RICHARD FREEMAN
DAMIAN RAESS
ISABELLE FERRERAS

Statistical Tests for the WageIndicator Web-Survey: A Preliminary Skirmish

Surveys that diverge from random samples of specified populations have become increasingly important in social science and business research. On the positive side, the Internet offers the possibility of obtaining information inexpensively from large samples. On the negative side, it has become increasingly difficult for standard surveys to obtain the sample response rates that they had obtained in the past. Historically, researchers have always had difficulty obtaining good response rates, so that many surveys and analyses are based on non-representative samples – in many cases, case studies.

This essay addresses the issue of data quality in web-surveys where the sample of web-respondents is non-random (Bethlehem 2008; Couper 2000). It focuses on the problem of generalizability from non-representative web-survey samples by discussing the example of the international WageIndicator web-survey – a large cross-country non-representative international survey (see Tijdens 2005)[1]. Specifically, it describes and explains the statistical tests we have conducted on the US web-based Paywizard[2] data. It also presents a plan for comparable tests in other WageIndicator countries that we think are a necessary first step to systematically assess the data reliability of both the national and international data set generated by the WageIndicator project.

To examine the possible bias in the data produced through the Paywizard (PW) survey, we do run a number of tests on data from the Paywizard and compare the results with those generated for a reference survey based on a representative sample (for similar strategies, see for example Faas and Schoen 2006; Fricker et al. 2005). The first exercise is thus to identify the reference survey for a national WageIndicator data set. In the US, for example, it is the Current Population Survey (CPS).

The proposed statistical strategy to probe the generalizability of the Paywizard data applies a three-staged series of tests to the classical labor economics model of the determinants of income. Our baseline model uses the basic socio-demographic variables gender, age and education as explanatory variables. Subsequent specifications may include employment characteristics such as employment contract (e.g. permanent vs. temporary; full- vs. part-time), industry, occupation, etc. In what follows we shall explicate the three families of tests. For illustration purposes, we present basic results for the United States (US).

1) Summary Statistics of Variables

The first test is to compare the distribution of the variables across the two data sets. To facilitate interpretation, a summary table of the results containing the means and standard deviations of the variables is desirable, though one should also look at the entire distribution to see if higher moments appear similar or dissimilar.

The rationale for this test is to examine whether the Paywizard is biased in terms of its representation of gender-, age-, education- and income-groups. Weighting can deal with bias in the means, but more sophisticated analysis is necessary if the distributions differ in other ways.

1 The WageIndicator is an international, continuous web-based survey about wages and employment issues. It was launched in The Netherlands in 2001. It has been growing rapidly, with currently more than 40 countries participating across 4 continents (Europe, Asia, Africa, North and South America). For more information, see www.wageindicator.org.

2 Paywizard is the specific name of the US version of the WageIndicator survey (see www.paywizard.org).

Our experience says that:

Income, a continuous variable (e.g. weekly gross wage), might be top-coded in the representative dataset, in which case the same top-coding rule should be applied to the national WageIndicator dataset.

For comparison of income, use the logarithmic function of income (logincome), rather than income, because income is a nonlinear function with independent variables and has a lognormal shape in most situations. Gender is a dichotomous variable, with male=0 and female=1. The metric variable age is censored to restrict respondents to the working age population, for instance defined as the 18 to 65 age group. Finally, education is a numeric variable with each category measuring the number of years of schooling corresponding to the highest grade of school completed (e.g. 14 categories ranging from group '1st-4th grade' or 2.5 years of schooling to group 'Doctorate' or 22 years).

For illustration purposes, we run the tests on a simplified, baseline model of the determinants of income, comparing sample characteristics between the CPS and the PW. Our simple model uses binary measures for three independent variables, as follows: gender ('male'=0; 'female'=1), age group ('16-34'=0: '35-75'=1), and education (1st grade thru high school graduation, 'low-medium'=0; some college and above, 'medium-high'=1). Due to top-coding of the earnings variable in the CPS, we restrict our dependent variable, hourly gross wage, to the values ranging from US$1.5 to US$150. We use 2007 data for both the CPS and the PW. The sample size is 158'745 for the CPS, 669 for the PW.[3]

Our main results are that:

Women, younger people, and skilled employees are over-represented in the Paywizard sample (Table 1). Highly-educated people, in particular, are strongly over-represented: nearly 90% of respondents have attended 'some college' in the PW vs. 60% in the CPS.

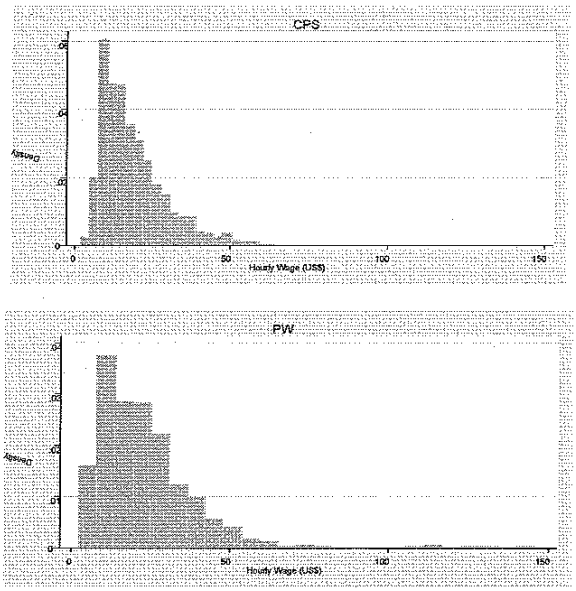
Average wage is about 30% higher in PW (US$ 24.2) compared to the CPS (US$ 18.7). This is partly explained by the higher share of high-educated workers in PW. Note also the greater incidence of very high pay (hourly wage between US$100-150) in the PW (Figure 1).

Table 1. Distribution over gender, age, and education in the CPS and PW, 2007, in %

	gender		age group		education	
	men	women	16-34	35-75	low-med	med-high
CPS	50.1	49.9	34.7	65.3	39.8	60.2
PW	47.5	52.5	41.7	58.3	12.3	87.7

3 We exclude from the samples the self-employed, the unemployed, the disabled, and students in full time education who work less than 16 hours per week.

Figure 1. Distribution over wage in the CPS and PW, 2007, in US$



2) Some Analysis of Correlations among Variables

The second step is to compare the correlation matrixes among variables of the model between the two data sets. We use the Pearson correlations coefficient for all variables.

We stress the correlations between variables for two reasons. First, even though the mean of a variable can be biased because of the non randomness of the sample, its product-moment correlation coefficient to a dependent variable could be the same (or not so much far off) as to the one of the representative sample. For example, if every person under the age of 12 screamed at a horror movie the correlation between age < 12 and screaming would be the same in a sample that had a biased age distribution as in a representative sample. Second, focusing on correlations is relevant because one of the crucial virtues of a non-representative survey is to find out patterns or relationships in the data that reflect behavior – much as the purpose of experimental psychology is to see how volunteers respond to situation, not to use them to make a statement about the population.

Our correlational analysis concludes that:

The associations involving income (logwage) are properly signed - income rises with age, with education, and, as expected women earn less than men (although the association is not statistically significant). Note that the size of the correlation coefficient between education and income is considerably smaller in the PW.

Women tend to be less educated than men in PW (though not significantly so) but more educated in the CPS.

Table 2A. Correlation matrix, CPS

	women	logwage	age group	education
women	1			
logwage	-0.16***	1		
age group	0.02***	0.26***	1	
education	0.08***	0.33***	0.03***	1

Note: ***significant at 1% or lower; ** significant at 5%; * significant at 10%

Table 2B. Correlation matrix, PW

	women	logwage	age group	education
women	1			
logwage	-0.06	1		
age group	-0.02	0.24***	1	
education	-0.03	0.09**	0.05	1

Note: ***significant at 1% or lower; ** significant at 5%; * significant at 10%

3) Analysis of Coefficients of Regression

The third test consists of comparing regression coefficients of the determinants of income with gender, age and education as regressors between the Paywizard and the Current Population Survey (i.e. between a national WageIndicator data set and its reference survey, a representative data set).

Logincome is a continuous variable, therefore we use OLS to determine coefficients. Because of the nonlinear functional form, we add the variable age^2 (age-square) on the right-hand-side, alongside age, gender and education. We compare unweighted, weighted, and quantile regressions. Regarding weighting, we recommend to weight the data from the Paywizard sample to render it comparable to the representative data. The data is weighted by education using all education-groups available. Further analyses may consider more fine-grained weighting, not just by education, but also by gender, age groups and occupation or industry (see de Pedraza et al. 2007).

Our regression results, summarized in Table 3, show that:[4]

Compared to the baseline results from the CPS, the unweighted regression coefficients in the PW are correctly signed but biased and often lacking statistical significance. This was expected because participation in the PW web-survey is volunteer (non-random).

Weighting does help in reducing the bias, although it does not eliminate the bias.

Median quantile regressions appear to narrow the bias even further than simple weighting -- suggesting that the weighted and unweighted results in PW might be driven by extreme values in the PW sample.

Table 3. Determinants of income (Dependent variable: log hourly gross wage)

	CPS unweighted	PW unweighted	PW weighted[5]	CPS quantile	PW quantile
women	-0.208***	-0.076	-0.113	-0.216***	-0.191**
	(0.002)	(0.051)	(0.069)	(0.003)	(0.079)
age	0.066***	0.084***	0.045**	0.070***	0.086***
	(0.001)	(0.016)	(0.019)	(0.001)	(0.024)
agesq	-0.001***	-0.001***	-0.000*	-0.001***	-0.001***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
education	0.364***	0.123	0.196**	0.374***	0.241**
	(0.003)	(0.078)	(0.084)	(0.003)	(0.120)
constant	1.205***	1.063***	1.788***	1.114***	0.921**
	(0.011)	(0.297)	(0.346)	(0.012)	(0.457)

4) Further steps in the analysis

The key value of the non-representative web survey is that it obtains information on variables not necessarily covered by any standard data set. In this context, the key methodological question is, if the web survey gives similar results to the reference survey (with random sample) on income (or on any variables present in both surveys), can one conclude that the results from the non-representative sample would generalize to the rest of the population for the variables that exist only through the web survey?

If the web survey gives dissimilar results on income (or on other variables present in both surveys) to the results in a reference data set, there should be ways to adjust the moment matrix. Yet this is very new and we are not clear on exactly which tests to conduct to render this possible and sound.

These are far-reaching statistical questions that we hope to address during an upcoming conference.

4 Note that here we use the continuous variable age, not the dichotomous variable age group

Concluding remarks

Web-surveys run the risk of bias due to the non-randomness of the sample. It is therefore essential to develop statistical analysis to test the generalizability of non-representative sample. This memo has described a set of statistical tests aimed at assessing the data quality of the web-based WageIndicator survey. Basic illustration of the tests on US data show significant bias in the non-representative Paywizard data, while techniques such as weighting and median quantile regression analysis (or any estimation that gives less weight to the extreme values) appear promising venues for reducing the selection bias. We hope that this and more elaborated tests yet to be defined will form the basis of a methods template that will be systematically applied by every national WageIndicator team on their data in order to generate a strong data set on wage and employment issues across the globe.

References

Bethlehem, J. (2008) 'How accurate are self-selection web surveys', Discussion Paper 08014, Statistics Netherlands, The Hague

Couper, M.P. (2000) 'Web Surveys: A Review of Issues and Approaches', *Public Opinion Quarterly*, Vol. 64, 4, 464-494.

de Pedraza, P., Tijdens K. and R. Muñoz de Bustillo (2007) 'Sample bias, weights and efficiency of weights in a continuous web voluntary survey,' AIAS working paper 60, Amsterdam: University of Amsterdam.

Faas, T. and H. Schoen (2006) 'Putting a Questionnaire on the Web is not Enough: A Comparison of Online and Offline Surveys Conducted in the Context of the German Federal Election 2002', *Journal of Official Statistics*, Vol. 22, 2, 177-190

Fricker S., Galesic M., Tourangeau R., and T. Yan (2005) 'An Experimental Comparison of Web and Telephone Surveys', *Public Opinion Quarterly*, Vol. 69, 3, 370-392.

Tijdens, K. (2005) 'Introduction to the *WageIndicator* dataset', available at: http://www.wageindicator.org/main/publications/publications2005

Richard B. Freeman holds the Herbert Ascherman Chair in Economics at Harvard University. He is currently serving as Faculty Director of the Labor and Worklife Program at the Harvard Law School. He directs the National Bureau of Economic Research / Sloan Science Engineering Workforce Projects, and is Senior Research Fellow in Labour Markets at the London School of Economics' Centre for Economic Performance. His recent publications include *Can Labor Standards Improve Under Globalization* (2004), *Emerging Labor Market Institutions for the 21st Century* (2005), *America Works: The Exceptional Labor Market* (2007), and *What Workers Say: Employee Voice in the Anglo American World* (2007). His forthcoming IZA Prize book is *Making Europe Work*. IZA Labor Economics Series (2009). Forthcoming co-edited books are *Reforming the Welfare State: Recovery and Beyond in Sweden* (2009); *Shared Capitalism: The Economic Issues* (2009); *International Comparison of the Structure of Wages* (2009), and *Science and Engineering Careers in the United States* (2009). Email: freeman@nber.org

Damian Raess received his PhD in Social Sciences from the University of Amsterdam in 2006. He specializes on the international and comparative political economy of labor, industrial relations and globalization. He was a visiting scholar in the department of political science at MIT (2007-2008), and is currently a visiting scholar in the Labor and Worklife Porgram, Harvard Law School, where he is Data Coordinator of the PayWizard project. His new project, generously funded by the Swiss National Science Foundation, is entitled *Labor in the Global Economy: Workers, Worker Representatives and Employers' Responses to Economic Globalization*. His articles have appeared in the *European Journal of Industrial Relations*, *Politics & Society*, and the *Review of International Political Economy*. Email: draess@law.harvard.edu

Isabelle Ferreras is a sociologist and political scientist. She is a tenured fellow of the Belgian National Science Foundation, an associate professor in the department of Social and Political Sciences, Université catholique de Louvain, and a senior research associate in the Labor and Worklife Program, Harvard Law School, where she is a member of the PayWizard project team directed by Prof. Richard Freeman. Research interests include economic sociology, sociology of work, industrial relations, political philosophy and democratic theory. Her recent book is *Critique politique du travail. Travailler à l'heure de la société des services* (Paris: Presses de Science Po, 2007). Email: isabelle.ferreras@uclouvain.be