Robinson's Ecological Correlations and the Behavior of Individuals: methodological corrections

Article *in* International Journal of Epidemiology · May 2011

DOI: 10.1093/ije/dyr081 · Source: PubMed

Manfred Grotenhuis
Radboud University
88 PUBLICATIONS **4,571** CITATIONS
SEE PROFILE

Rob Eisinga
Radboud University
151 PUBLICATIONS **5,142** CITATIONS
SEE PROFILE

Title A correction to Robinson's *Ecological Correlations and the Behavior of Individuals*

Author, Degrees and Affiliation

Manfred te Grotenhuis, PhD, Department of Sociology, Radboud University Nijmegen, The Netherlands.

Rob Eisinga, PhD, Department of Sociology, Radboud University Nijmegen, The Netherlands.

S V Subramanian, PhD, Department of Society, Human Development and Health, Harvard School of Public Health

Corresponding Authors

Manfred te Grotenhuis, Department of Sociology, Faculty of Social Sciences Radboud University Nijmegen PO Box 9104, 6500 HE Nijmegen, The Netherlands. E-mail: m.tegrotenhuis@maw.ru.nl

S V Subramanian, PhD, Department of Society, Human Development and Health, Harvard School of Public Health 677 Huntington Avenue, Boston MA 02115, USA. Email: svsubram@hsph.harvard.edu

Sixty years ago the late William S. Robinson (1913-1996) published his *Ecological Correlations and the Behaviors of Individuals*.(**1**) The paper became an all-time classic and it is one of the most influential methodological papers in social sciences, including epidemiology. To underscore its impact on epidemiology, this classic was reprinted in this journal,(**2**) along with an original re-analysis of Robinson's data and conclusion from a multilevel and historical perspective,(**3**) with discussions.(**4-7**) In this letter, we identify, and correct, an error in Robinson's original analysis.

Robinson used data from a U.S. Census Bureau 1933 publication, Table 10 (pp. 1229).(**8**) We coded these data from the original paper records and have made them available at http://www.ru.nl/mt/rob/downloads. Robinson opened his paper with a scatter diagram (his Figure 1) in which the 'percent illiterate' is plotted against the 'percent negro', using observations of the U.S. Census Bureau's nine geographic divisions in 1930. The illiteracy-race relationship was positive and the 'ecological' Pearson correlation was 0.946. This figure is a weighted correlation coefficient, using the number of individuals in each division as weights. The data were taken from the *row margins* of a 2 (race) by 2 (illiteracy) by 9 (division) table (his Table 2). Robinson then aggregated this table into a 2 (race) by 2 (illiteracy) table (his Table 1) and used the counts in the *interior cells* to calculate an individual correlation of 0.203. Here we have the well-known illustration of a potential ecological fallacy as the ecological (division-level) correlation is substantially different than the individual correlation.

Somewhat less known, Robinson also demonstrated that the weighted ecological correlation equals the weighted difference between the overall individual correlation and the average of all within-division individual correlations (Equation 1, pp. 340).(**2**) This relation between ecological and individual correlation holds as long as the weighted ecological correlation is calculated from the total margins of the underlying individual data table. This is important to note because when Robinson analyzed the data at state-level, this mathematical relationship did not hold any more. There are two reasons for this remarkable fact. First, Robinson unwarily used state-level data which were *not* the result of the underlying individual data table. Second, for reason not known to us, Robinson used *unweighted* ecological correlations. Robinson's Figure 1, Table 1, and Table 2 are based on a (rounded) total of 97,272,000. However, the 1930 census reports a total population of 98,723,047 U.S. citizens (pp. 1219).(**8**) Upon scrutinizing U.S. Bureau of the Census,(**8-10**) we discovered that Robinson excluded 1,449,824 non-whites, e.g., Mexicans, Indians, and Chinese. Thus, Robinson's Figure 1 relates to the percent black and the percent illiterate amongst black and white people, i.e., without 1.45 million non-whites. Among these discarded non-whites, 362,643 of them were not able to read and write, either in English or in other languages; hence 25% of all non-whites was illiterate. That is far more than the percent illiterate among whites (2.7), foreign-born (9.9), and blacks (16.3). Furthermore, the non-whites were not distributed at random across the nine census divisions; instead, they were highly concentrated in the 'Mountain' division and more specifically in New Mexico and Arizona, see Table 1.

Table 1 Total illiteracy and black and white illiteracy in New Mexico and Arizona, 1930

State	Total	Total Illiterates	% Illiterates	Total Black & Whites	Total Black & White Illiterates	% Black & White Illiterates
New Mexico	314,370	41,845	13.3	252,718	19,403	7.7
Arizona	335,029	33,969	10.1	222,392	1,877	0.8

Source: U.S. Census Bureau, Chapter 13, Table 10 (pp. 1229) (**8**)

Table 1 shows that in New Mexico, and especially in Arizona, the percent illiterate is very different depending on whether non-whites are included or not. In Arizona the total number of illiterates was 33,969 (10.1% of the population) whereas the total number of white and black illiterates was only 1,877 (0.8% of the black & white population). We must note that the U.S. Census Bureau Table 10 (**10**) only reports the percent illiterate in the total population for every state and includes no warning that the subtotals for blacks and whites for each state do not sum up to the grand total. Other U.S. Census Bureau tables and publications however does include such a warning, see for instance, U.S. Bureau of the Census (pp. 35),(**8**) and U.S. Bureau of the Census (pp. 2).(**9**) This mismatch between grand total and underlying subtotals had no consequences for all division-level analyses since Robinson aggregated these from the correct underlying individual level table. However, for the state-level analysis, he used the U.S. Census Bureau Table 10 total columns that *included* the non-whites while the underlying individual table still excluded non-whites.

Robinson argued that he used divisions only for simplification (pp. 338).(**2**) He then continued with his Figure 2 that gives the relationship between percent illiterate and percent blacks at state-level. To construct this figure Robinson used the percentages by state which are based on the *entire* state population of 10 years old and over, *including* non-whites. Robinson’s state-level ecological correlation was 0.773. The individual table (with a correlation of 0.203) was still based on the black and white population only. As a consequence, the mathematical relationship between the individual correlation (0.203) and ecological correlation (0.773) did not hold. This of course is logical because *per state* his individual Table 1 has marginal counts (non-whites excluded) that are not equal to the counts used in his original Figure 2 (non-whites included) and this is especially so for Arizona and New Mexico, as we showed in **Table 1**. To restore the relationship between the individual and ecological correlation (i.e., according to Robinson's Equation 1), we deleted all non-whites at state-level and re-calculated the ecological correlation. We also discovered that Robinson did not weight his ecological correlation. This is important because correlations may be sensitive to weighting if the total number of lower level units (here individuals) varies substantially across the higher level units (states). This is especially so if some of the data points are extreme cases that do not fit the linear trend.(**11**)

Robinson remarked in his paper that weighting the ecological correlation is less substantial (pp. 339),(**2**) but holds true only for the division-level analysis, and not the state-level analysis. We calculated an unweighted correlation coefficient with non-whites excluded at the individual and the state level. The unweighted state-level correlation was 0.874 in comparison with Robinson's 0.773. Upon correcting Robinson's erroneous original Figure 2 and subsequently weighting the data for population size, the discrepancy in the correlations between state and individual level were even more marked with state-level correlation being 0.913, and the individual level correlation being 0.203, and importantly conforms with Robinson's Equation 1 (pp. 340),(**2**) that the ecological correlation is the weighted difference between the individual correlation and the average of all within-*state* correlations. It is interesting to note that the weighting effect is almost completely attributable to Arizona and New Mexico. We elaborate on this at http://www.ru.nl/mt/rob/downloads.

Robinson's demonstration of the negative correlation between illiteracy and nativity at the division-level and a positive correlation at the individual level (see his Table 3/Figure 3) was also problematic. In these division-level analyses, the non-whites again are excluded, so his Figure 3 portrays the association between percent foreign-born and percent illiterates among whites and blacks. We may add that Robinson used the phrase 'foreign-born people', but in fact they are all foreign-born whites because in 1930 the U.S. Census Bureau did not publish state-wise total number of illiterate foreign-born U.S. citizens. In his analyses, Robinson found a negative ecological correlation of −0.619 at division-level and an individual level correlation of 0.118. Indeed, these two correlations are related according to Robinson's Equation 1. He additionally claimed an ecological correlation of −0.526 at state-level, but this ecological correlation is again based on the complete census data (non-whites included) and was unweighted. Weighting the data produced a correlation of −0.509 and omitting the non-whites gave a weighted ecological correlation of −0.462. Hence, after correcting the errors of not weighting and including non-whites at state-level while excluding them at the individual level, the difference between the individual and state-level ecological correlation is somewhat less but still clear.

We personally favor the illiteracy-nativity relationship as an illustration for potential ecological fallacies. At state-level a negative relationship exists (−0.462), so the higher the percentage of white immigrants in a state, the lower the percentage of black and white illiterates. This might suggest that among these immigrants there were relatively less illiterates compared to the native population. Robinson, however, showed that the opposite was true: among white immigrants the average illiteracy was higher compared to natives (individual correlation was 0.118). Clearly, immigrants settled in states where illiteracy was low. We present the scatter diagram (see **Figure 1**) in which the illiteracy-nativity relation is shown at state-level (which Robinson did not present at all), corrected for the incorrect inclusion of non-whites on the ecological level which lead to the weighted correlation of −0.462.

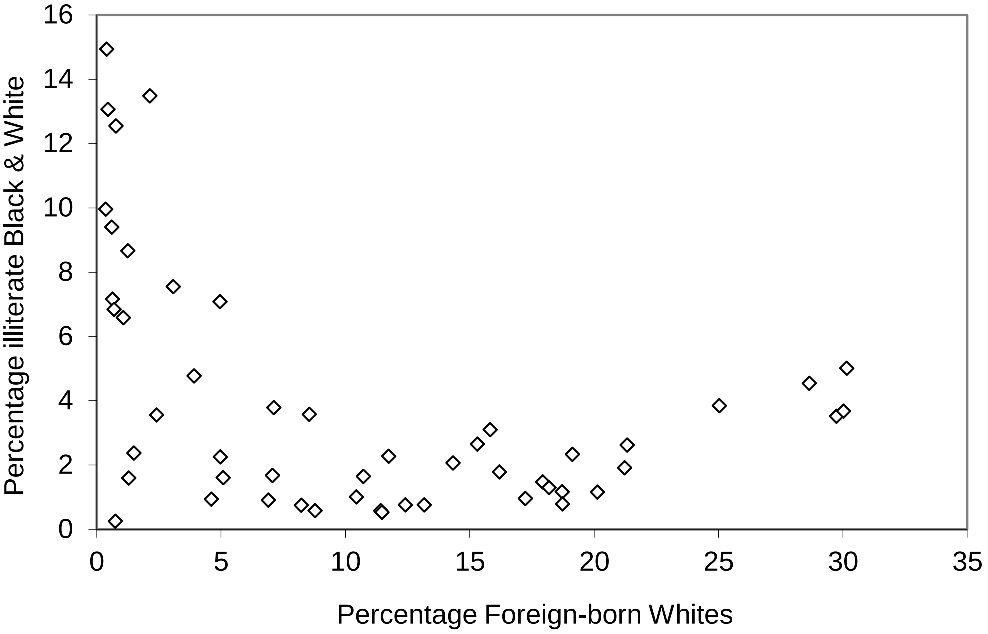

Figure 1 Relationship between percent literacy among blacks and whites and the percentage foreign-born whites in 1930 across 48 U.S. states and the District of Columbia

In this figure we clearly find a negative trend, although the trend is non-linear, which already was showing somewhat in the original Figure 3 at the division-level, suggesting that the Pearson correlation coefficient may not the best measurement for the association.

In this letter we showed that Robinson erroneously included the non-whites in calculating the ecological illiteracy-race and illiteracy-nativity correlations at state-level while he excluded them in calculating the underlying individual correlations. In addition to this data glitch, he did not weight his ecological correlations at state-level. An educated guess would be that Robinson did not note that the interior cells he used fail to sum up to the state-level totals and maybe he thought that weighting would be irrelevant at state-level since that appeared to be the case at the division-level. Fortunately, these errors do not alter the methodological contributions of Robinson's paper in demonstrating the discrepancy in results based on individual and ecological data. We remain intrigued as to how these inadvertent errors in one of the most cited and influential methodological papers went undetected for over 60 years, reiterating our belief in the value of replication for scientific inquiry and research.

References

1. Robinson WS. Ecological correlations and the behaviour of individuals. *American Sociological Review* 1950;**15**(3): 351-7.

2. Robinson WS. Ecological correlations and the behavior of individuals. *International journal of epidemiology* 2009 Apr;**38**(2): 337-41.

3. Subramanian SV, Jones K, Kaddour A, Krieger N. Revisiting Robinson: the perils of individualistic and ecologic fallacy. *International journal of epidemiology* 2009 Apr;**38**(2): 342-60; author reply 70-3.

4. Firebaugh G. Commentary: 'Is the social world flat? W.S. Robinson and the ecologic fallacy'. *International journal of epidemiology* 2009 Apr;**38**(2): 368-70; author reply 70-3.

5. Oakes JM. Commentary: Individual, ecological and multilevel fallacies. *International journal of epidemiology* 2009 Apr;**38**(2): 361-8; author reply 70-3.

6. Subramanian SV, Jones K, Kaddour A, Krieger N. Response: The value of a historically informed multilevel analysis of Robinson's data. *International journal of epidemiology* 2009 Jan 28;**38**(2): 370-3.

7. Wakefield J. Multi-level modelling, the ecologic fallacy, and hybrid study designs. *International journal of epidemiology* 2009 Apr;**38**(2): 330-6.

8. US Bureau of the Census. *Fifteenth Census of the United States: 1930. Population, Volume II, General Report. Statistics by Subjects.* Washington: United States Government Printing Office; 1933. Accessed at http://www2.census.gov/prod2/decennial/documents/16440598v2ch16.pdf

9. US Bureau of the Census. *Fifteenth Census of the United States: 1930. Population, Volume III, Part 2. Reports by States, Showing the Composition and Characteristics of the Population for Counties, Cities, and Townships or other Minor Civil Divisions*. Washington: United States Government Printing Office; 1932. Accessed at http://www2.census.gov/prod2/decennial/documents/10612982v3p2ch01.pdf

10. US Bureau of the Census. *Statistical Abstracts of the United States 1931*. Washington: United States: Government Printing Office; 1931. Accessed at http://www2.census.gov/prod2/statcomp/documents/1931-02.pdf

11. Belsley DA, Kuh E, Welsch RE. *Regression diagnostics: identifying influential data and sources of collinearity*. New York: John Wiley; 1980.