Microfinance Risk Management, L.L.C.
6970 Chippewa St. #1W
Saint Louis, MO 63109–3060, U.S.A.
http://www.microfinance.com
+1 (314) 752–9022

March 21, 2006

To: Nigel Biggar, Grameen Foundation U.S.A.

From: Mark Schreiner

Re: Response to IRIS questions on poverty scorecards

This memo responds to your request for comments on IRIS' latest questions on approaches to measuring poverty status.

I have been in communication with IRIS since Jan. 2004 when Chris Dunford of Freedom from Hunger shared my short concept note outlining the basic ideas of poverty scoring with national survey data. For example, I pointed out that their proposed approach for dealing with measurement error in expenditure (IRIS, 2005c) ended up counting some people *above* the poverty line as poor. In another instance, I explained why "MAXR" tends to select indicators that correctly classify the very rich but has greater difficulty distinguishing between borderline cases, reducing "Poverty Accuracy". Finally, I encouraged IRIS to make the assumptions behind their performance criterion explicit, something that has occurred with the new "BPAC". Along with our phone call with Thierry and Christiaan, many questions (including some in the current request) have been discussed, and I like to think that this has helped improve poverty measurement. I am hopeful that Schreiner *et al.* (2004) has been useful for IRIS, much as Grootaert and Braithwaite (1998) has been for us. In this spirit, I hope this exchange will improve poverty measurement and help the poor. The main points are:

- Accuracy matters, but "practicality" also matters (and probably matters more)
- Small technical differences probably "don't make no nevermind"
- Logit is standard, in theory and practice
- Quantile and two-step methods can indeed improve classification of the poor

After discussing these points, the memo addresses IRIS' specific questions.

In its search for simple tools useful to its affiliates, GFUSA—like IRIS—has produced scorecards that are quantitative, data-based, and empirically validated, fulfilling the Congressional mandate and worthy of USAID's imprimatur.

Accuracy matters, but "practicality" also matters (and probably matters more)

Our main goal is not to maximize scorecard accuracy but rather to maximize the likelihood of the scorecard's being used. When scoring projects fail, it is rarely because scorecards are inaccurate and almost always because front-line users refuse (or do not bother) to use the scorecards or to use them properly (Schreiner, 2002). The challenge is not technical but human and organizational, not statistics but change management. "Accuracy" is easier to achieve—and less important—than "practicality".

We aim to make poverty scoring something that program managers understand, trust, and therefore use routinely and properly. While we value accuracy, we are often willing to trade it off against simplicity, ease-of-use, and "face validity".

We also trade accuracy against cost. For scorecard construction, we have a budget of a month and $10,000 per country. For scorecard use, programs are more likely to collect data, compute scores, and pay attention to results if the process avoids "extra" work.

To this end, we focus on indicators that are simple, verifiable, difficult-to-falsify, and (ideally) already collected. We avoid indicators such as "total assets" or "expenditure on footwear and clothing" that are both next-to-impossible to measure accurately and also extremely time-consuming to collect.

We also want to measure *changes* in poverty through time (hence GFUSA's name "progress index"). Thus, we sometimes forego indicators (such as education of the female head) that rarely change (even if poverty changes) in favor of indicators that do tend to change with poverty (such as possession of a fridge or color TV).

This "practicality" focus leads to one-page scorecards computable by front-line workers in the field as they collect the data (Figures 1, 2, and 3). This is accomplished via:

- Few indicators (10, not 15 plus other "control" variables)
- Categorical indicators such as "type of roof" (not numerical such as "total assets")
- User-friendly weights (all positive integers, no negative or decimal weights, and no need to take logarithms or exponents)
- Simple scores ranging from 0 (most likely poor) to 100 (least likely poor)

Among other things, this simplicity enables "rapid targeting", such as determining (in a day) who in a village is eligible for microfinance services, work-for-food, etc.

IRIS's scorecards may include these features, but they have not been reported yet.

Figure 1: Philippines poverty scorecard (source: Schreiner, 2005a)

Indicator			Values		Points
1. Does the family own a gas stove or gas range?			No	Yes	
			0	13	
2. How many people in the family are aged 0 to 17?	≥5	3 or 4	1 or 2	Zero	
	0	6	15	26	
3. How many television sets does the family own?		Zero	1	≥2	
		0	9	20	
4. What are the house's outer walls made of?			Light (cogon, nipa, or sawali, bamboo, anahaw)	Strong (iron, aluminum, tile, concrete, brick, stone, wood, asbestos)	
			0	4	
5. Do any family members have salaried employment?			No	Yes	
			0	7	
6. How many radios does the family own?		Zero	1	≥2	
		0	3	12	
7. Does the family own a sala set?			No	Yes	
			0	8	
8. What is the house's roof made of?			Light (Salvaged, makeshift, cogon, nipa, or anahaw)	Strong (Galvanized iron, aluminum tile, concrete, brick, stone, or asbestos)	
			0	2	
9. What kind of toilet facility does the family have in the house?			None, open pit, closed pit, or other	Water sealed	
			0	3	
10. Do all children in the family of ages 6 to 11 go to school?		No	Yes	No children ages 6-11	
		0	2	4	

Source: Calculations based on the 2002 APIS by Microfinance Risk Management, L.L.C. Total:

Figure 2: Poverty scorecard for México (source: Schreiner, 2005b)

Pregunta				Respuesta		Puntos
1. ¿Qué combustible utiliza para cocinar?				Leña	Gas	
				0	13	
2. ¿Esta vivienda cuenta con una regadera?				No	Sí	
				0	6	
3. ¿De qué material es la mayor parte de los pisos de la vivienda?			Tierra	Cemento o firme	Madera, loseta o mosiaco	
			0	6	12	
4. En los últimos tres meses, ¿Compró calzado para una persona de 17 años o más?				No	Sí	
				0	4	
5. En los últimos tres meses, ¿Compró prendas de vestir para una persona de 17 años o más?				No	Sí	
				0	5	
6. ¿Cuántos miembros del hogar son de edades de 0 a 17 años?	Cuatro o más	Tres	Dos	Uno	Cero	
	0	11	14	22	29	
7. ¿Cuenta con teléfono fijo o teléfono celular?				No	Sí	
				0	8	
8. ¿Cuántos miembros del hogar son "obreros o empleados"?			Cero	Uno	Dos o más	
			0	4	9	
9. ¿Cuenta con automóvil, camioneta, etc.?				No	Sí	
				0	6	
10. ¿Cuenta con un horno de microondas?				No	Sí	
				0	8	
Fuente: Cálculos de Microfinance Risk Management con datos de ENIGH.					Total:	

Figure 3: Poverty scorecard for Bosnia-Herzegovina (source: Schreiner *et al.*, 2004)

Indicator	**Value**	**Weight**
1. Ownership of car	No	0
	Yes	12
2. Education level of female household head/partner/spouse	≤ Primary	0
	> Primary	4
3. Number of household members	6 or more	0
	5	8
	4	11
	3	19
	2	27
	1	34
4. Ownership of stereo CD player	No	0
	Yes	8
5. Location of residence	Rural or peri-urban	0
	Urban	6
6. Average times eats meat each week with main meal	Rarely (0-2)	0
	Sometimes (3-5)	8
	Often (6 or more)	20
7. Average times eats sweets each week with main meal	Rarely (0-2)	0
	Sometimes (3-5)	8
	Often (6 or more)	16
Minimum possible score (most-likely poor)		0
Maximum possible score (least-likely poor)		100

Small technical differences probably "don't make no nevermind"

In scoring, the well-known "flat max" phenomenon generally means that different scorecards constructed in different ways with different indicators and different weights all often end up with about the same "Total Accuracy". To paraphrase Wainer (1976), it may be that great efforts at technical refinements "don't make no nevermind".[1]

This is so because most (good) indicators are highly correlated with each other. For example, someone with a straw roof probably does not have indoor plumbing either. So if we know the roof is straw, we gain little—in terms of measuring poverty—from checking whether there is indoor plumbing. Likewise, if we know there is indoor plumbing, we gain little from checking whether the roof is straw.

The "flat max" shows up clearly in the scorecard comparisons in Schreiner *et al.* (2004), Schreiner (2005b), and IRIS (2005a). Going from 5 to 10 indicators improves "Total Accuracy" some, but after that decreasing returns set in very quickly.

The "flat max" also applies to different scorecard-construction approaches. In all 12 countries and for both single-step and two-step methods in IRIS (2005a), the best scorecard correctly classified only 2 or 3 more cases out of 100 than the worst scorecard.

With such small differences, I would suggest that different approaches be judged less on accuracy and more on other criteria, such as cost and simplicity/"practicality". My review of the scoring literature comes to the same general conclusion.

[1] For more on the "flat max", see Lovie and Lovie (1986); Kolesar and Showers (1985); Stillwell, Barron, and Edwards (1983); and Dawes (1979).

Logit is standard, in theory and practice

GFUSA and IRIS seek to classify people as "very poor" or "not very poor".[2] This is a "head count" poverty measure, concerned only with being above or below the poverty line, regardless of *distance* from the poverty line.

Before discussing IRIS' questions, it is helpful to review the precise distinctions between the different approaches to scoring poverty.

Logit

The standard approach to classification—both in the academic literature (Greene, 1993) and in scoring practice—is logistic regression.[3] Logit accounts for the yes/no, either/or nature of head-count poverty status.

The output of a Logit scorecard is a *probability* of being "very poor". A specific person has a likelihood of being "very poor" that is greater than 0 percent but less than 100 percent, even though true poverty status is always either "very poor" or "not very poor". If, however, 100 people each have a poverty likelihood of 75 percent, then about 75 should turn out to be "very poor". Like weather forecasts, poverty likelihoods may be right or wrong for a given person, but they should be right on average.

The share of all clients who are "very poor" is simply the average poverty likelihood of all clients. For example, if 1 client has a poverty likelihood of 20 percent, 1 has a poverty likelihood of 30 percent, and 1 has a poverty likelihood of 50 percent, then the share of clients who are "very poor" is 33.3 percent, as 33.3 = (20 + 30 + 50) / 3.

Progress out of poverty is measured as changes in average poverty likelihood. For example, Schreiner (2005c) used the scorecard in Figure 3 to find that, from one loan to the next for repeat borrowers at Prizma Mikro, average poverty likelihood fell by 0.7 percentage points (7 of 1,000 repeat clients crossed the poverty line).[4]

With Logit, clients close to the poverty line tend to have poverty likelihoods close to 50 percent. Thus, if measurement error causes true expenditure and measured expenditure fall on a different sides of the poverty line, Logit assigns a poverty likelihood of about 50 percent, and measurement error tends to average out over the portfolio. Furthermore, Logit guarantees that average estimated poverty likelihood in the sample used to construct the scorecard is the same as the true poverty rate in that sample.

---

[2] Actually, GFUSA seeks to classify them as "very poor", "poor", or "not poor", but we will stick here to two classes.

[3] Logit is essentially equivalent to the Probit and "linear probability" in IRIS (2005a).

[4] There is no implication that borrowing from Prizma *caused* the changes in poverty.

Truncated estimates from OLS/Quantile regression

OLS/quantile methods differ from Logit in that their output is not a *probability of being poor* (say, 75 percent) but rather a *level of expenditure* (say, $0.87/person/day). The estimate of expenditure is then converted to a poverty status by comparing it with the poverty line, with people below the line classified as "very poor".

Besides its intuitiveness, an advantage of the OLS/quantile approach is that it uses information (unlike Logit) about the distance between expenditure and the poverty line. Furthermore, it can be straightforwardly and intuitively extended to three classes.

A disadvantage is that it ignores estimated expenditure's sampling distribution. The point estimate of the level of expenditure is just the mean of an asymptotically Normal distribution that extends across both sides of the poverty line. The OLS/quantile approach looks just at the mean, not the whole distribution, as if the likelihood of being "very poor" were 100 percent simply because the mean of the distribution of estimated expenditure is below the poverty line. If true expenditure is not distributed symmetrically about the poverty line, then estimated poverty shares are inaccurate. This consequence follows directly from OLS/quantile's (unlike Logit) not being designed for classification (yes/no, either/or) problems.

Of course, the extent of the inaccuracy is an empirical question, so it may not matter much. I have not run tests on specific data, but I suspect that the error is small.

Another disadvantage of OLS/quantile is that—to my knowledge—is that it rarely appears in the literature and its properties are not documented.

Overall, either Logit or OLS/quantile could perform better in a given country and data set. (In any case, the "flat max" suggests that any differences are probably small).

Quantile and two-step methods can indeed improve classification of the poor

IRIS (2005a) compares OLS/quantile versus Logit/Probit/"linear probability" for 12 LSMS countries. By the "BPAC" criteria, the quantile approach is the most accurate.

Looking at "Total Accuracy" for single-step methods, Logit/Probit/"linear probability" is (barely) better in 10 of 11 countries (1 country had a tie). For two-step methods, Logit/Probit/"linear probability" is (barely) better in all 12 countries. The "flat max" makes this a dead heat in terms of "Total Accuracy", so I would look at other criteria.

An issue is that the "not very poor" outnumber the "very poor". In this case, "Total Accuracy" tends to tip the scorecard toward indicators that correctly classify "very rich", no-brainer cases but that are less accurate for more difficult cases.

For example, people with post-graduate degrees are highly unlikely to be "very poor". To measure their poverty, a whole scorecard is hardly necessary. But "has a post-graduate degree" is irrelevant for less obvious cases, as almost none of "not very rich" have a post-graduate degree. Thus, a 10-indicator scorecard effectively uses only 9 indicators to classify most people. This reduces accuracy among the "not very rich".

When IRIS (2005c) noted this issue and asked for suggestions, I pointed out that the scoring industry typically deals with "unbalanced sample proportions" by giving more weight to "very poor" cases so that they are no longer so outnumbered. Some ways to do this:

- Explicitly weigh cases by poverty status (Salford Systems, 2000)
- Introduce explicit costs of misclassification into the regression criterion
- Adjust indicator-selection criteria (such as MAXR) to value the "very poor" more
- Hand-select indicators relevant for borderline cases
- Choose a cut-off threshold to improve accuracy among the "very poor"
- Use a method (such as quantile) that places less importance on global accuracy and on local accuracy (such as among the "very poor")
- Use a two-step method which first identifies "no-brainer" cases with one set of indicators and then classifies the remaining, more difficult cases with a second set of indicators (Hand and Vinciotti, 2002; Shapire, 2001; Grootaert and Braithwaite, 1998; Myers and Forgy, 1963).

Compared to "Total Accuracy", "BPAC" places greater weight on correct classification of the "very poor". This seems quite reasonable to me (and I applaud the explicitness of the assumptions behind "BPAC" as spelled out in IRIS, 2005b). Of course, accuracy is not the only criterion, but it is important.

In the "BPAC" tests reported in IRIS (2005a), quantile regression is the most accurate, and two-step methods are sometimes (but not always, and even then not usually by a wide margin) more accurate than one-step methods.

This is because quantile places greater weight on classifying "very poor" cases. Furthermore, the exact quantile through which to run the regression was apparently chosen so as to minimize the absolute value of "PIE", an element of "BPAC".

In second place is Logit/Probit/"linear probability". Of course, a better comparison would select a cut-off threshold for the Probit poverty likelihoods to optimize "BPAC" (as was done for quantile). IRIS (2005a) does not report what cut-off was selected for the Probit tests, nor whether that cut-off optimized "BPAC".

Are there downsides to quantile and/or two-step methods? For users, they are more complex (but only if the method is two-step and if the computing is done on paper). Two-step methods may also require collecting more indicators, and they run more risk of overfitting (see below).

On the whole, my judgment is that (in part because of the "flat max") a simple one-step approach provides sufficient (if lower) accuracy and equal (or better) "practicality".

Among one-step methods, quantile appears best, followed by Logit/Probit/"linear probability". A full comparison, however, requires optimizing the Probit cut-off.

In this context, the remainder of this memo addresses IRIS' specific questions.

1. IRIS searched for econometric models that maximize accuracy and BPAC (and noted that 1-step logit and probit were almost always among the worst performing models). It is unclear how logit was selected for the GF scorecard as the econometric model that maximizes accuracy. Was there a search for a better model and if so what did it yield? Will a two-step method be considered?

Logit was selected because the problem is one of classification, and Logit is designed for that. Furthermore, Logit is the scoring-industry standard.

Logit was not selected to maximize accuracy (and at the time, "BPAC" did not exist). Instead, it was selected with an eye toward quick, inexpensive development of simple, easy-to-use scorecards that would gain the trust and acceptance of users, fit on a single piece of paper, and be computable on-the-spot by front-line workers.

The "flat max"—and evidence in IRIS (2005a), Schreiner (2005a), and Schreiner *et al.* (2004), as well as my own review of the scoring literature and experience with scoring for microfinance over the past 7 years—suggest that searching for an optimal method is unlikely to offer large pay-offs. I do not anticipate testing two-step scorecards, but it would be an interesting exercise.

2. As part of its selection of the best models, IRIS's method includes a sensitivity analysis of the weights, as the best possible econometric model automatically produces the best possible weights for the indicators selected by MaxR or ANOVA. How does the GF approach select the variables and weights included in scorecards (regression, expert opinion,…?). Is selection of variables done with C-statistic? Is there a sensitivity analysis on the weights?

Schreiner (2005a and 2005b) are memos describing poverty scorecards for managers of GFUSA affiliates. As such, they only briefly discuss technical issues of scorecard construction, mostly just noting the use of statistics and national expenditure surveys.

MAXR (or "stepwise") selects indicators one at a time, maximizing at each step a measure of accuracy known as "R^2", without input from the scorecard builder. Of course, maximizing R^2 is not the same as optimizing "BPAC". Thanks to the "flat max", however, the method will probably do well by reasonable criteria such as "BPAC".

Stepwise (often jokingly called "unwise") has well-known, relevant weaknesses:[5]

- Regular statistical tests are biased. In particular, R^2 is too high
- Stepwise scorecards are overfit (that is, they do not generalize well to new data)
- Overfitting is worse when:
    - Indicators are highly correlated (as with poverty indicators)
    - There are many indicators (as in LSMS/national survey data)

Overfitting is a classic scoring pitfall. It occurs because a scorecard is built from one set of data (say, an LSMS survey) and then applied to different data (say, a microlender's clients). Some of the apparent patterns in the construction data that are reflected in the scorecard are not true patterns at all but rather random sampling variation. These patterns are not repeated when the scorecard is applied, harming accuracy.

As a simple example of overfitting, suppose a scorecard predicts hair color based on color of clothing. While there are real correlations to be detected, if the construction sample were "the next ten people to walk in the door", pure chance could produce spurious "patterns" (such as brown hair with blue shirts). If the scorecard then predicts "if blue shirt, then brown hair", it will be inaccurate for the next people to walk in the door, as the random pattern is unlikely to be repeated.

Many indicators have about the same accuracy. In a given sample some—by chance—will be slightly more accurate and will be selected by stepwise. In a different sample, however, they would no longer be "best".

[5] See #12 on STAT-L FAQ, `http://www-personal.umich.edu/~dronis/statfaq.htm`.

Thus, automated indicator selection (MAXR/stepwise) tends to produce non-robust (but more or less equally accurate) scorecards. With different data, the set of "best" indicators and weights could change a lot. This is the definition of sensitivity.

Thus, rather than being a "sensitivity analysis", stepwise creates sensitivity.

The solution is to build scorecards based not solely on in-sample goodness-of-fit but also on theory, experience, and common sense. Why use "blue shirts" as an indicator of "brown hair" if there is not sensible reason to link the two? As stated by Ira Bernstein:[6] "Stepwise is no substitute for understanding the statistics, the data, and the domain. In general, because overfitting is a real issue, using theory and diagnostics to choose variables that are somehow 'non-optimal' on the current data can nonetheless produce models that generalize better (and . . . are easier to explain to lay people)."

This is exactly why we have not sought to maximize accuracy. Or more precisely, we try to maximize accuracy when scoring is used ("out-of-sample") by *not* trying to maximize accuracy during scorecard construction ("in-sample").[7] Indeed, even simple scorecards with 0/1 weights (such as Schreiner *et al.*, 2004 and Kolesar and Showers, 1985) based on experience can be as accurate out-of-sample (thanks to the "flat max" and the avoidance of overfitting) as fancy, data-based econometric models.

[6] Quoted in STAT-L FAQ. Bernstein is rather harsh on stepwise, saying it "allows us to not think about the problem", and "I don't know what knowledge we would lose if all papers using stepwise regression were to vanish".

[7] Accuracy in all scorecards discussed here is overstated to some unknown degree, as they have been tested on the same data that was used to construct them.

Of course, data usually contain information lacking in theory or experience; the trick is to separate real patterns from random ones. My scorecard-construction technique melds the two approaches:

- Using the "c" criterion (area under a Receiver Operator Characteristic curve, the standard way to measure scoring accuracy), select 100–200 candidate indicators from a pool of 500–1000
- For each indicator, create a one-indicator scorecard, ranking them by "c".
- Select one of the top 1–5 indicators based on:
    - "Face validity" (theory, experience, and common sense)
    - Likelihood of changing over time as poverty status changes
    - Verifiability, and susceptibility to strategic falsification
    - Cost to collect
    - Simplicity
    - Accuracy
    - Other criteria, such as similarity to other indicators already in the scorecard
- Add the selected indicator to the scorecard, and repeat the last three steps

This amounts to a "MAXC" stepwise process, with human judgment used to reduce overfitting and increase "face validity" for users.

For the Philippines (Schreiner, 2005b), GFUSA-affiliate managers reviewed the 10-indicator scorecard and asked—for various reasons—to replace five indicators with others, four of which appeared in the 15-indicator scorecard. Making the changes encouraged acceptance at hardly any cost—thanks to the "flat max"—to accuracy.

3. Because of its mandate and its concern about measurement errors, the IRIS/USAID approach explored econometric models that equalize undercoverage and leakage (quantile), therefore minimizing the concern about errors at the sample level. Since it calculates individual poverty rates, GF is supposedly taking these two errors into account. How does it do this?

The "MAXC" stepwise process (adjusted by human judgment) uses the "c" criterion to summarize accuracy. Higher "c" implies less total undercoverage and leakage. "c" (or functions of it) is a standard measure of accuracy in the scoring industry.

Logit scorecards produce estimates of poverty likelihood. If, for operational reasons, users want to classify clients as "very poor" or "not very poor", then they must select a cut-off above which clients are labeled "not very poor" and below which they are labeled as "very poor". Cut-offs can be set—if desired—to equalize undercoverage and leakage.

More generally, labeling clients based on their poverty likelihood in relation to a given cut-off leads to four possible outcomes:

- Successes:
  - "True very poor": Labeled as "very poor", and truly "very poor"
  - "True not very poor": Labeled as "not very poor", and truly "not very poor"
- Mistakes:
  - "False very poor": Labeled as "very poor", but truly "not very poor"
  - "False not very poor": Labeled as "not very poor", but truly "very poor"

	Labeled "very poor"	Labeled "not very poor"
Truly "very poor"	True very poor	False not very poor
Truly "not very poor"	False very poor	True not very poor

The standard way in the scoring industry to account for errors is to assign a cost to each outcome, and then build a scorecard (and/or choose a cut-off) to minimize cost. For example, the implicit cost matrix for the "Total Accuracy" criterion is:

	Labeled "very poor"	Labeled "not very poor"
Truly "very poor"	0	1
Truly "not very poor"	1	0

We ask GFUSA affiliates to determine the cost matrix appropriate for their goals, and we then inform them of the optimal cut-offs.

4. GF does not mention BPAC, but it can be computed from their total and poverty accuracy numbers. It would be interesting to compare with the IRIS/USAID BPAC numbers when GF completes the Bangladesh scorecard.

Yes.

The scorecards for Mexico and the Philippines were built before IRIS developed BPAC.

The ideal test of pure accuracy (ignoring "practicality") would work as follows:

- IRIS and GFUSA receive data on some share (say, two-thirds) of households surveyed for a country
- One-third of all households are held-out for a later "out-of-sample" test of accuracy
- Scorecards are built based on the two-thirds of cases
- Scorecards are used to classify the (now revealed) one-third of held-out cases
- Accuracy is compared on various criteria

5. On the tool development side, there are two possible sources of differences in cost between the GF and IRIS/USAID approaches: 1) GF and IRIS rely on HH surveys (LSMS, SDA, country-level). If these are not available or not reliable, both approaches need to collect new household data, at the same high cost. 2) IRIS tested different econometric models to maximize accuracy; if GF's choice of the logit method came from a similar search for the best performing method, the two approaches have similar costs here too. On the tool implementation side, the per capita cost of applying the GF and IRIS/USAID tools in the field is probably similar since they include roughly the same number of questions. One difference may be that GF doesn't need to enter data to compute overall poverty rates because the assessment is done on an individual level, whereas the calculation of collective poverty rates in the IRIS/USAID requires data entry and computerized calculation of poverty rates over the sample. Is GF referring to development or implementation costs when suggesting that the cost of its scorecard is lower than that of the IRIS/USAID tool?

As the question points out, there is in principle no great difference in costs in terms of:

- Acquiring data to build scorecards
- Searching for approaches that maximize accuracy
- Implementation (either approach can, in principle, be automated or put on a single piece of paper for on-the-spot computation)

In a previous comment in a message to Nigel Biggar and Laura Foose, I said our approach was "probably much less expensive" than IRIS'. I suspect that this comment prompted the current question.

The comment refers not to development costs nor implementation costs but rather project costs. I may be mistaken, but I imagine that IRIS has a grant of several hundred thousand dollars and 2–3 years to develop its (as-yet unreported) scorecards. GFUSA developed its two scorecards (and documented them, and started to use them in the field) in about six months for less than $20,000. While IRIS' project has a wider scope and must deal with greater procedural complexities, it is seems reasonable to think about benefits and costs.

6. *GF is planning to develop scorecards for India, Pakistan, Bangladesh, Haiti, Bolivia, Morocco and Egypt. According to the information about these countries reviewed by IRIS, recent (post 2000) LSMS data is only available in two Indian states (UP and Bihar), and these are included in the IRIS/USAID analysis. The most recent LSMSs in Morocco and Pakistan date from before 1995 (and are hence probably unusable). It appears that no LSMS or SDA data is available in the other four countries (and the rest of India), so has GF gained access to reliable and complete household survey data for these areas?*

This is a question for Nigel.

References

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

Greene, William H. (1993) *Econometric Analysis: Second Edition*, New York, NY: MacMillan, ISBN 0–02–346391–0.

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, Washington, D.C., `http://www.worldbank.org/html/dec/Publications/Workpapers/WPS1900series/wps1942/wps1942.pdf`.

Hand, David J.; and Veronica Vinciotti. (2002) "Scorecard Construction with Unbalanced Class Sizes", Department of Mathematics, Imperial College.

IRIS. (2005a) "Accuracy Results for 12 Poverty Assessment Tool Countries", `http://www.povertytools.org/documents/Accuracy%20Results%20for%2012%20Countries.pdf`.

_____. (2005b) "Note on Assessment and Improvement of Tool Accuracy", `http://www.povertytools.org/documents/Assessing%20and%20Improving%20Accuracy.pdf`.

_____. (2005c) "Issues in Assessing the Accuracy of Poverty Assessment Tools", `http://www.povertytools.org/documents/Listserv%20Question%206.pdf`.

Kolesar, Peter; and Janet L. Showers. (1985) "A Robust Credit Screening Model Using Categorical Data", *Management Science*, Vol. 31, No. 2, pp. 123–133.

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

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.

Salford Systems. (2000) *CART for Windows User's Guide*.

Schapire, Robert E. (2001) "The Boosting Approach to Machine Learning: An Overview", AT&T Labs, `http://research.att.com/~shapire`.

Schreiner, Mark. (2005a) "Additional analysis of poverty scorecard for Philippines", memo for Grameen Foundation U.S.A.

_____. (2005b) "Ficha de puntaje con tres clases de pobreza, áreas rurales de México", memo for Grameen Foundation U.S.A.

_____. (2005c) "Initial analysis of poverty scores for Prizma Mikro", memo for Prizma Mikro.

_____. (2002) "Scoring: The Next Breakthrough in Microfinance?" Consultative Group to Assist the Poorest Occasional Paper No. 7, Washington, D.C., `http://www.cgap.org/docs/OccasionalPaper_07.pdf`.

_____; Matul, Michal; Pawlak, Ewa; and Sean Kline. (2004) "The Power of Prizma's Poverty Scorecard: Lessons for Microfinance", manuscript, Microfinance Risk Management, `http://www.microfinance.com`.

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.

Wainer, Howard. (1976) "Estimating Coefficients in Linear Models: It Don't Make No Nevermind", *Psychological Bulletin*, Vol. 83, pp. 213–217.