APPENDIX 1

METHODOLOGY

Variable of interest

The results are presented as literacy rate of people 15 years of age and over.

Factors

Literacy rates were calculated for each province and classified by four determinant factors: gender, area of residence, age and ethnicity. The categories of the four factors are:

gender	male, female
area of residence	urban, rural
age	young—from 15 years to 34 years, old—35 years and older
ethnicity	Kinh, non Kinh

Age 35 was chosen as a break point for age because the most recent literacy campaign has targeted the 15 to 34 age group[1].

Ethnicity was classified as Kinh or non- Kinh because Kinh is the major ethnic group, 86.3% of the population in Vietnam (Population census 1999) The remaining 13.5% included 50 different ethnic groups.

Assessing the impact of the determinant factors within provinces using odds ratios

The impact of each factor was assessed using odds ratios both raw and adjusted. The odds ratio is a measure of association between an outcome variable and a determinant factor, where the literacy data can be structured, as a two-by-two table of counts for each determinant with literacy as outcome: for example with gender as the determinant, and with observed counts as shown below:

		Determinant	
		male	female
Outcome	literate	a	b
	illiterate	c	d
	odds	a/c	b/d

The odds ratio is the ratio of the odds that men are literate, a/c, divided by the odds that women are literate, b/d. Therefore, the odds ratio for gender where the outcome variable is literacy is

$$oddsratio = \frac{a}{c} : \frac{b}{d} = \frac{ad}{bc}$$

In addition, the odds ratios were adjusted for the other likely confounding factors. The likely confounding variables were the other factors from the list: for example, for gender, the confounding factors are area of residence, age and ethnicity.

[1] Pham Minh Hac et al, *Education for All*, 2000

To calculate adjusted odds ratio the Mantel Haenszel estimate of the adjusted odds ratios were used with the following formula (Mantel, Haenszel 1959; McNeil 1996)

$$OR_{MH} = \frac{\sum a_g d_g / n_g}{\sum b_g c_g / n_g}$$ where g specifies the stratum, n_g the number in the stratum.

A confidence interval for the adjusted odds ratio may be obtained by using a formula given by Robins et al [1986] as below:

$$SE\left[\ln\left(OR_{MH}\right)\right] = \sqrt{\frac{\sum P_g R_g}{2R_+^2} + \frac{\sum\left(P_g S_g + Q_g R_g\right)}{2R_+ S_+} + \frac{\sum Q_g S_g}{2S_+^2}}$$

where $R_+ = \sum R_g$, $S_+ = \sum S_g$, $P_g = \frac{a_g + d_g}{n_g}$,

$Q_g = \frac{b_g + c_g}{n_g}$, $R_g = \frac{a_g d_g}{n_g}$ and $S_g = \frac{b_g c_g}{n_g}$.

Finally the 95% confidence intervals are calculated as;

$$95\%\ CI = OR_{MH} * \exp(\pm 1.96 * SE[\ln(OR_{MH})])$$

We can calculate the adjusted odds ratio for each of the 61 provinces. The method is illustrated for Ha Noi. The following table shows the sample data for Ha Noi.

area		Urban							
ethnicity		Kinh				Non Kinh			
age		young		old		young		old	
gender		male	female	male	female	male	female	male	female
Ha Noi	literate	3672	3750	3999	4147	19	38	21	20
	illiterate	16	9	22	210	0	0	0	1
	Total	3688	3759	4021	4357	19	38	21	21
area		Rural							
ethnicity		Kinh				Non Kinh			
age		young		old		young		old	
gender		male	female	male	female	male	female	male	female
Ha Noi	literate	3733	3586	3051	3116	7	7	6	4
	illiterate	36	25	78	620	0	0	0	0
	Total	3769	3611	3129	3736	7	7	6	4

We will calculate the odds ratio for literacy versus gender adjusted for area, ethnicity and age. There are 8 stratum defined by area, ethnicity and age.

From the data we obtain the following

area	urban				rural				
Kinh	Kinh		Non Kinh		Kinh		Non Kinh		
age	young	old	young	old	young	old	young	old	total
n	7447	8378	57	42	7380	6865	14	10	30193
P	0.49	0.50	0.33	0.52	0.51	0.53	0.50	0.60	4.00
Q	0.51	0.50	0.67	0.48	0.49	0.47	0.50	0.40	4.00
R	4.44	100.24	0.00	0.50	12.65	275.55	0.00	0.00	393.37
S	8.06	10.89	0.00	0.00	17.49	35.40	0.00	0.00	71.84
P*R	2.19	50.36	0.00	0.26	6.44	147.35	0.00	0.00	206.60
P*S	3.98	5.47	0.00	0.00	8.91	18.93	0.00	0.00	37.29
Q*R	2.24	49.88	0.00	0.24	6.21	128.20	0.00	0.00	186.77
Q*S	4.07	5.42	0.00	0.00	8.59	16.47	0.00	0.00	34.55

And based on the above data, the gender and urban/rural adjusted odds ratio is calculated as;

$$OR_{MH} = \frac{\sum a_g d_g / n_g}{\sum b_g c_g / n_g} = = \frac{393.37}{71.84}$$
$$= 5.48$$

Then SE(ln(OR_{MH})) = 0.0893. Finally the 95% confidence intervals for Ha Noi are calculated by;

95% CI = OR_{MH}*exp(±1.96* SE[ln(OR_{MH})])

= 5.48 * exp(±1.96* 0.0.0893)

= (4.60, 6.52)

Assessing the impact of the determinant factors within provinces using logistic regression

An approximation of the adjusted odds ratio can be obtained using logistic regression modeling. The coefficient of the parameter for the factor in a logistic regression model using only the main effects of the factors is almost equal to the log of the adjusted odds ratio. We illustrate using the Ha Noi data:

	Logistic estimates	Exp(est)	Mantel Hainzsel	Log - MH
beta	1.706	5.507	5.475	0.032
95% lower limit	1.530	4.618	4.600	0.018
95% upper limit	1.882	6.567	6.523	0.044

The logistic regression model gives estimates of the effects of gender on the log odds ratio. Hence to obtain estimates for the adjusted odd ratio, we need to calculate exp(estimate). The table shows that the logistic regression estimates are slightly greater than the Mantel Haenzsel estimates, but by less than half a percent of the MH value. These estimates were obtained using the PLUM algorithm as implemented in SPSS 11.5.

We used more complex logistic regression models which included second order interaction terms as well. These were fitted for each province separately. In the case of provinces with small (less than 5%) ethnic minority populations, the factor Kinh was omitted. In each province, the dominant or terms of greatest statistical significance involved Age. The final modeling

involved the young adult population with factors Gender, Area of residence and Kinh in the case of provinces with large ethnic minority populations. Parsimonious models were obtained in which each term of the second order was statistically significant, or if no second order terms were significant, the main effects terms were statistically significant. Because of the very large numbers To be of statistical significance, the p-value had to be less than 0.01. These models were used to justify the selection of subgroups with low literacy levels requiring special attention.

Maps

We have used MapInfo Version 7 to create the range thematic maps with five different colors.

References:

Kleinbaum, B.G. & Klein, M. 2002 Logistic Regression-A Self Learning Text, 2nd edition Springer-Verlag

Mantel, N. & Haenszel, W. 1959, "Statistical aspects of the analysis of data from retrospective studies of disease, *Journal of the National cancer Institute*, 22, pages 719-748.

McNeil, D. 1996, Epidemiological Research Methods, John Wiley & Sons, page 105.

Robins J.M., Breslow N.E., Greenland S., 1986, "Estimators of the Mantel-Haenszel variance consistent in both sparse data and large-strata limiting models", *Biometrics*, 92: pages 311-323.