	Households correctly excluded.	

To set an optimal cut-off, a program would:

- Assign benefits and costs to possible outcomes, based on its values and mission
- Tally total net benefits for each cut-off using Figure 12 for a given poverty line
- Select the cut-off with the highest total net benefit

The most difficult step is assigning benefits and costs to targeting outcomes. Any program that uses targeting—with or without scoring—should thoughtfully consider

how it values successful inclusion or exclusion versus errors of undercoverage and leakage. It is healthy to go through a process of thinking explicitly and intentionally about how possible targeting outcomes are valued.

A common choice of benefits and costs is "Total Accuracy" (IRIS Center, 2005; Grootaert and Braithwaite, 1998). With this, total net benefit is the number of households correctly included or correctly excluded:

Total Accuracy =	1	x	Households correctly included	–
	0	x	Households mistakenly undercovered	–
	0	x	Households mistakenly leaked	+
	1	x	Households correctly excluded.	

Figure 12 shows "Total Accuracy" for all cut-offs for the Ethiopia scorecard. For the USD1.25/day 2005 PPP line in the validation sample, total net benefit is greatest (74.8) for a cut-off of 15–19, with about three in four Ethiopian households correctly classified.

"Total Accuracy" weighs successful inclusion of households below the line the same as successful exclusion of households above the line. If a program valued inclusion more (say, twice as much) than exclusion, it could reflect this by setting the benefit for inclusion to 2 and the benefit for exclusion to 1. Then the chosen cut-off would maximize (2 x Households correctly included) + (1 x Households correctly excluded).[19]

[19] Figure 12 also reports "BPAC", the Balanced Poverty Accuracy Criteria adopted by USAID as its criterion for certifying poverty-assessment tools. After normalizing by the number of people below the poverty line, the formula is:
BPAC = (Inclusion – |Undercoverage – Leakage|) x [100 ÷ (Inclusion+Undercoverage)].

As an alternative to assigning benefits and costs to targeting outcomes and then choosing a cut-off to maximize total net benefit, a program could set a cut-off to achieve a desired poverty rate among targeted households. The third column of Figure 13 ("% targeted who are poor") shows, for the Ethiopia scorecard applied to the validation sample, the expected poverty rate among households who score at or below a given cut-off. For the example of the USD1.25/day 2005 PPP line, targeting households who score 35–39 or less would target 58.2 percent of all Ethiopian households and produce a poverty rate among those targeted of 40.8 percent.

Figure 13 also reports two other measures of targeting accuracy. The first is a version of coverage ("% of poor who are targeted"). For the example of the USD1.25/day 2005 PPP line and a cut-off of 35–39, 87.2 percent of all poor households are covered.

The final targeting measure in Figure 13 is the number of successfully targeted poor households for each non-poor household mistakenly targeted (right-most column). For the USD1.25/day 2005 PPP line and a cut-off of 35–39, covering 0.7 poor households means leaking to 1 non-poor households.

10. Conclusion

Pro-poor organizations in Ethiopia can use the scorecard to estimate the likelihood that a household has expenditure below a given poverty line, to estimate the poverty rate of a group of households at a point in time, and to estimate changes in the poverty rate of a group of households between two points in time. The scorecard can also be used for targeting.

The scorecard is inexpensive to use and can be understood by non-specialists. It is designed to be practical for local pro-poor organizations who want to improve how they monitor and manage their social performance in order to speed up their participants' progress out of poverty.

The scorecard is built with a sub-sample of data from the 2004/5 HICE and 2004 WMS, tested with a different sub-sample, and calibrated to four poverty lines (USD1.00/day 2005 PPP, USD1.25/day 2005 PPP, USD1.75/day 2005 PPP, and USD2.50/day 2005 PPP).

Accuracy and precision are reported for estimates of households' poverty likelihoods, groups' poverty rates at a point in time, and changes in groups' poverty rates over time. Of course, the scorecard's estimates of changes in poverty rates are not the same as estimates of program impact. Targeting accuracy is also reported.

When the scorecard is applied to the validation sample, the absolute difference between estimates versus true poverty rates for groups of households at a point in time is always less than 2.1 percentage points and averages—across the four poverty lines—

about 1.3 percentage points. For $n = 16{,}384$ and 90-percent confidence, the precision of these differences is +/–0.6 percentage points or less, and for $n = 1{,}024$, precision is +/–2.5 percentage points or less.

For targeting, programs can use the results reported here to select a cut-off that fits their values and mission.

Although the statistical technique is innovative, and although technical accuracy is important, the design of the scorecard here focuses on transparency and ease-of-use. After all, a perfectly accurate scorecard is worthless if programs feel so daunted by its complexity or its cost that they do not even try to use it. For this reason, the scorecard is kept simple, using 11 indicators that are inexpensive to collect and that are straightforward to verify. Points are all zeros or positive integers, and scores range from 0 (most likely below a poverty line) to 100 (least likely below a poverty line). Scores are related to poverty likelihoods via simple look-up tables, and targeting cut-offs are likewise simple to apply. The design attempts to facilitate adoption by helping managers understand and trust scoring and by allowing non-specialists to generate scores quickly in the field.

In sum, the scorecard is a practical, objective way for pro-poor programs in Ethiopia to monitor poverty rates, track changes in poverty rates over time, and target services. The same approach can be applied to any country with similar data from a national expenditure survey.

References

Adams, Niall M.; and David J. Hand. (2000) "Improving the Practice of Classifier Performance Assessment", *Neural Computation*, Vol. 12, pp. 305–311.

Baesens, Bart; Van Gestel, Tony; Viaene, Stijn; Stepanova, Maria; Suykens, Johan A. K.; and Jan Vanthienen. (2003) "Benchmarking State-of-the-Art Classification Algorithms for Credit Scoring", *Journal of the Operational Research Society*, Vol. 54, pp. 627–635.

Bigsten, Arne; Kebede, Bereket; Shimeles, Abebe; and Mekonnen Taddesse. (2002) "Growth and Poverty Reduction in Ethiopia: Evidence from Household Panel Surveys", Göteborg University Working Papers in Economics No. 65, `gupea.ub.gu.se/dspace/bitstream/2077/2799/1/gunwpe0065.pdf`, accessed 8 April 2009.

Bollen, Kenneth A.; Glanville, Jennifer L.; and Guy Stecklov. (2007) "Socio-Economic Status, Permanent Income, and Fertility: A Latent-Variable Approach", *Population Studies*, Vol. 61, No. 1, pp. 15–34.

Caire, Dean. (2004) "Building Credit Scorecards for Small Business Lending in Developing Markets", `microfinance.com/English/Papers/Scoring_SMEs_Hybrid.pdf`, accessed 9 April 2009.

Ravallion, Martin; Chen; Shaohua; and Prem Sangraula. (2008) "Dollar a Day Revisited", World Bank Policy Research Working Paper No. 4620, `www-wds.worldbank.org/external/default/WDSContentServer/IW3P/IB/2008/08/26/000158349_20080826113239/Rendered/PDF/WPS4703.pdf`, accessed 13 April 2009

Coady, David; Grosh, Margaret; and John Hoddinott. (2002) "The Targeting of Transfers in Developing Countries: Review of Experience and Lessons" , `hdl.handle.net/10986/14902`, retrieved 3 November 2015.

Cochran, William G. (1977) *Sampling Techniques, Third Edition*.

Dawes, Robyn M. (1979) "The Robust Beauty of Improper Linear Models in Decision Making", *American Psychologist*, Vol. 34, No. 7, pp. 571–582.

Dekker, Marleen. (2006) "Estimating Wealth Effects without Expenditure Data: Evidence from Rural Ethiopia", *Ethiopian Journal of Economics*, Vol. 15, No. 1, pp. 35–54.

Demombynes, Gabriel; Elbers, Chris; and Peter Lanjouw. (2007) "How Good a Map? Putting Small-Area Estimation to the Test", World Bank Policy Research Working Paper No. 4155, www-wds.worldbank.org/servlet/WDSContentServer/WDSP/IB/2007/03/26/000016406_20070326150728/Rendered/PDF/wps4155.pdf, accessed 10 April 2009.

Dercon, Stefan. (2002) *The Impact of Economic Reforms on Rural Households in Ethiopia: A Study from 1989 to 1995.*

_____. (2000) "Changes in Poverty and Social Indicators in Ethiopia in the 1990s: (At Last) Some Good News from Ethiopia", www.economics.ox.ac.uk/members/stefan.dercon/CHANGES%20IN%20POVERTY%20AND%20SOCIAL%20INDICATORS.pdf, accessed 9 April 2009.

_____. (1997) "Poverty and Deprivation in Ethiopia", www.econ.ox.ac.uk/members/stefan.dercon/poverty%20profile.pdf, accessed 9 April 2009.

Devereux, Stephen; and Kay Sharp. (2003) "Is Poverty Really Falling in Rural Ethiopia?" chronicpoverty.org/pubfiles/devereux.pdf, accessed 9 April 2009.

Efron, Bradley; and Robert J. Tibshirani. (1993) *An Introduction to the Bootstrap.*

Elbers, Chris; Lanjouw, Peter; and Phillippe George Leite. (2008) "Brazil within Brazil: Testing the Poverty Map Methodology in Minas Gerais", World Bank Policy Research Working Paper No. 4513, www-wds.worldbank.org/servlet/WDSContentServer/WDSP/IB/2008/02/26/000158349_20080226134003/Rendered/PDF/wps4513.pdf, accessed 10 April 2009.

Filmer, Deon; and Lant Pritchett. (2001) "Estimating Wealth Effects without Expenditure Data—or Tears: An Application to Educational Enrollments in States of India", *Demography*, Vol. 38, No. 1, pp. 115–132.

Filmer, Deon; and Kinnon Scott. (2008) "Assessing Asset Indices", World Bank Policy Research Working Paper No. 4605, papers.ssrn.com/sol3/papers.cfm?abstract_id=1149108, accessed 9 April 2009.

Friedman, Jerome H. (1997) "On Bias, Variance, 0–1 Loss, and the Curse-of-Dimensionality", *Data Mining and Knowledge Discovery*, Vol. 1, pp. 55–77.

Fuller, Rob. (2006) "Measuring the Poverty of Microfinance Clients in Haiti", microfinance.com/English/Papers/Scoring_Poverty_Haiti_Fuller.pdf, accessed 9 April 2009.

Goodman, Leo A.; and Kruskal, William H. (1979) *Measures of Association for Cross Classification*.

Grootaert, Christiaan; and Jeanine Braithwaite. (1998) "Poverty Correlates and Indicator-Based Targeting in Eastern Europe and the Former Soviet Union", World Bank Policy Research Working Paper No. 1942, dx.doi.org/10.1596/1813-9450-1942, retrieved 15 May 2016.

Grosh, Margaret; and Judy L. Baker. (1995) "Proxy Means Tests for Targeting Social Programs: Simulations and Speculation", LSMS Working Paper No. 118, poverty2.forumone.com/library/view/5496/, accessed 9 April 2009.

Gwatkin, Davidson R.; Rutstein, Shea; Johnson, Kiersten; Suliman, Eldaw; Wagstaff, Adam; and Agbessi Amouzou. (2007) "Socio-Economic Differences in Health, Nutrition, and Population: Ethiopia", Country Reports on HNP and Poverty, go.worldbank.org/T6LCN5A340, accessed 9 April 2009.

Hagos, Fitsum; and Stein Holden. (2003) "Rural Household Poverty Dynamics in Northern Ethiopia, 1997–2000: Analysis of the Determinants of Poverty", chronicpoverty.org/pdfs/Hagos.pdf, accessed 9 April 2009.

Hand, David J. (2006) "Classifier Technology and the Illusion of Progress", *Statistical Science*, Vol. 22, No. 1, pp. 1–15.

Hoadley, Bruce; and Robert M. Oliver. (1998) "Business Measures of Scorecard Benefit", *IMA Journal of Mathematics Applied in Business and Industry*, Vol. 9, pp. 55–64.

International Comparison Project. (2008) "Tables of Results", siteresources.worldbank.org/ICPINT/Resources/icp-final-tables.pdf, accessed 1 April 2009.

IRIS Center. (2008) "Client Assessment Survey—Ethiopia", povertytools.org/USAID_documents/Tools/Current_Tools/USAID_PAT_ETH_9_2008.xls, accessed 9 April 2009.

_____. (2007a) "Manual for the Implementation of USAID Poverty Assessment Tools", povertytools.org/training_documents/Manuals/USAID_PAT_Manual_Eng.pdf, accessed 9 April 2009.

_____. (2007b) "Introduction to Sampling for the Implementation of PATs", povertytools.org/training_documents/Sampling/Introduction_Sampling.ppt, accessed 9 April 2009.

_____. (2005) "Notes on Assessment and Improvement of Tool Accuracy", povertytools.org/other_documents/AssessingImproving_Accuracy.pdf, accessed 9 April 2009.

Johnson, Glenn. (2007) "Lesson 3: Two-Way Tables—Dependent Samples", http://www.stat.psu.edu/online/development/stat504/03_2way/53_2way_compare.htm, accessed 9 April 2009.

Koenker, Roger; and Kevin F. Hallock. (2001) "Quantile Regression", *Journal of Economic Perspectives*, Vol. 15, No. 4, pp. 143–156.

Kolesar, Peter; and Janet L. Showers. (1985) "A Robust Credit Screening Model Using Categorical Data", *Management Science*, Vol. 31, No. 2, pp. 124–133.

Lindelow, Magnus. (2006) "Sometimes More Equal Than Others: How Health Inequalities Depend on the Choice of Welfare Indicator", *Health Economics*, Vol. 15, pp. 263–279.

Loening, Josef L.; Durevall, Dick; and Yohannes Ayalew Birru. (2009) "Inflation Dynamics and Food Prices in an Agricultural Economy: The Case of Ethiopia", University of Gothenburg Working Papers in Economics No. 347, gupea.ub.gu.se/dspace/bitstream/2077/19464/1/gupea_2077_19464_1.pdf, accessed 9 April 2009.

Lovie, Alexander D.; and Patricia Lovie. (1986) "The Flat Maximum Effect and Linear Scoring Models for Prediction", *Journal of Forecasting*, Vol. 5, pp. 159–168.

Martinelli, César; and Susan W. Parker. (2007) "Deception and Misreporting in a Social Program", ciep.itam.mx/~martinel/lies4.pdf, accessed 9 April 2009.

Matul, Michal; and Sean Kline. (2003) "Scoring Change: Prizma's Approach to Assessing Poverty", Microfinance Centre for Central and Eastern Europe and the New Independent States Spotlight Note No. 4, mfc.org.pl/sites/mfc.org.pl/files/spotlight4.PDF, accessed 9 April 2009.

McNemar, Quinn. (1947) "Note on the Sampling Error of the Difference between Correlated Proportions or Percentages", *Psychometrika*, Vol. 17, pp. 153–157.

Ministry of Finance and Economic Development. (2006) *Ethiopia: Building on Progress, A Plan for Accelerated and Sustained Development to End Poverty (PASDEP), 2005/06–2009/10, Volume I: Main Text*, planipolis.iiep.unesco.org/upload/Ethiopia/Ethiopia_PASDEP_2005_2010.pdf, accessed 9 April 2009.

_____. (2002) "Poverty Profile of Ethiopia: Analysis based on the 1999/00 HICE and WM Survey Results".

Moffitt, Robert. (1991) "Program Evaluation with Non-experimental Data", *Evaluation Review*, Vol. 15, No. 3, pp. 291–314.

Montgomery, Mark; Gragnolati, Michele; Burke, Kathleen A.; and Edmundo Paredes. (2000) "Measuring Living Standards with Proxy Variables", *Demography*, Vol. 37, No. 2, pp. 155–174.

Myers, James H.; and Edward W. Forgy. (1963) "The Development of Numerical Credit Evaluation Systems", *Journal of the American Statistical Association*, Vol. 58, No. 303, pp. 779–806.

Narayan, Ambar; and Nobuo Yoshida. (2005) "Proxy Means Tests for Targeting Welfare Benefits in Sri Lanka", Report No. SASPR–7, siteresources.worldbank.org/EXTSAREGTOPPOVRED/Resources/493440-1102216396155/572861-1102221461685/Proxy+Means+Test+for+Targeting+Welfare+Benefits.pdf, accessed 9 April 2009.

Onwujekwe, Obinna; Hanson, Kara; and Julia Fox-Rushby. (2006) "Some Indicators of Socio-Economic Status May Not Be Reliable and the Use of Indices with These Data Could Worsen Equity", *Health Economics*, Vol. 15, pp. 639–644.

Rutstein, Shea Oscar; and Kiersten Johnson. (2004) "The DHS Wealth Index", DHS Comparative Reports No. 6, measuredhs.com/pubs/pdf/CR6/CR6.pdf, accessed 9 April 2009.

Sahn, David E.; and David Stifel. (2003) "Exploring Alternative Measures of Welfare in the Absence of Expenditure Data", *Review of Income and Wealth*, Series 49, No. 4, pp. 463–489.

_____. (2000) "Poverty Comparisons Over Time and Across Countries in Africa", *World Development*, Vol. 28, No. 12, pp. 2123–2155.

SAS Institute Inc. (2004) "The LOGISTIC Procedure: Rank Correlation of Observed Responses and Predicted Probabilities", in *SAS/STAT User's Guide, Version 9*, support.sas.com/documentation/cdl/en/statug/59654/HTML/default/statug_logistic_sect035.htm, accessed 9 April 2009.

Schreiner, Mark. (2009a) "Simple Poverty Scorecard Poverty-Assessment Tool: Peru", SimplePovertyScorecard.com/PER_2007_ENG.pdf, accessed 9 April 2009.

_____. (2009b) "Simple Poverty Scorecard Poverty-Assessment Tool: Philippines", SimplePovertyScorecard.com/PHL_2002_ENG.pdf, accessed 10 April 2009.

_____. (2008a) "Simple Poverty Scorecard Poverty-Assessment Tool: Peru", SimplePovertyScorecard.com/PER_2003_ENG.pdf, accessed 10 April 2009.

_____. (2008b) "Simple Poverty Scorecard Poverty-Assessment Tool: India", SimplePovertyScorecard.com/IND_2005_ENG.pdf, accessed 9 April 2009.

_____. (2006a) "Is One Simple Poverty Scorecard Poverty-Assessment Tool Enough for India?", microfinance.com/English/Papers/Scoring_Poverty_India_Segments.pdf, accessed 9 April 2009.

_____. (2006b) "Simple Poverty Scorecard Poverty-Assessment Tool: Bangladesh", SimplePovertyScorecard.com/BGD_2000_ENG.pdf, accessed 9 April 2009.

_____. (2005a) "La Herramienta del Índice de Calificación de la Pobreza™: Mexico", SimplePovertyScorecard.com/MEX_2002_SPA.pdf, accessed 9 April 2009.

_____. (2005b) "IRIS Questions on the Simple Poverty Scorecard Poverty-Assessment Tool", microfinance.com/English/Papers/Scoring_Poverty_Response_to_IRIS.pdf, accessed 9 April 2009.

_____. (2002) *Scoring: The Next Breakthrough in Microfinance?* CGAP Occasional Paper No. 7, microfinancegateway.org/files/3276_076.pdf, accessed 9 April 2009.

_____; Matul, Michal; Pawlak, Ewa; and Sean Kline. (2004) "Poverty Scoring: Lessons from a Microlender in Bosnia-Herzegovina", microfinance.com/English/Papers/Scoring_Poverty_in_BiH_Short.pdf, accessed 9 April 2009.

Sillers, Don. (2006) "National and International Poverty Lines: An Overview", pdf.usaid.gov/pdf_docs/Pnadh069.pdf, retrieved 31 May 2012.

Stifel, David; and Luc Christiaensen. (2007) "Tracking Poverty over Time in the Absence of Comparable Consumption Data", *World Bank Economic Review*, Vol. 21, No. 2, pp. 317–341.

Stillwell, William G.; Barron, F. Hutton; and Ward Edwards. (1983) "Evaluating Credit Applications: A Validation of Multi-Attribute Utility Weight Elicitation Techniques", *Organizational Behavior and Human Performance*, Vol. 32, pp. 87–108.

Tafesse, Getahun. (2003) "The Roles of Agricultural Growth and Poverty Reduction in Ethiopia", `ftp://ftp.fao.org/es/ESA/Roa/pdf/3_Poverty/Poverty_Ethiopia.pdf`, accessed 9 April 2009.

Tarozzi, Alesandro. (2008) "Can Census Data Alone Signal Heterogeneity in the Estimation of Poverty Maps?", `www.econ.duke.edu/~taroz/TarozziHet2008.pdf`, accessed 9 April 2009.

_____; and Angus Deaton. (2007) "Using Census and Survey Data to Estimate Poverty and Inequality for Small Areas", `princeton.edu/~deaton/downloads/20080301SmallAreas_FINAL.pdf`, accessed 9 April 2009.

Toohig, Jeff. (2008) "PPI Pilot Training Guide", `progressoutofpoverty.org/toolkit`, accessed 9 April 2009.

Vyas, Seema; and Lilani Kumaranayake. (2006) "Constructing Socio-Economic Status Indices: How to Use Principal Components Analysis", *Health Policy Planning*, Vol. 21, No. 6, pp. 459–468.

Wagstaff, Adam; and Naoko Watanabe. (2003) "What Difference Does the Choice of SES Make in Health Inequality Measurement?", *Health Economics*, Vol. 12, No. 10, pp. 885–890.

Wainer, Howard. (1976) "Estimating Coefficients in Linear Models: It Don't Make No Nevermind", *Psychological Bulletin*, Vol. 83, pp. 223–227.

Woldehanna, Tassew; Hoddinott, John; and Stefan Dercon. (2008) "Poverty and Inequality in Ethiopia: 1995/96–2004/05", `ssrn.com/abstract=1259341`, accessed 9 April 2009.

Zeller, Manfred. (2004) "Review of Poverty Assessment Tools", `pdf.usaid.gov/pdf_docs/PNADH120.pdf`, retrieved 1 February 2011.

_____; Sharma, Manohar; Henry, Carla; and Cécile Lapenu. (2006) "An Operational Method for Assessing the Poverty Outreach Performance of Development Policies and Projects: Results of Case Studies in Africa, Asia, and Latin America", *World Development*, Vol. 34, No. 3, pp. 446–464.

Figure 2: Sample sizes and household poverty rates by sub-sample and poverty line

		% with expenditure below a poverty line			
		International (2005 PPP)			
Sub-sample	**Households**	**$1.00/day**	**$1.25/day**	**$1.75/day**	**$2.50/day**
All Ethiopia	21,297	13.1	27.2	52.3	78.4
Construction					
Selecting indicators and weights	7,177	12.9	27.3	52.7	78.3
Calibration					
Associating scores with likelihoods	6,997	12.9	27.1	52.7	78.2
Validation					
Measuring accuracy	7,123	13.3	27.2	51.5	78.8
Change between construction and calibration to validation (percentage points)					
		-0.4	0.0	1.2	-0.6

Source: 2004/5 HICE and 2004 WMS.

Figure 3: Poverty lines and poverty rates by region

		Poverty line (per capita) and poverty rate (%)			
Region	**Type of rates**	**$1.00/day (2.81 Birr/day)**	**$1.25/day (3.51 Birr/day)**	**$1.75/day (4.91 Birr/day)**	**$2.50/day (7.02 Birr/day)**
Tigray	Household level	16.3	34.1	54.6	74.8
	Person level	22.3	44.1	66.2	83.7
Affar	Household level	10.5	23.7	40.5	65.2
	Person level	15.6	34.7	55.2	78.6
Amhara	Household level	16.7	29.6	57.4	84.0
	Person level	21.8	37.4	66.1	89.5
Oromiya	Household level	8.8	22.3	48.5	76.8
	Person level	12.1	29.0	57.9	84.7
Somali	Household level	16.0	31.3	53.0	74.5
	Person level	22.5	41.9	65.0	85.5
Benshangul-Gumuz	Household level	8.2	22.0	41.3	72.1
	Person level	13.0	31.5	52.8	83.1
S.N.N.P.R	Household level	15.2	31.5	55.8	80.4
	Person level	20.5	40.0	66.2	87.7
Harari	Household level	6.1	18.2	29.1	49.2
	Person level	9.5	25.7	37.5	60.0
Addis Ababa	Household level	12.8	23.9	35.3	57.2
	Person level	18.0	32.6	45.2	66.8
Dire Dawa	Household level	9.8	24.6	44.2	63.9
	Person level	15.2	34.9	56.5	75.4
All Ethiopia	Household level	13.1	27.2	52.3	78.4
	Person level	17.4	34.8	61.8	85.7

Source: 2004/5 HICE and 2004 WMS.

Figure 4: Poverty indicators by uncertainty coefficient

Uncertainty coefficient	Indicator (Answers ordered starting with those most strongly linked with higher poverty likelihoods)
118	How many people are in the household? (Six or more; Five; Four; Three; Two or one)
64	Do all children ages 6 to 12 attend school? (No; Yes; No children ages 6 to 12)
27	How many household members have agriculture/animal husbandry/fishing/forestry as their main occupation? (Three or more; Two; One; None)
13	Does the household currently own any mattresses and/or beds? (No; Yes)
12	Does the household currently own any radios? (No; Yes)
12	What is the highest grade completed by the female head/spouse? (None, first grade, second grade, or no data; Third grade or higher)
11	Does the female head/spouse know how to read and write? (No; Yes)
9	What is the main source of light for the dwelling unit? (Firewood, candles, or other; Kerosene; Electricity (private or shared))
9	Does the household currently own any table and chairs? (No; Yes)
8	What is the highest grade completed by the male head/spouse? (None or no data; Grade 1 to 5; Grade 6 to 8; Grade 9 or higher)
8	Does the household currently own any watches or clocks? (No; Yes)
7	Does the male head/spouse know how to read and write? (No; Yes)
7	What is the main construction material of the walls of the dwelling unit? (Wood and grass, mud and stone, or other; Wood and mud, reeds and bamboo, cement and stone, hollow blocks, or bricks)
7	Does the household currently own any blankets/*gabi*? (No; Yes)
6	What is the main source of cooking fuel? (Mainly firewood (purchase or collected), animal dung, or other; Crop residue; Charcoal, kerosene, butane gas, electricity, or does not use fuel)

Figure 4 (cont.): Poverty indicators by uncertainty coefficient

Uncertainty coefficient	Indicator (Answers ordered starting with those most strongly linked with higher poverty likelihoods)
6	What is the main source of drinking water in the dry season? (River, lake, pond, or other; Protected/unprotected well/spring, or tap outside the compound; Tap in compound (private or shared), or tap inside the house)
5	What is the main construction material of the roof of the dwelling unit? (Thatch and grass, wood and mud, reeds and bamboo, or other; Corrugated iron sheets or clay)
5	What type of toilet facility does the household use? (Pit latrine (shared), field/forest, container (household utensils), or other; Pit latrine (private); Flush toilet (private or shared))
5	Does the household currently own any jewelry (gold/silver)? (No; Yes)
5	Does the household currently own any axes/*gejera* or pick axes/*geso*? (No; Yes)
4	Does the household currently own any sickles/*mecha*? (No; Yes)
4	What is the main source of drinking water in the rainy season? (River, lake, pond, or other; Protected well/spring; Unprotected well/spring or tap water outside the compound; Rainwater, tap in compound (private or shared), or tap inside the house)
4	Does the household currently own any pack animals? (No; Yes)
4	Does the household currently own any plows? (No; Yes)
3	Does the household currently own any stoves (gas or electric)? (No; Yes)
3	Do any household members work in salaried jobs as their main occupation? (No; Yes)
3	Does the household currently own any *mofer* or *kember*? (No; Yes)
3	Does the household currently own a plow and/or plowing animals? (No; Yes)
2	Does the household currently own any cattle, sheep, or goats? (No; Yes)
2	Excluding kitchen and toilets, how many rooms does the dwelling unit have? (One; Two; Three or more)

Source: 2004/5 HICE and 2004 WMS, $1.25/day line in 2005 PPP.

$1.00/Day 2005 PPP Poverty Line Tables

(and tables pertaining to all four poverty lines)

Figure 5 ($1.00/day 2005 PPP line): Estimated poverty likelihoods associated with scores

If a household's score is . . .	. . . then the likelihood (%) of being below the poverty line is:
0–4	38.3
5–9	59.6
10–14	38.3
15–19	29.2
20–24	24.0
25–29	17.5
30–34	13.4
35–39	8.0
40–44	8.3
45–49	4.9
50–54	1.9
55–59	3.0
60–64	0.8
65–69	0.7
70–74	0.0
75–79	0.3
80–84	0.0
85–89	0.0
90–94	0.0
95–100	0.0

Based on the 2004/5 HICE and 2004 WMS.

Figure 6 ($1.00/day 2005 PPP line): Derivation of estimated poverty likelihoods associated with scores

Score	Households below poverty line		All households at score		Poverty likelihood (estimated, %)
0–4	69	÷	179	=	38.3
5–9	582	÷	976	=	59.6
10–14	989	÷	2,583	=	38.3
15–19	2,624	÷	8,978	=	29.2
20–24	2,534	÷	10,580	=	24.0
25–29	2,269	÷	12,941	=	17.5
30–34	1,474	÷	10,964	=	13.4
35–39	877	÷	10,973	=	8.0
40–44	761	÷	9,216	=	8.3
45–49	357	÷	7,285	=	4.9
50–54	158	÷	8,366	=	1.9
55–59	157	÷	5,150	=	3.0
60–64	33	÷	4,388	=	0.8
65–69	24	÷	3,300	=	0.7
70–74	0	÷	1,930	=	0.0
75–79	3	÷	1,178	=	0.3
80–84	0	÷	676	=	0.0
85–89	0	÷	315	=	0.0
90–94	0	÷	23	=	0.0
95–100	0	÷	0	=	0.0

Surveyed cases weighted to represent Ethiopia's households.
Number of cases normalized to total to 100,000.

Figure 7 (All poverty lines): Distribution of household poverty likelihoods across ranges demarcated by poverty lines

	Likelihood of having expenditure in range demarcated by poverty lines per day per capita				
	<$1.00/day	≥$1.00/day and <$1.25/day	≥$1.25/day and <$1.75/day	≥$1.75/day and <$2.50/day	≥$2.50/day
Score	<Birr 2.81	≥Birr 2.81 and <Birr 3.51	≥Birr 3.51 and <Birr 4.91	≥Birr 4.91 and <Birr 7.02	≥Birr 7.02
0–4	38.3	49.3	0.0	12.4	0.0
5–9	59.6	23.4	12.5	4.5	0.0
10–14	38.3	25.3	18.6	15.6	2.2
15–19	29.2	29.1	25.3	11.7	4.7
20–24	24.0	23.7	31.1	17.0	4.3
25–29	17.5	21.0	32.7	22.6	6.2
30–34	13.4	14.9	32.4	28.8	10.4
35–39	8.0	10.5	39.1	27.7	14.7
40–44	8.3	10.1	28.6	33.1	19.9
45–49	4.9	13.7	24.8	32.1	24.4
50–54	1.9	5.5	19.4	39.3	33.9
55–59	3.0	1.9	18.6	30.3	46.1
60–64	0.8	2.2	9.9	31.5	55.7
65–69	0.7	1.5	6.0	27.1	64.8
70–74	0.0	0.5	6.9	21.2	71.4
75–79	0.3	0.8	5.7	10.1	83.2
80–84	0.0	3.4	4.5	7.6	84.5
85–89	0.0	9.3	0.8	18.3	71.6
90–94	0.0	0.0	0.0	0.0	100.0
95–100	0.0	0.0	0.0	0.0	100.0

Note: All poverty likelihoods in percentage units. All dollar poverty lines in units of 2005 PPP.

Figure 8 ($1.00/day 2005 PPP line): Bootstrapped differences between estimated and true poverty likelihoods for households in a large sample (n = 16,384) from the validation sample, with confidence intervals

	Difference between estimate and true value			
		Confidence interval (+/− percentage points)		
Score	Diff.	90-percent	95-percent	99-percent
0–4	–43.1	27.3	28.6	29.3
5–9	+38.8	5.4	6.7	8.8
10–14	+1.5	4.0	4.8	6.3
15–19	–2.6	2.3	2.5	3.2
20–24	–0.8	1.8	2.1	2.7
25–29	+2.0	1.3	1.6	2.1
30–34	–0.5	1.4	1.7	2.2
35–39	–1.0	1.2	1.4	1.9
40–44	+3.7	0.9	1.1	1.4
45–49	+2.1	0.8	1.0	1.3
50–54	–2.1	1.5	1.6	1.8
55–59	+0.8	1.0	1.2	1.6
60–64	+0.5	0.2	0.2	0.2
65–69	+0.7	0.0	0.0	0.0
70–74	–0.1	0.1	0.1	0.1
75–79	–5.4	4.6	5.0	6.1
80–84	+0.0	0.0	0.0	0.0
85–89	+0.0	0.0	0.0	0.0
90–94	+0.0	0.0	0.0	0.0
95–100	+0.0	0.0	0.0	0.0

Based on scorecard applied to the validation sample.

Figure 9 (All poverty lines): Differences, precision of differences, and sample-size α for bootstrapped estimates of poverty rates for groups of households at a point in time for the scorecard applied to the validation sample

	Poverty line			
	International (2005 PPP)			
	$1.00/day	**$1.25/day**	**$1.75/day**	**$2.50/day**
Estimate minus true value	0.5	1.4	2.1	-1.3
Precision of difference	0.4	0.6	0.6	0.5
α factor	1.03	1.00	0.96	0.90

Precision is measured as 90-percent confidence intervals in units of +/− percentage points.
Differences and precision estimated from 1,000 bootstraps of size n = 16,384.
α is estimated from 1,000 bootstrap samples of n = 256, 512, 1,024, 2,048, 4,096, 8,192, and 16,384.

Figure 10 ($1.00/day 2005 PPP line): Differences and precision of differences for bootstrapped estimates of poverty rates for groups of households at a point in time, by sample size, scorecard applied to validation sample

	Difference between estimate and true value			
		Confidence interval (+/− percentage points)		
Sample size (n)	Diff.	90-percent	95-percent	99-percent
1	+0.7	55.9	57.9	78.3
4	+1.6	26.0	32.0	41.4
8	+0.9	19.5	22.4	29.0
16	+0.4	13.8	17.2	24.1
32	+0.4	10.0	12.2	16.2
64	+0.4	7.3	8.9	11.6
128	+0.5	4.9	5.9	7.7
256	+0.5	3.6	4.3	5.7
512	+0.4	2.4	2.9	3.7
1,024	+0.5	1.8	2.2	2.8
2,048	+0.5	1.3	1.6	1.9
4,096	+0.5	0.9	1.1	1.5
8,192	+0.5	0.6	0.7	1.0
16,384	+0.5	0.4	0.5	0.7

Figure 11 (All poverty lines): Possible types of outcomes from targeting by poverty score

True poverty status	Targeting segment: Targeted	Targeting segment: Non-targeted
Below poverty line	**Inclusion** Under poverty line Correctly Targeted	**Undercoverage** Under poverty line Mistakenly Non-targeted
Above poverty line	**Leakage** Above poverty line Mistakenly Targeted	**Exclusion** Above poverty line Correctly Non-targeted

Figure 12 ($1.00/day 2005 PPP line): Households by targeting classification and score, along with "Total Accuracy" and BPAC, scorecard applied to validation sample

Score	Inclusion: < poverty line correctly targeted	Undercoverage: < poverty line mistakenly non-targeted	Leakage: ≥ poverty line mistakenly targeted	Exclusion: ≥ poverty line correctly non-targeted	Total Accuracy Inclusion + Exclusion	BPAC See text
0–4	0.1	13.2	0.0	86.6	86.8	–97.6
5–9	0.4	13.0	0.8	85.9	86.3	–88.5
10–14	1.4	12.0	2.4	84.3	85.7	–61.6
15–19	4.4	9.0	8.4	78.3	82.7	+28.1
20–24	7.1	6.2	16.2	70.5	77.6	–21.6
25–29	9.4	4.0	26.9	59.8	69.2	–101.5
30–34	10.9	2.4	36.3	50.4	61.3	–172.4
35–39	12.0	1.3	46.2	40.5	52.5	–246.4
40–44	12.5	0.8	54.9	31.8	44.3	–311.7
45–49	12.8	0.5	61.9	24.8	37.6	–364.3
50–54	13.1	0.2	69.9	16.8	29.9	–424.5
55–59	13.2	0.1	74.9	11.7	25.0	–462.3
60–64	13.3	0.1	79.3	7.4	20.6	–494.9
65–69	13.3	0.0	82.6	4.1	17.4	–519.7
70–74	13.3	0.0	84.5	2.2	15.4	–534.1
75–79	13.3	0.0	85.7	1.0	14.3	–542.6
80–84	13.3	0.0	86.3	0.3	13.7	–547.7
85–89	13.3	0.0	86.6	0.0	13.4	–550.1
90–94	13.3	0.0	86.7	0.0	13.3	–550.2
95–100	13.3	0.0	86.7	0.0	13.3	–550.2

Inclusion, undercoverage, leakage, and exclusion normalized to sum to 100.

Figure 13 ($1.00/day 2005 PPP line): Households below the poverty line and all households at a given score or at or below a given score cut-off, scorecard applied to validation sample

Targeting cut-off	% all households who are targeted	% targeted who are poor	% of poor who are targeted	Poor households targeted per non-poor household targeted
0–4	0.2	82.0	1.1	4.6:1
5–9	1.2	32.8	2.8	0.5:1
10–14	3.7	36.8	10.3	0.6:1
15–19	12.7	34.2	32.7	0.5:1
20–24	23.3	30.4	53.2	0.4:1
25–29	36.2	25.9	70.4	0.3:1
30–34	47.2	23.1	81.8	0.3:1
35–39	58.2	20.6	90.0	0.3:1
40–44	67.4	18.6	93.9	0.2:1
45–49	74.7	17.1	95.9	0.2:1
50–54	83.0	15.8	98.5	0.2:1
55–59	88.2	15.0	99.3	0.2:1
60–64	92.6	14.3	99.6	0.2:1
65–69	95.9	13.9	99.6	0.2:1
70–74	97.8	13.6	99.7	0.2:1
75–79	99.0	13.5	100.0	0.2:1
80–84	99.7	13.4	100.0	0.2:1
85–89	100.0	13.3	100.0	0.2:1
90–94	100.0	13.3	100.0	0.2:1
95–100	100.0	13.3	100.0	0.2:1

$1.25/Day 2005 PPP Poverty Line Tables

Figure 5 ($1.25/day 2005 PPP line): Estimated poverty likelihoods associated with scores

If a household's score is . . .	. . . then the likelihood (%) of being below the poverty line is:
0–4	87.6
5–9	82.9
10–14	63.6
15–19	58.3
20–24	47.7
25–29	38.5
30–34	28.4
35–39	18.5
40–44	18.4
45–49	18.6
50–54	7.4
55–59	5.0
60–64	3.0
65–69	2.2
70–74	0.5
75–79	1.1
80–84	3.4
85–89	9.3
90–94	0.0
95–100	0.0

Based on the 2004/5 HICE and 2004 WMS.

Figure 6 ($1.25/day 2005 PPP line): Derivation of estimated poverty likelihoods associated with scores

Score	Households below poverty line		All households at score		Poverty likelihood (estimated, %)
0–4	157	÷	179	=	87.6
5–9	810	÷	976	=	82.9
10–14	1,642	÷	2,583	=	63.6
15–19	5,236	÷	8,978	=	58.3
20–24	5,045	÷	10,580	=	47.7
25–29	4,982	÷	12,941	=	38.5
30–34	3,108	÷	10,964	=	28.4
35–39	2,027	÷	10,973	=	18.5
40–44	1,691	÷	9,216	=	18.4
45–49	1,358	÷	7,285	=	18.6
50–54	615	÷	8,366	=	7.4
55–59	256	÷	5,150	=	5.0
60–64	130	÷	4,388	=	3.0
65–69	73	÷	3,300	=	2.2
70–74	10	÷	1,930	=	0.5
75–79	12	÷	1,178	=	1.1
80–84	23	÷	676	=	3.4
85–89	29	÷	315	=	9.3
90–94	0	÷	23	=	0.0
95–100	0	÷	0	=	0.0

Surveyed cases weighted to represent Ethiopia's households.
Number of cases normalized to total to 100,000.

Figure 8 ($1.25/day 2005 PPP line): Bootstrapped differences between estimated and true poverty likelihoods for households in a large sample (n = 16,384) from the validation sample, with confidence intervals

	Difference between estimate and true value			
		Confidence interval (+/– percentage points)		
Score	**Diff.**	**90-percent**	**95-percent**	**99-percent**
0–4	+6.1	12.2	14.8	17.4
5–9	+31.5	6.9	8.1	11.2
10–14	+2.0	4.1	4.8	6.6
15–19	+2.7	2.3	2.7	3.4
20–24	+0.7	2.2	2.5	3.2
25–29	–0.0	1.9	2.2	2.8
30–34	–4.4	3.2	3.4	3.7
35–39	+0.5	1.6	1.9	2.5
40–44	+3.4	1.6	2.0	2.6
45–49	+9.4	1.4	1.7	2.4
50–54	–0.4	1.2	1.5	1.9
55–59	+0.2	1.4	1.7	2.1
60–64	+1.8	0.6	0.7	0.9
65–69	+2.0	0.2	0.2	0.3
70–74	+0.4	0.1	0.1	0.2
75–79	–4.6	4.3	4.6	5.7
80–84	+3.4	0.0	0.0	0.0
85–89	+9.3	0.0	0.0	0.0
90–94	+0.0	0.0	0.0	0.0
95–100	+0.0	0.0	0.0	0.0

Based on scorecard applied to the validation sample.

Figure 10 ($1.25/day 2005 PPP line): Differences and precision of differences for bootstrapped estimates of poverty rates for groups of households at a point in time, by sample size, scorecard applied to validation sample

	Difference between estimate and true value			
		Confidence interval (+/− percentage points)		
Sample size (n)	Diff.	90-percent	95-percent	99-percent
1	+0.4	69.8	69.9	87.8
4	+1.7	36.8	44.3	54.3
8	+1.3	25.5	30.8	39.9
16	+1.1	19.0	22.4	29.1
32	+1.1	13.0	15.2	19.7
64	+1.0	9.6	11.5	14.5
128	+1.1	6.7	8.3	10.8
256	+1.3	4.6	5.6	7.1
512	+1.3	3.1	3.9	5.1
1,024	+1.4	2.2	2.8	3.5
2,048	+1.4	1.6	1.9	2.4
4,096	+1.4	1.1	1.3	1.7
8,192	+1.4	0.8	1.0	1.3
16,384	+1.4	0.6	0.7	0.9

Figure 12 ($1.25/day 2005 PPP line): Households by targeting classification and score, along with "Total Accuracy" and BPAC, scorecard applied to validation sample

Score	Inclusion: < poverty line correctly targeted	Undercoverage: < poverty line mistakenly non-targeted	Leakage: ≥ poverty line mistakenly targeted	Exclusion: ≥ poverty line correctly non-targeted	Total Accuracy Inclusion + Exclusion	BPAC See text
0–4	0.1	27.1	0.0	72.7	72.9	–98.8
5–9	0.7	26.6	0.5	72.3	72.9	–93.4
10–14	2.3	24.9	1.4	71.3	73.6	–77.8
15–19	7.4	19.9	5.4	67.4	74.8	–26.2
20–24	12.5	14.7	10.8	62.0	74.5	+31.5
25–29	17.8	9.4	18.4	54.4	72.2	+32.4
30–34	21.5	5.7	25.7	47.1	68.6	+5.6
35–39	23.7	3.5	34.4	38.4	62.1	–26.5
40–44	25.2	2.0	42.2	30.6	55.9	–54.9
45–49	26.0	1.2	48.6	24.1	50.2	–78.7
50–54	26.8	0.5	56.3	16.5	43.2	–106.8
55–59	27.0	0.2	61.2	11.6	38.6	–124.7
60–64	27.1	0.1	65.4	7.3	34.5	–140.4
65–69	27.2	0.1	68.7	4.1	31.2	–152.4
70–74	27.2	0.0	70.6	2.2	29.3	–159.5
75–79	27.2	0.0	71.8	1.0	28.2	–163.6
80–84	27.2	0.0	72.4	0.3	27.6	–166.1
85–89	27.2	0.0	72.8	0.0	27.2	–167.3
90–94	27.2	0.0	72.8	0.0	27.2	–167.4
95–100	27.2	0.0	72.8	0.0	27.2	–167.4

Inclusion, undercoverage, leakage, and exclusion normalized to sum to 100.

Figure 13 ($1.25/day 2005 PPP line): Households below the poverty line and all households at a given score or at or below a given score cut-off, scorecard applied to validation sample

Targeting cut-off	% all households who are targeted	% targeted who are poor	% of poor who are targeted	Poor households targeted per non-poor household targeted
0–4	0.2	82.7	0.5	4.8:1
5–9	1.2	56.5	2.4	1.3:1
10–14	3.7	61.6	8.5	1.6:1
15–19	12.7	57.9	27.1	1.4:1
20–24	23.3	53.7	46.0	1.2:1
25–29	36.2	49.2	65.6	1.0:1
30–34	47.2	45.6	79.0	0.8:1
35–39	58.2	40.8	87.2	0.7:1
40–44	67.4	37.4	92.7	0.6:1
45–49	74.7	34.9	95.6	0.5:1
50–54	83.0	32.2	98.3	0.5:1
55–59	88.2	30.6	99.3	0.4:1
60–64	92.6	29.3	99.7	0.4:1
65–69	95.9	28.3	99.8	0.4:1
70–74	97.8	27.8	99.9	0.4:1
75–79	99.0	27.5	100.0	0.4:1
80–84	99.7	27.3	100.0	0.4:1
85–89	100.0	27.2	100.0	0.4:1
90–94	100.0	27.2	100.0	0.4:1
95–100	100.0	27.2	100.0	0.4:1

$1.75/Day 2005 PPP Poverty Line Tables

Figure 5 ($1.75/day 2005 PPP line): Estimated poverty likelihoods associated with scores

If a household's score is . . .	. . . then the likelihood (%) of being below the poverty line is:
0–4	87.6
5–9	95.5
10–14	82.2
15–19	83.6
20–24	78.8
25–29	71.2
30–34	60.8
35–39	57.6
40–44	46.9
45–49	43.4
50–54	26.8
55–59	23.5
60–64	12.8
65–69	8.2
70–74	7.4
75–79	6.8
80–84	7.9
85–89	10.1
90–94	0.0
95–100	0.0

Based on the 2004/5 HICE and 2004 WMS.

Figure 6 ($1.75/day 2005 PPP line): Derivation of estimated poverty likelihoods associated with scores

Score	Households below poverty line		All households at score		Poverty likelihood (estimated, %)
0–4	157	÷	179	=	87.6
5–9	932	÷	976	=	95.5
10–14	2,123	÷	2,583	=	82.2
15–19	7,506	÷	8,978	=	83.6
20–24	8,336	÷	10,580	=	78.8
25–29	9,211	÷	12,941	=	71.2
30–34	6,662	÷	10,964	=	60.8
35–39	6,316	÷	10,973	=	57.6
40–44	4,325	÷	9,216	=	46.9
45–49	3,163	÷	7,285	=	43.4
50–54	2,240	÷	8,366	=	26.8
55–59	1,212	÷	5,150	=	23.5
60–64	563	÷	4,388	=	12.8
65–69	270	÷	3,300	=	8.2
70–74	144	÷	1,930	=	7.4
75–79	80	÷	1,178	=	6.8
80–84	54	÷	676	=	7.9
85–89	32	÷	315	=	10.1
90–94	0	÷	23	=	0.0
95–100	0	÷	0	=	0.0

Surveyed cases weighted to represent Ethiopia's households.
Number of cases normalized to total to 100,000.

Figure 8 ($1.75/day 2005 PPP line): Bootstrapped differences between estimated and true poverty likelihoods for households in a large sample (n = 16,384) from the validation sample, with confidence intervals

	Difference between estimate and true value			
		Confidence interval (+/– percentage points)		
Score	Diff.	90-percent	95-percent	99-percent
0–4	–12.4	6.2	6.2	6.2
5–9	+3.8	3.5	4.2	5.5
10–14	–7.8	5.2	5.5	6.0
15–19	+3.8	1.8	2.2	2.9
20–24	+1.4	1.9	2.2	2.9
25–29	+0.2	1.8	2.2	2.8
30–34	–5.3	3.6	3.8	4.2
35–39	+7.4	2.1	2.5	3.2
40–44	+4.6	2.4	2.8	3.8
45–49	+7.4	2.6	3.1	4.1
50–54	+0.6	2.2	2.5	3.3
55–59	+9.0	2.2	2.6	3.3
60–64	–0.6	2.3	2.8	3.5
65–69	–1.4	2.4	2.9	3.8
70–74	+4.1	1.8	2.0	2.8
75–79	+1.1	3.7	4.2	5.7
80–84	+7.9	0.0	0.0	0.0
85–89	+10.1	0.0	0.0	0.0
90–94	+0.0	0.0	0.0	0.0
95–100	+0.0	0.0	0.0	0.0

Based on scorecard applied to the validation sample.

Figure 10 ($1.75/day 2005 PPP line): Differences and precision of differences for bootstrapped estimates of poverty rates for groups of households at a point in time, by sample size, scorecard applied to validation sample

	Difference between estimate and true value			
		Confidence interval (+/− percentage points)		
Sample size (n)	Diff.	90-percent	95-percent	99-percent
1	–0.3	72.2	77.6	87.7
4	+0.9	39.6	46.6	59.4
8	+1.4	27.8	34.0	42.5
16	+1.6	19.4	23.5	30.2
32	+1.8	13.6	16.8	22.4
64	+1.9	9.8	11.5	14.1
128	+1.9	6.8	8.2	11.1
256	+2.0	4.7	5.7	7.8
512	+2.0	3.4	4.0	5.5
1,024	+2.1	2.5	3.0	3.8
2,048	+2.1	1.8	2.1	2.7
4,096	+2.1	1.3	1.4	2.0
8,192	+2.1	0.8	1.0	1.4
16,384	+2.1	0.6	0.7	1.0

Figure 12 ($1.75/day 2005 PPP line): Households by targeting classification and score, along with "Total Accuracy" and BPAC, scorecard applied to validation sample

Score	Inclusion: < poverty line correctly targeted	Undercoverage: < poverty line mistakenly non-targeted	Leakage: ≥ poverty line mistakenly targeted	Exclusion: ≥ poverty line correctly non-targeted	Total Accuracy Inclusion + Exclusion	BPAC See text
0–4	0.2	51.3	0.0	48.5	48.7	–99.3
5–9	1.1	50.4	0.1	48.4	49.5	–95.7
10–14	3.4	48.1	0.3	48.2	51.6	–86.1
15–19	10.7	40.8	2.1	46.5	57.1	–54.6
20–24	18.9	32.6	4.4	44.1	63.0	–18.1
25–29	28.2	23.3	8.1	40.5	68.7	+25.1
30–34	35.2	16.2	12.0	36.6	71.8	+60.2
35–39	40.9	10.6	17.3	31.2	72.1	+66.4
40–44	44.8	6.7	22.6	26.0	70.8	+56.2
45–49	47.5	4.0	27.2	21.3	68.8	+47.2
50–54	49.7	1.8	33.4	15.1	64.8	+35.1
55–59	50.5	1.0	37.7	10.8	61.3	+26.8
60–64	51.1	0.4	41.5	7.0	58.1	+19.3
65–69	51.4	0.1	44.5	4.0	55.4	+13.5
70–74	51.4	0.0	46.4	2.1	53.6	+9.9
75–79	51.5	0.0	47.5	1.0	52.5	+7.7
80–84	51.5	0.0	48.2	0.3	51.8	+6.4
85–89	51.5	0.0	48.5	0.0	51.5	+5.8
90–94	51.5	0.0	48.5	0.0	51.5	+5.7
95–100	51.5	0.0	48.5	0.0	51.5	+5.7

Inclusion, undercoverage, leakage, and exclusion normalized to sum to 100.

Figure 13 ($1.75/day 2005 PPP line): Households below the poverty line and all households at a given score or at or below a given score cut-off, scorecard applied to validation sample

Targeting cut-off	% all households who are targeted	% targeted who are poor	% of poor who are targeted	Poor households targeted per non-poor household targeted
0–4	0.2	100.0	0.3	Only poor targeted
5–9	1.2	91.6	2.1	10.9:1
10–14	3.7	91.1	6.6	10.2:1
15–19	12.7	83.8	20.7	5.2:1
20–24	23.3	81.1	36.7	4.3:1
25–29	36.2	77.8	54.7	3.5:1
30–34	47.2	74.7	68.5	2.9:1
35–39	58.2	70.2	79.4	2.4:1
40–44	67.4	66.5	87.1	2.0:1
45–49	74.7	63.6	92.3	1.7:1
50–54	83.0	59.8	96.5	1.5:1
55–59	88.2	57.2	98.1	1.3:1
60–64	92.6	55.1	99.2	1.2:1
65–69	95.9	53.6	99.8	1.2:1
70–74	97.8	52.6	99.9	1.1:1
75–79	99.0	52.0	100.0	1.1:1
80–84	99.7	51.6	100.0	1.1:1
85–89	100.0	51.5	100.0	1.1:1
90–94	100.0	51.5	100.0	1.1:1
95–100	100.0	51.5	100.0	1.1:1

$2.50/Day 2005 PPP Poverty Line Tables

Figure 5 ($2.50/day 2005 PPP line): Estimated poverty likelihoods associated with scores

If a household's score is . . .	. . . then the likelihood (%) of being below the poverty line is:
0–4	100.0
5–9	100.0
10–14	97.8
15–19	95.3
20–24	95.7
25–29	93.8
30–34	89.6
35–39	85.3
40–44	80.1
45–49	75.6
50–54	66.1
55–59	53.9
60–64	44.3
65–69	35.3
70–74	28.7
75–79	16.8
80–84	15.5
85–89	28.4
90–94	0.0
95–100	0.0

Based on the 2004/5 HICE and 2004 WMS.

Figure 6 ($2.50/day 2005 PPP line): Derivation of estimated poverty likelihoods associated with scores

Score	Households below poverty line		All households at score		Poverty likelihood (estimated, %)
0–4	179	÷	179	=	100.0
5–9	976	÷	976	=	100.0
10–14	2,526	÷	2,583	=	97.8
15–19	8,553	÷	8,978	=	95.3
20–24	10,129	÷	10,580	=	95.7
25–29	12,141	÷	12,941	=	93.8
30–34	9,821	÷	10,964	=	89.6
35–39	9,355	÷	10,973	=	85.3
40–44	7,378	÷	9,216	=	80.1
45–49	5,504	÷	7,285	=	75.6
50–54	5,528	÷	8,366	=	66.1
55–59	2,774	÷	5,150	=	53.9
60–64	1,945	÷	4,388	=	44.3
65–69	1,163	÷	3,300	=	35.3
70–74	553	÷	1,930	=	28.7
75–79	198	÷	1,178	=	16.8
80–84	105	÷	676	=	15.5
85–89	89	÷	315	=	28.4
90–94	0	÷	23	=	0.0
95–100	0	÷	0	=	0.0

Surveyed cases weighted to represent Ethiopia's households.
Number of cases normalized to total to 100,000.

Figure 8 ($2.50/day 2005 PPP line): Bootstrapped differences between estimated and true poverty likelihoods for households in a large sample (n = 16,384) from the validation sample, with confidence intervals

	Scorecard applied to the validation sample Difference between estimate and true value			
		Confidence interval (+/− percentage points)		
Score	Diff.	90-percent	95-percent	99-percent
0–4	+0.0	0.0	0.0	0.0
5–9	+3.9	2.6	3.2	4.0
10–14	+2.0	1.9	2.3	3.1
15–19	–0.8	0.9	1.1	1.5
20–24	–2.5	1.5	1.5	1.6
25–29	–1.7	1.2	1.3	1.4
30–34	–1.1	1.2	1.4	1.9
35–39	+0.2	1.5	1.9	2.5
40–44	–4.6	3.1	3.3	3.5
45–49	+1.1	2.4	2.9	3.8
50–54	+0.8	2.4	2.8	3.4
55–59	+0.0	3.3	3.9	5.1
60–64	+0.5	3.5	4.0	5.1
65–69	–16.6	10.2	10.7	11.6
70–74	–4.9	5.2	6.5	8.5
75–79	+3.7	4.9	6.2	7.4
80–84	–3.4	7.4	8.7	11.7
85–89	+28.4	0.0	0.0	0.0
90–94	+0.0	0.0	0.0	0.0
95–100	+0.0	0.0	0.0	0.0

Based on scorecard applied to the validation sample.

Figure 10 ($2.50/day 2005 PPP line): Differences and precision of differences for bootstrapped estimates of poverty rates for groups of households at a point in time, by sample size, scorecard applied to validation sample

	Difference between estimate and true value			
		Confidence interval (+/− percentage points)		
Sample size (n)	Diff.	90-percent	95-percent	99-percent
1	–0.4	63.1	75.0	83.3
4	–1.4	30.3	38.0	51.2
8	–1.3	21.2	25.2	32.2
16	–1.7	14.8	16.9	22.6
32	–1.7	10.8	12.8	16.9
64	–1.6	7.3	8.9	11.4
128	–1.5	5.3	6.1	8.1
256	–1.4	3.8	4.6	5.8
512	–1.4	2.7	3.1	3.9
1,024	–1.3	1.8	2.2	2.9
2,048	–1.3	1.3	1.6	2.1
4,096	–1.3	1.0	1.1	1.4
8,192	–1.3	0.7	0.8	1.0
16,384	–1.3	0.5	0.6	0.7

Figure 12 ($2.50/day 2005 PPP line): Households by targeting classification and score, along with "Total Accuracy" and BPAC, scorecard applied to validation sample

Score	Inclusion: < poverty line correctly targeted	Undercoverage: < poverty line mistakenly non-targeted	Leakage: ≥ poverty line mistakenly targeted	Exclusion: ≥ poverty line correctly non-targeted	Total Accuracy Inclusion + Exclusion	BPAC See text
0–4	0.2	78.6	0.0	21.2	21.4	–99.5
5–9	1.1	77.7	0.0	21.1	22.3	–97.1
10–14	3.6	75.2	0.1	21.1	24.7	–90.7
15–19	12.3	66.5	0.4	20.7	33.0	–68.3
20–24	22.6	56.2	0.7	20.5	43.1	–41.8
25–29	34.9	43.9	1.3	19.9	54.8	–9.7
30–34	44.7	34.2	2.5	18.6	63.3	+16.6
35–39	53.9	24.9	4.3	16.9	70.8	+42.2
40–44	61.4	17.4	6.0	15.2	76.7	+63.5
45–49	66.7	12.1	8.0	13.2	79.9	+79.4
50–54	72.1	6.8	11.0	10.2	82.3	+86.1
55–59	74.7	4.1	13.5	7.7	82.4	+82.9
60–64	76.6	2.3	16.0	5.2	81.7	+79.7
65–69	78.0	0.8	17.9	3.3	81.3	+77.3
70–74	78.6	0.2	19.2	1.9	80.5	+75.6
75–79	78.7	0.1	20.3	0.9	79.6	+74.3
80–84	78.8	0.0	20.8	0.3	79.2	+73.5
85–89	78.8	0.0	21.2	0.0	78.8	+73.1
90–94	78.8	0.0	21.2	0.0	78.8	+73.1
95–100	78.8	0.0	21.2	0.0	78.8	+73.1

Inclusion, undercoverage, leakage, and exclusion normalized to sum to 100.

Figure 13 ($2.50/day 2005 PPP line): Households below the poverty line and all households at a given score or at or below a given score cut-off, scorecard applied to validation sample

Targeting cut-off	% all households who are targeted	% targeted who are poor	% of poor who are targeted	Poor households targeted per non-poor household targeted
0–4	0.2	100.0	0.2	Only poor targeted
5–9	1.2	96.9	1.4	31.5:1
10–14	3.7	96.7	4.6	29.7:1
15–19	12.7	96.5	15.6	27.5:1
20–24	23.3	97.1	28.7	33.1:1
25–29	36.2	96.4	44.3	26.7:1
30–34	47.2	94.6	56.7	17.6:1
35–39	58.2	92.7	68.4	12.6:1
40–44	67.4	91.2	77.9	10.3:1
45–49	74.7	89.3	84.6	8.4:1
50–54	83.0	86.8	91.4	6.6:1
55–59	88.2	84.7	94.8	5.5:1
60–64	92.6	82.7	97.1	4.8:1
65–69	95.9	81.4	99.0	4.4:1
70–74	97.8	80.3	99.7	4.1:1
75–79	99.0	79.5	99.9	3.9:1
80–84	99.7	79.1	100.0	3.8:1
85–89	100.0	78.8	100.0	3.7:1
90–94	100.0	78.8	100.0	3.7:1
95–100	100.0	78.8	100.0	3.7:1