Analysing the Data

Laurie and Winifred Bauer

Once all the data from the questionnaires had been entered into Excel files, the analysis process began. The first files tackled were some of those where the patterns had been very obvious in the course of entering the data. However, after three or four of these had been analysed, the remainder of the questions were analysed in numerical order.

In the original Excel files, the schools were ordered according to their Ministry of Education number. This system has a regional component, but other criteria are also used for the listing. Composite schools all appear first, in approximate north to south order, and then alphabetical within major regions, followed by other schools in north to south regional order, alphabetical within those regions. However, since the original Ministry of Education numbering, some schools have closed, and their numbers have been allocated to new schools, so that these schools sometimes have numbers quite unexpected given the general principles. The first step in the analysis was to establish a north-to-south order for the participating schools, with west-to-east as a second criterion, while keeping schools in particular regions together as far as possible. For this reason, in the centre of the North Island in particular, strict north-to-south order was abandoned, so that all Taranaki schools were together and all Hawkes Bay schools were together. Having the schools clustered in this way made it easier to see whether forms were regionally patterned or not.

A working copy was made of the original data file, so that the full data set was not lost. After the schools had been re-ordered, the first step in the analysis involved moving columns with similar responses to adjacent positions. Thus, for instance, in the question about two students riding on a one-seater bicycle, all responses which preserved the <l> in the root (e.g. *doubling, a double, doubles*) were grouped, and all responses without the <l> were clustered (e.g. *dubbing, a dub, dubs*).

Any one-off responses were then deleted, but only after the researcher was satisfied that they did not closely resemble any other response. (A single response of *dubs* would thus not have been discarded, on the grounds that it was sufficiently like other *dub-* forms to be of interest. However, the one response *skitching* bore no resemblance to any other form in any questionnaire, and was therefore discarded as being of no further interest.) Similarly, any forms with very small numbers of reports (fewer than 5) were discarded if those reports were scattered round the country. However, if the reports came from schools in close proximity, they were retained.

At the next stage of the analysis, a decision was made about whether or not it was appropriate to group forms which were similar but not the same. So in the data above, a decision had to be made about whether *doubling, a double, doubles* were to count as reports of the same form or not. This decision had to be made very carefully, on a case-by-case basis. While singulars and plurals, or different tenses of a verb were often merged, and forms like *crap* and *crappy* or *It rocks* and *rocky* were often treated as evidence of the use of the same root, this was always considered carefully, and the two were not merged if there appeared to be reason

to keep them apart. Thus *white rabbits* and *white rabbit* were not merged, since it seemed possible that *white rabbit* was an innovation, and we might need to know where this innovation appeared. Often in cases like this, two passes were made over the data, one which treated the two separately and one which merged them. In the question about two people riding on a one-seater bike, it was very clear from the data that the crucial variable was the presence or absence of the <l>, rather than an *–ing* form as opposed to an *–s* form (most often, the same school would report using both the *–ing* and the *–s* form of a root). The number of forms in the data was thus reduced as far as possible by grouping of this kind.

When no further simplifications and eliminations were justified, the forms reported were mapped. There were often too many forms in one set of data to put on one map, and in these cases, the forms were divided into groups on the basis of frequency, and mapped on as many maps as was necessary. In general, it was possible to put 8-10 low frequency forms on one map, but only 4-5 higher frequency forms.

The maps were created as Microsoft Word tables. While there were problems inherent in this, it presented a manageable form of mapping without requiring any expenditure on programming. However, these maps proved troublesome to transfer cross-platform, and the map border remained somewhat unstable.

Initially, the rural schools were mapped on a two-page map of the country, and the urban schools were mapped into separate tables. The data was entered by allocating a particular character to each form, and typing that symbol into the appropriate box. Where one school reported several of the forms on a particular map, all the necessary symbols were put in that box. These maps were then printed in black and white, and then hand-coloured with a variety of coloured pens. The urban areas were filled with a grid pattern, and these were coloured to reproduce the distribution shown in the large-scale tables, thus giving a composite map. A visual inspection of the map then revealed any likely cases of regionalisation, including urban and rural differences. In the most dramatic cases, the colours all occurred in blocks. In cases where there was no regionalisation, the maps were a mosaic of different colours.

On a number of occasions, after the initial mapping process, it was seen that certain forms were regionalised, and others not, and if necessary, additional maps were produced which contained only the forms of interest.

These original maps, while satisfactory for the initial analysis, were not suitable for publication, and for the sets of data with clearly regionalised results, other versions of the map were produced. In particular, a composite map outline was made with an insert grid for each of the urban centres with more than two participating schools. Where the box on the main map also contained a rural school, the rural school was included in the insert for practical reasons. This composite map could then be shaded in black and white or in colour for publication purposes.

While the mapping process is quite time-consuming, maps are the obvious and most readily absorbed method of presenting regionalised data. This was supported by tables which included counts of the reports of forms in particular regions, which enabled the analyst to see whether the percentages were in the vicinity of the expected (e.g. the Northern Region contained 38% of the schools sampled, and this figure could be compared with the percentage of the reports of the individual form, with wide deviations from 38% indicating either a Northern

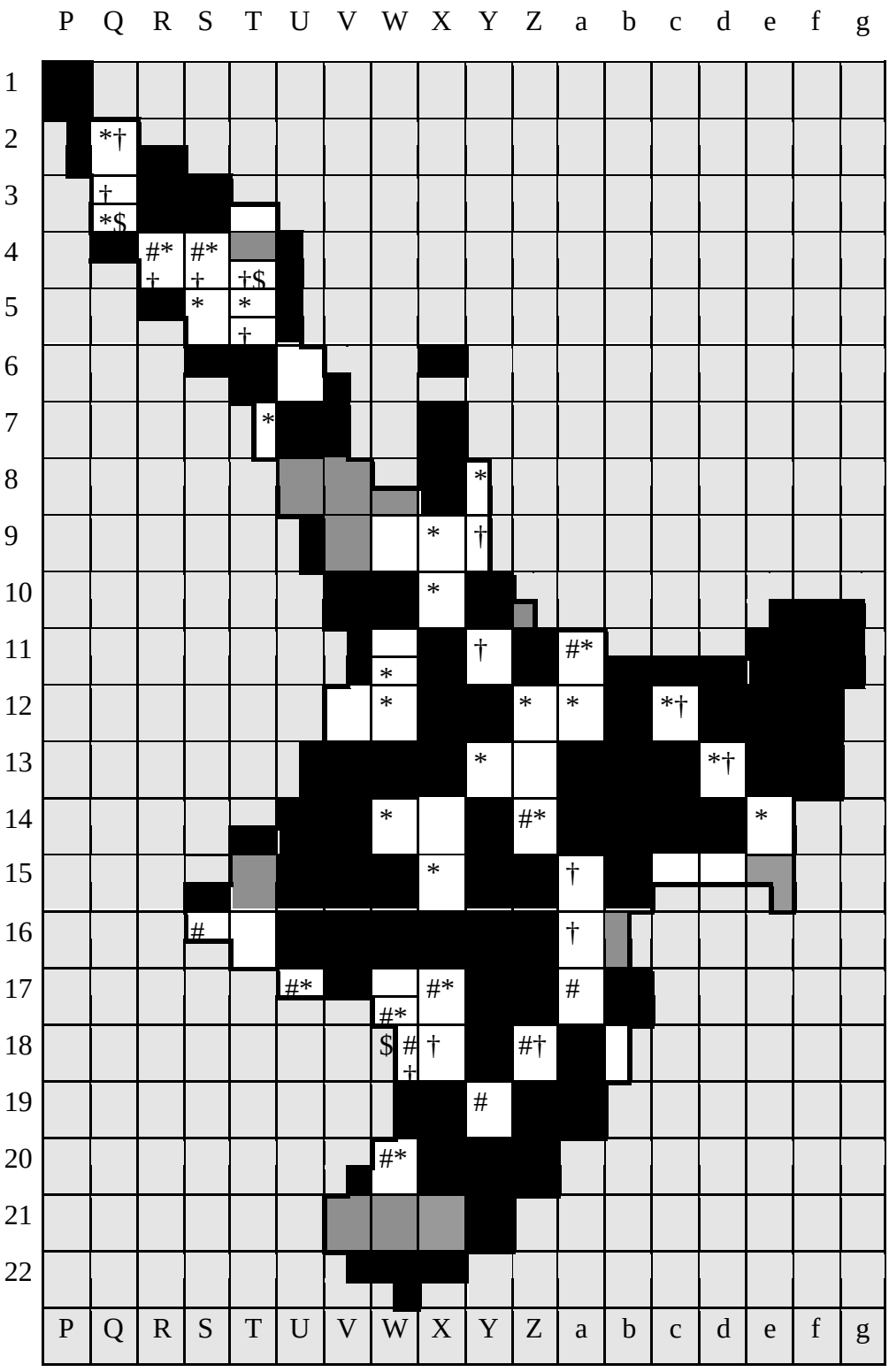
Region form or a non-Northern form). The analysis of the most obviously regionalised forms had established that there were two different divisions which were likely to be important: the Island division, and a division into three regions which we named Northern, Central and Southern. At this stage in the analysis, there were certain areas whose affiliation to one or other of these regions was uncertain: Taranaki, for instance, sometimes appeared to be Northern, but at other times appeared to be Central. At this stage of the analysis, these doubtful regions were excluded from the tables, so that the tables included only the clear-cut schools. After the statistical analysis had determined the overall affiliation of these schools (all but one!), the figures in the tables were adjusted accordingly. The tables presented in the documents discussing the data sets thus do not reveal the earlier stage in the analysis.

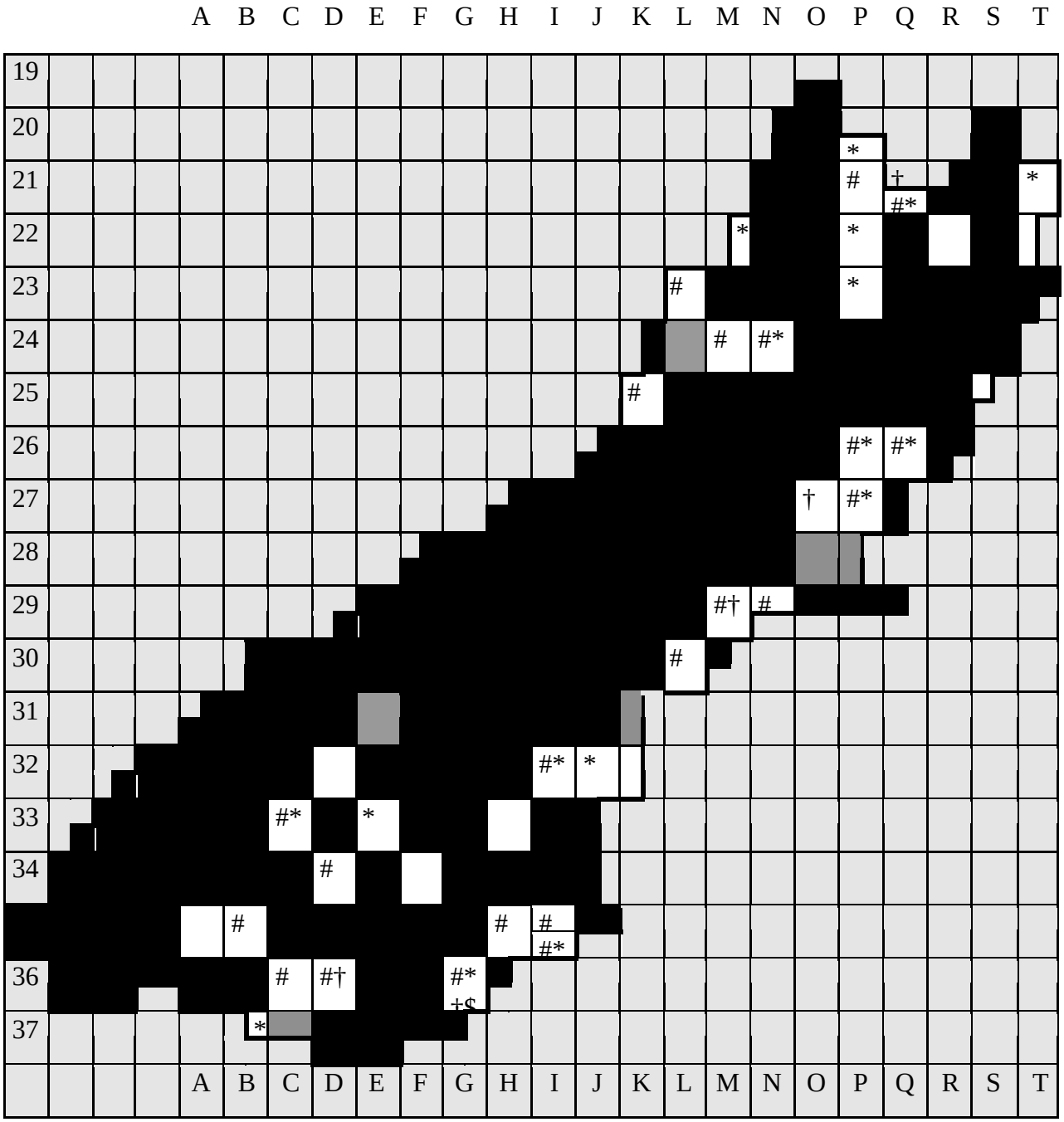
In addition, it was quite often noted as the data was analysed that certain forms were reported principally by schools at one or other end of the decile scale. In these cases, a further copy of the file was made and schools were ordered by decile, and calculations were made of the percentage of schools in each decile which reported the form in question. Excel then produced graphs of these results. These graphs enabled the identification of forms which were unevenly distributed across the deciles, and thus likely to be socially variable.

When all the data had been analysed in this way, the forms which patterned according to one or more of these factors (region, decile, urban-rural) were identified for each question, and formed the input to the statistical analysis.

One example of the original maps is given below. The composite maps are illustrated in many of the documents discussing the data.

Q8: 3rd the …

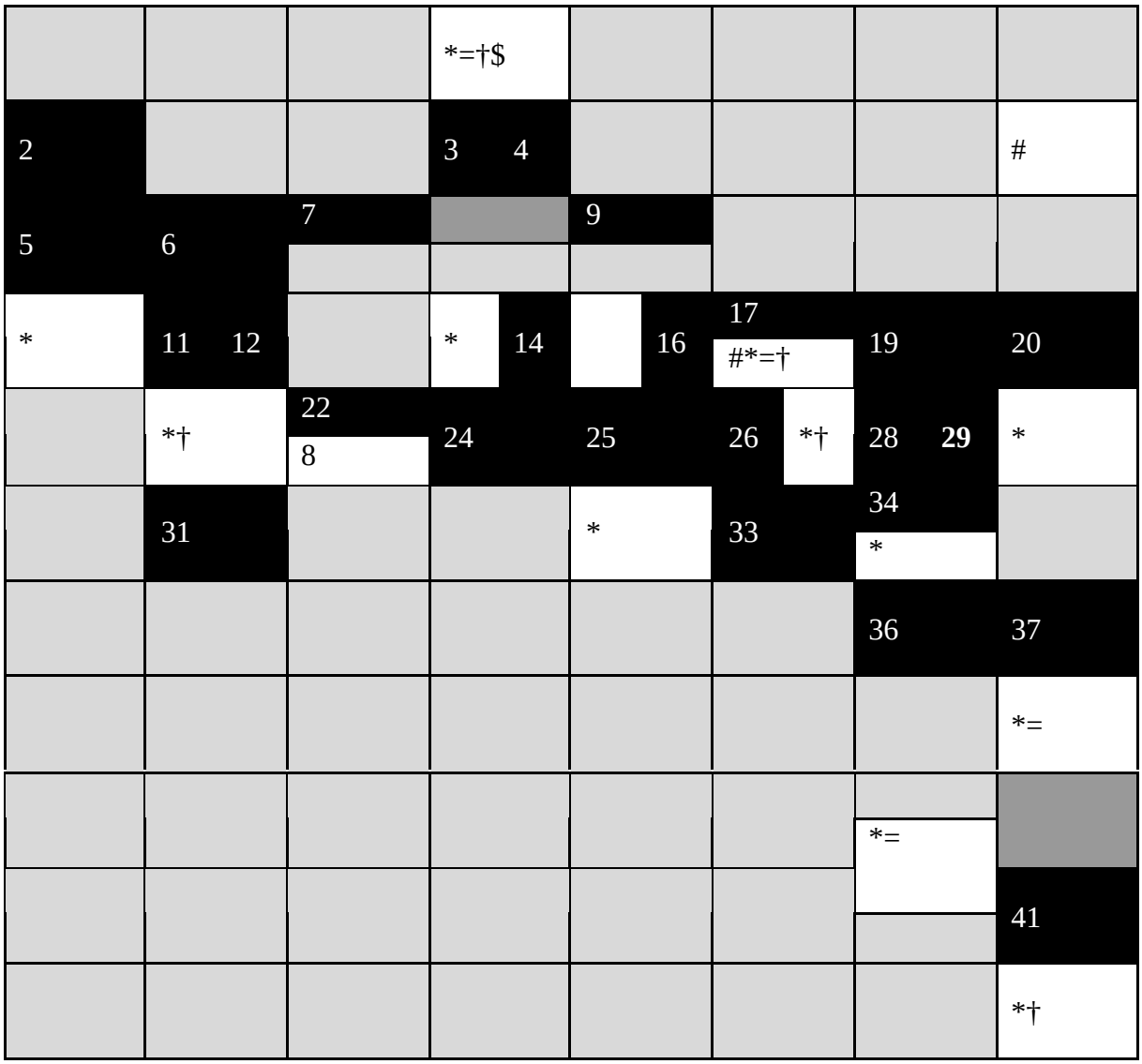

P Q R S T U V W X Y Z a b c d e f g
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
P Q R S T U V W X Y Z a b c d e f g



Key

#	golden princess	*	golden eagle		none of these		See Urban Map
=	hairy chest	†	nerd		no data supplied		No data available
$	turd	£	golden prince				

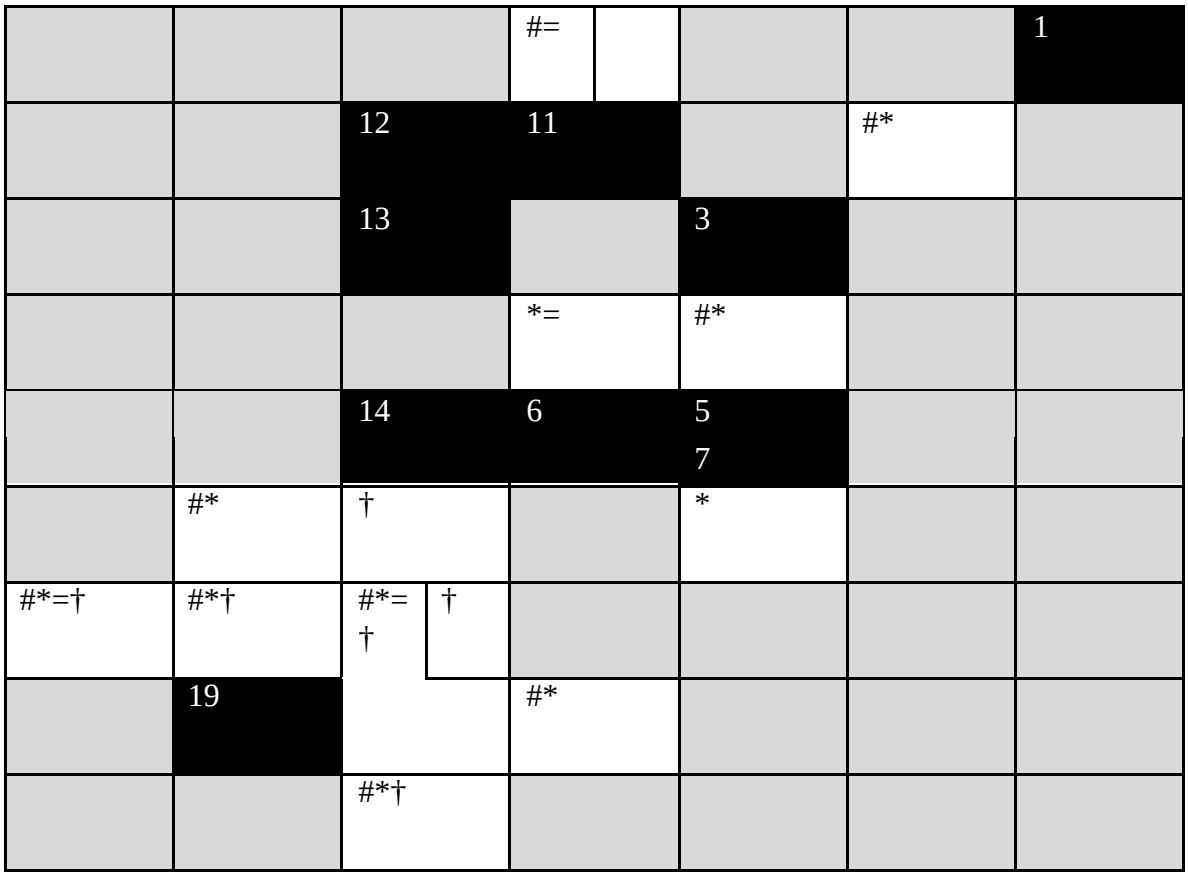
Auckland (U8, V8, V9, W8)



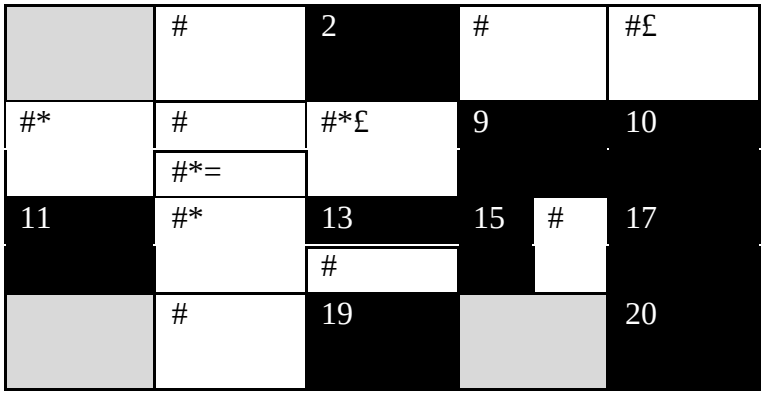
Hastings (b16)

New Plymouth (T15)

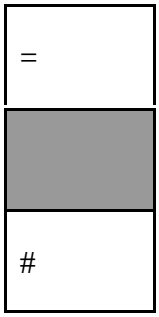
Wellington (V21, W21)



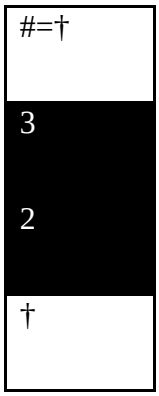
Christchurch (O28, P28)



Timaru (K31)



Invercargill (C37)



Key:

#	golden princess
*	golden eagle
=	hairy chest
†	nerd
$	turd
£	golden prince
	none of these
	no data supplied
	No participating school
	Land/sea outside urban area

