Sixty years ago the late William S. Robinson (1913–96) published his *Ecological Correlations and the Behavior of Individuals*. The paper became an all-time classic and it is one of the most influential methodological papers in social sciences. To underscore its impact on epidemiology, this classic was reprinted in this journal, along with an original re-analysis of Robinson’s data and conclusion from a multilevel and historical perspective, with discussions. In this research letter, we identify, and correct, an error in Robinson’s original analysis.

Robinson used data from a US Census Bureau 1933 publication, Table 10 (p. 1229). 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 US Census Bureau’s nine geographical divisions in 1930. We note that Robinson used ‘negro’ in the original paper, while in our own results we will use the term ‘Black’. The illiteracy–race relationship was positive and the ‘ecological’ Pearson’s 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) by 9 (division) 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, p. 340). 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 tables. This is important to note because when Robinson analysed 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. Secondly, for reasons not known to us, Robinson used unweighted ecological correlations. Robinson’s Figure 1, Tables 1 and 2 are based on a (rounded) total of 97,272,000. However, the 1930 census reports a total population of 98,723,047 US citizens (p. 1219). Upon scrutinizing US Bureau of the Census, we discovered that Robinson excluded 1,449,824 non-Whites, which are foremost Mexicans (1,002,241), other categories are Indians (238,981), Chinese (63,932), Japanese (97,273), Filipino (42,964), Hindu (2833), Korean (1446) and other (694). Thus, Robinson’s Figure 1 relates to the percentage of Black and the percentage of illiterate among 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 percentage of illiterates 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.

Table 1 shows that in New Mexico, and especially in Arizona, the percentage of illiterates 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 1877 (0.8% of the Black and White population). We must note that the US Census Bureau Table 10 only reports the percentage of 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 US Census Bureau tables and publications, however, do include such a warning, see for instance, US Bureau of the Census (p. 35), and US Bureau of the Census (p. 2). 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 US Census Bureau Table 10 total columns that included the non-Whites, whereas the underlying individual table still excluded non-Whites.

Robinson argued that he used divisions only for simplification (p. 338). He then continued with his
Figure 2 that gives the relationship between percentage of illiterate and percentage of Blacks at state level. To construct this figure Robinson used the percentages by state which are based on the entire state population of 10 year olds, 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 (p. 339) but this holds true only for the division-level analysis, and not for 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 was 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 (p. 340), 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 percentage of foreign-born and percentage of illiterates among Whites and Blacks. We may add that Robinson used the phrase ‘foreign-born’, but in fact they are all foreign-born Whites because in 1930 the US Census Bureau did not publish state-wise total number of illiterate foreign-born US 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, whereas excluding them at the individual level, the difference between the individual- and state-level ecological correlation is somewhat less but still clear.

We personally favour 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 with the native population. Robinson, however, showed that the opposite was true: among White immigrants the average illiteracy was higher compared with natives (individual correlation was 0.118). Clearly, immigrants settled in states where illiteracy was low. We present the scatter diagram (Figure 1) in

Table 1 Total illiteracy and Black and White illiteracy in New Mexico and Arizona, 1930

<table>
<thead>
<tr>
<th>State</th>
<th>Total</th>
<th>Total illiterates</th>
<th>Percentage of illiterates</th>
<th>Total Black and Whites</th>
<th>Total Black and White illiterates</th>
<th>Percentage of Black and White illiterates</th>
</tr>
</thead>
<tbody>
<tr>
<td>New Mexico</td>
<td>314370</td>
<td>41845</td>
<td>13.3</td>
<td>252718</td>
<td>19403</td>
<td>7.7</td>
</tr>
<tr>
<td>Arizona</td>
<td>335029</td>
<td>33969</td>
<td>10.1</td>
<td>222392</td>
<td>1877</td>
<td>0.8</td>
</tr>
</tbody>
</table>

Source: US Census Bureau, Chapter 13, Table 10 (p. 1229).
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 \( r = 0.462 \). In this figure, we clearly find a negative trend, although the trend is non-linear, which already was somewhat evident in the original Figure 3 at the division level, suggesting that the Pearson’s correlation coefficient may not be the best measurement for the association.

In this note we showed that Robinson erroneously included the non-Whites in calculating the ecological illiteracy–race and illiteracy–nativity correlations at state level, whereas 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. 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, re-iterating our belief in the value of replication for scientific inquiry and research. Fortunately, however, these errors do not alter the methodological contributions of Robinson’s paper in demonstrating the discrepancy in results based on individual and ecological data.

References


doi:10.1093/ije/dyr081

Advance Access publication 19 May 2011
© The Author 2011; all rights reserved.