2nd Annual International Conference on Cognition, Language and Special Education Research, *Educating: Weaving research into practice,* 3-5 December 2004

Demographic profiling for educational researchers: Using SPSS Optimal Scaling to identify distinct groups of participants

Bob Funnell, Fiona Bryer, and Peter Grimbeek

Griffith University, Centre for Learning Research

Abstract

The French sociologist, Pierre Bourdieu, pioneered new methods for examining the way in which social actors occupy positions in a field, or social space. These methods also have application to educational research. Bourdieu's analysis of the French public's judgments about taste in art and wine, and their social class, yielded the useful insight that social actors make choices to differentiate them from those in other classes. He pioneered what subsequently has been termed correspondence analysis to spatially represent the groupings produced by such choices. Market researchers have used software (including correspondence software) that capitalises on Bourdieu's insights to assist them in identifying types of consumers (e.g., Generation X). A multivariable version of Correspondence software, SPSS Optimal Scaling, turns out to yield interesting insights when applied to data sets collected in the course of educational research. Importantly, Optimal Scaling is able to utilise a nonparametric factor analytic procedure to analyse and spatially represent the interactions between ranges of categorical, ordinal, and interval level variables. This paper discusses the origin of and potential uses of this statistical software in educational research.

Re-evaluating the role of demographic data in educational research

A routine aspect of educational surveys is the collection of conceptually relevant information about the personal and social characteristics of participants. This data is usually reported in passing and addressed via a series of comments about frequencies, percentages, averages, and so on. Subsequent reference to such information in the discussion of the results of educational surveys typically relates to reports of group differences. However, this approach to the analysis of demographic data grossly undervalues its potential contribution.

It is surprising to realise that even the descriptive uses of demographic data outlined above was once perceived to be a radical innovation. Bourdieu (1979), initially trained in philosophy, was the first anthropologist to systematically use demographic statistics in fieldwork when he recorded the frequencies of all types of marriage in the Kabyle and other regions of Algeria. Surprisingly, this data showed that, within the universe of all marriages, those to the parallel first cousin, regarded as the norm in the anthropological canon, occurred in a low percentage of cases. Bourdieu intended only to check a pattern in one area: He did not intend to disprove a theory. He had employed statistics in an area where concepts had been established through qualitative methods and reinforced in diagrams and charts showing these norms for marriage. A major revision followed his research that, over time, has influenced wider theories about marriage strategies.

How people occupy social space

Bourdieu's early use of correspondence analysis to examine social spaces represented an even more radical step in the fields of anthropology and social science. A social space can be understood in terms of relationships that, over time, have been established between persons, groups, and institutions.

Bourdieu (1984), in his book, *Distinction*, took France to be a social space. Instead of measuring numbers in and mobility between different class occupations, Bourdieu conducted a detailed examination of the material and educational assets available to members of different occupations, what he called their economic and cultural capital. The size of a type of capital determined the area one can live in, how the residence will be furnished, where one's children can attend school, and for how long. In a survey, he asked about the minutiae of social life, about food, clothing, preferences in music, movies, art, newspapers, magazines, sport, voting, etc.

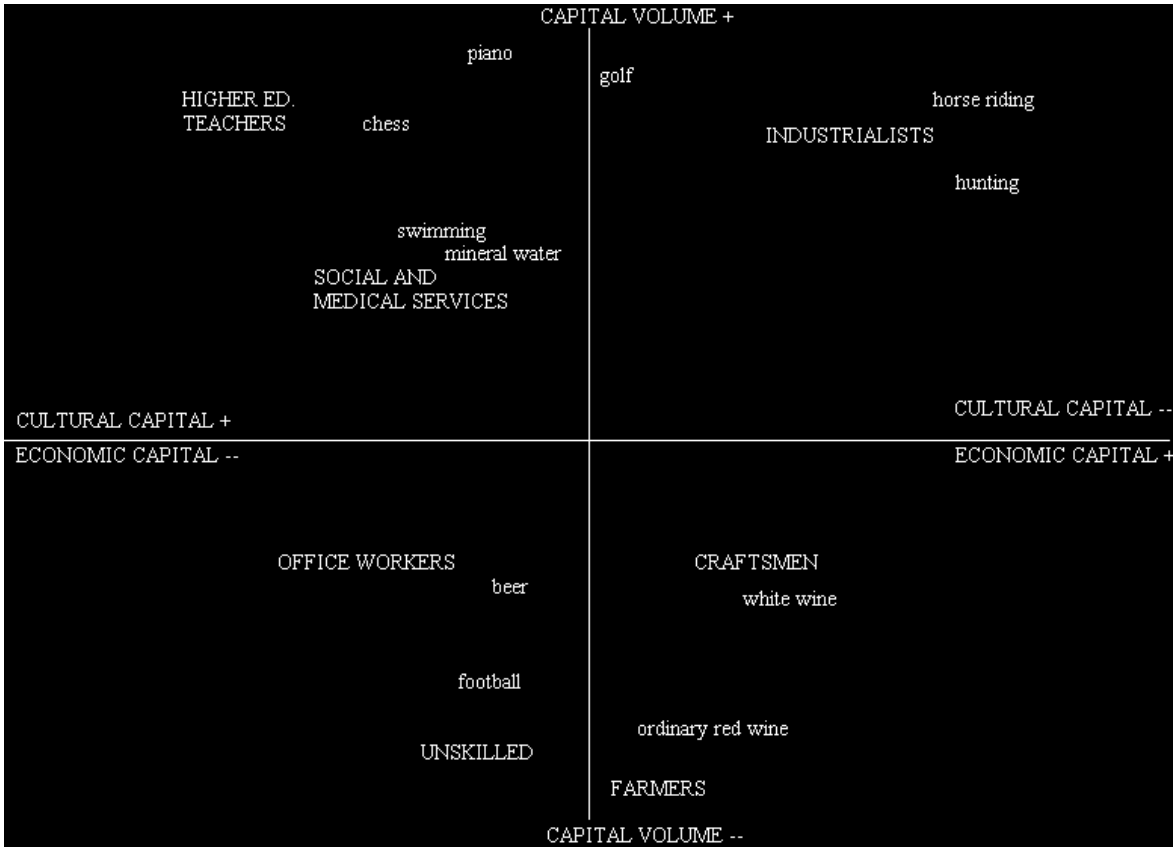

Figure 1. Approximation of a social space for cultural consumption (adapted from Bourdieu, 1984)

As shown in Figure 1, Bourdieu (1984) used correspondence analysis on these survey results to assist the meaningful construction of the space. Correspondence analysis is a statistical method for transforming a cross-tabulation of data into a graphical display. The variables are weighted negatively and positively within two dimensions and represented as dot points in a four-quadrant space. The graphic display, such as the one shown in Figure 1, is a guide to visually represent participant positions in social space. It brings together clusters of occupations and categories with similar weightings and separates them from others with different weightings. This method allowed Bourdieu to show that to occupy one place in a social space is to differ from others occupying another place within it. The analysis of

correspondences plotted factors influencing various places in the space (see Figure 1), and the relationships revealed in Figure 1 show how groups of people occupy social space.

It is the intersection of these variables that tells a story in Bourdieu's partition and apportionment of social space in terms of variables indicative of economic and cultural capital. Bourdieu's four quadrants distinguished between four French groups and characterised the bases of their grouping: higher education teachers with large amounts of cultural capital, office workers with some cultural capital, craftsmen with some financial capital, and industrialists with large amounts of economic capital. Figure 1 also shows that individual taste in relation to alcoholic beverage did essential work in differentiating these groups. Whereas office workers with minimal financial and some cultural capital preferred beer, craftsmen with minimal cultural but some financial capital preferred wine. That is, the choice of beer versus wine was "part and parcel" of their differentiation of social space.

In summary, Bourdieu showed that correspondence analysis could be used to introduce factors that would not ordinarily be included in a study. His analysis of taste in relation to choices between fine art and alcoholic beverage did essential work in differentiating groups of people who had accumulated financial versus educational social capital. This analysis could not have been built by isolating dependent and independent variables, but only via a technique that shows the structural relationships between the factors being analysed.

Two points were made in this section. First, statistics employed in the early stages of research have a potential to bring things to the surface that could not be anticipated in established theory. Second, a range of categories can be combined in single analyses. Market researchers and others have adapted the notion of social space to denote, for instance, generational groupings of participants who share common experiences. The universal currency of the term "Generation X" (Coupland, 1991) provides an ad hoc demonstration of the marketing advantages attendant on such labeling.

Optimal Scaling and demographic data

A modern version of Correspondence Analysis is Optimal Scaling. This approach can accommodate demographic data across various kinds of participation and various measurement assumptions. Current SPSS (2003) software, Optimal Scaling, allows for choice between analyses that treat all variables as having nominal measurement properties at best (Multinominal) or analyses that take into account potential differences in measurement properties (Non-multinominal). A scenario involving Chinese immigration to Australia (Wang, Davies, Grimbeek, & Loke, 2004) illustrates the uses of descriptive data when appropriate analytic procedures are applied to survey data in various ways.

It has been noted that the demographic data routinely collected by researchers has the capacity to reflect the differing ways in which a range of people occupies social space. That is to say, a field of participants can be described in terms of the extent to which the participants "clump" across measures such as gender, age, country of birth, etc. Response categories in relation to a demographic measure such as country of birth—a passive descriptor of a participant—are not exactly equivalent to a participant actively selecting wine versus beer. In the case of migrants, however, this demographic might well reflect social choices on the part of the host country if nothing else.

The data points in this scenario also involve measures of differing sorts in terms of measurement properties. When speaking about wine versus beer, the choice is, at best, either

categorical or nominal: That is, participants prefer one or the other according to their social position. When speaking about age, the measure has ordinal or interval properties (e.g., age group, age in years), and so on. The importance of this distinction is that it influences the kind of analysis of associations between variables undertaken. Analyses in Optimal Scaling can either treat all variables as nominal (Multinominal) or can allow for potential differences in measurement properties (Non-multinominal). A rationale for the multinominal analysis is that it requires a minimal set of assumptions about the properties of variables, and a rationale for the non-multinominal analysis is that the acknowledgment of measurement properties makes the analysis more sensitive and powerful. In either case, some form of clustering (equivalent to nonparametric factor analytic procedures) of response categories across variables occurs.

The type of Optimal Scaling analysis required and the desirable format for variable response categories can be related to subsequent analyses. Whereas Bourdieu complemented his spatial representations with contingency table analyses, latter-day researchers are likely to follow Optimal Scaling and similar analysis by using parametric analytic procedures to test the significance of associations between demographic variables and other outcome scores.

Optimal Scaling analyses can provide a rationale for collapsing across low frequency response categories. Researchers collapse categories within variables as a prelude to entering demographic variables as independent variables (IVs) in univariate or multivariate parametric analyses. They do so to minimise the likelihood of such analyses becoming unstable (and, hence, unreliable) because of the presence of one or more cells (a cell being equivalent to a single or conjoined response category such as older females) with fewer than 6-10 respondents or one or more cells with fewer than 10% of participants. Hence, collapsing across response categories can be a desirable intermediary step after preliminary categorical level analyses, aided by frequency and contingency table analyses of participant numbers.

Any collapsing of response categories affects follow-up examinations of the social space unfolded via demographic data. Collapsing across response categories affects analyses because it reduces the apparent complexity of the data. This reduction in complexity could be considered to be a loss or a gain depending on the value of that more complex description.

In summary, there is more than one way to represent the underlying demographic realities, and there is no absolutely preferred approach. Constraints in this process include the decision to collapse variables for the sake of subsequent analyses and the decision to enter variables into analysis with due recognition of their measurement properties. These points are illustrated in analyses of demographic data from a study of Chinese Australian migrants (Wang et al., 2004).

Optimal Scaling of selected demographic variables related to a sample of Chinese migrants to Australia

Henry Wang (MEd student) surveyed a sample of 600 Chinese Australians, all members of community groups (Wang et al., 2004). Most of these were female (71%, n = 489), the largest group of participants was 40-49 years of age (42%, n = 253), and about 60% (n = 359) described themselves as unemployed or retired. About 2/3 (n = 401) came from Taiwan. The range of years in which these migrants arrived in Australia spanned a 34-year period, 1970-2004. The process under consideration is how best to assemble these bits and pieces of demographic information into a joint demographic profile of their social space.

First analysis: Multinominal scaling

SPSS (2003) allows a data analyst to plot the relationships between two variables (Correspondence analysis) or many variables (Optimal scaling). A starting point is to use Optimal Scaling to generate a spatial representation somewhat akin to that favoured by Pierre Bourdieu (1979). This initial view of the sample is achieved by entering variables into Optimal scaling as multinominal variables. That is, all the variables are treated as nominal, and the number of categories (as a range) per variable is specified (e.g., there were five age groups, and four groups of migration years).

Multinominal[1] Optimal scaling (see Figure 1) does Homogeneity analysis, which quantifies nominal (categorical) data by assigning numerical values to the cases (objects) and categories. The goal of Homogeneity analysis is to describe the relationships between two or more nominal variables in a low-dimensional space (i.e., minimum of two dimensions) containing the variable categories as well as the objects in those categories. Objects within the same category are plotted close to each other, whereas objects in different categories are plotted far apart. Each object is as close as possible to the category points for categories that contain that object. Homogeneity analysis is also known in the literature as Multiple Correspondence Analysis. Homogeneity analysis can also be viewed as principal components analysis of nominal data. Homogeneity analysis is preferred over standard principal components analysis either when linear relationships between the variables don't hold or when variables are measured at a nominal level.

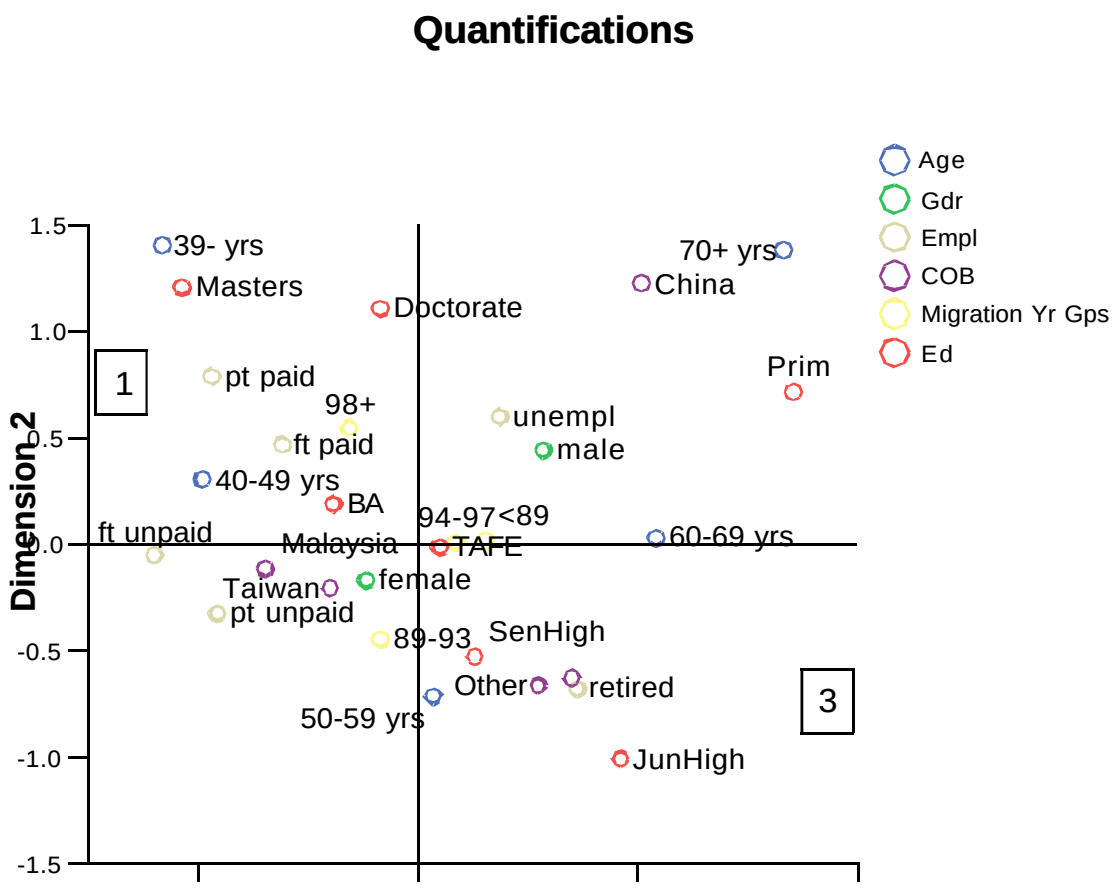

Figure 2. Spatial representation of associations between demographic variables, all treated as nominal

The initial Optimal Scaling procedure[2] used in this Chinese-Australian sample made the minimal assumption that all response categories were nominal. With this procedure, associations were plotted between six demographics (age group, migration year, gender, employment status, country of birth, and educational level) with the twin aims of (a) facilitating the steps of collapsing across response categories and (b) selecting variables for use in subsequent analyses. Marital status was omitted because a very large majority (90%) stated that they were married.

As shown above in Figure 2, if one considers quadrants 1 and 2, participants who migrated in 1998 or more recently, with bachelor, master, and doctoral qualifications, and with full-time and part-time paid work, in the age group categories, 39-and-under and 40-49 cluster in quadrant 1. Likewise, female participants who migrated in 1989-1993 from Taiwan and Malaysia, with part-time and full-time unpaid work cluster in quadrant 2. (Note anticlockwise numbering of quadrants.)

If one considers quadrants 3 and 4, participants from Hong Kong (not visibly indicated on map) or other places, with junior or senior high school qualifications cluster in quadrant 3. Finally, unemployed male participants from China in the 70-or-older age group, with primary school qualifications cluster in quadrant 4. It is notable that the response categories for the years 1994-1997 and for the years prior to 1989 are aligned with the midpoint between quadrants 3 and 4, as are participants with TAFE qualifications in the age group 60-69, indicating that participants from quadrants 3 and 4 share many of these characteristics.

This clustering of demographic categories in Figure 2 supported collapsing the six age groups into four categories at most (<50, 50-59, 60-69, <69). However, since fewer than 10% of participants were in the over-69 age bracket, and since the 60-69 age bracket overlapped the 70-plus age bracket in terms of quadrants, it was more functional from an analytic point of view to collapse age groups into three categories (<50, 50-59, >59). It also supported collapsing employment status into four categories at most (paid, unpaid, unemployed, retired). It also supported collapsing country of birth into three categories at most (Malaysia/Taiwan, Hong Kong/other, China). It supported collapsing the categories for educational qualification into three categories at most (primary, high/TAFE, tertiary). However, given the overlap for TAFE and the minimal percentage (4%) of participants with primary school qualifications, it was more functional from an analytic perspective to collapse educational qualifications into two categories (pre-tertiary, tertiary). Finally, although two of the four migration year groupings overlap (1994-97, <1989), the gap between these two categories made collapsing conceptually indefensible. So, migration year remained a four-category variable (<89, 89-93, 94-97, 98+).

Second analysis: Nominal and ordinal scaling

This second variety of the Optimal Scaling procedure involves indicating that some variables are other than nominal (i.e., age groups, migration year). In this case, the form of Optimal Scaling undertaken was a Categorical Principal Components Analysis (CATPCA).

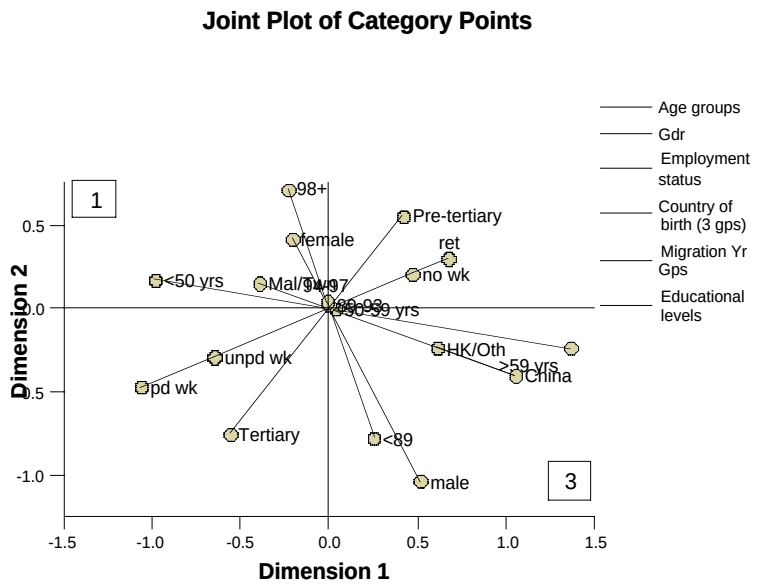

Figure 3. Collapsed categories and analysis with measurement properties taken into account

Inspection of Figure 3 made it clear that this sample contains a cluster of participants from Hong Kong, other places, and China. For this reason, the country-of-origin variable could be collapsed accordingly. Likewise, participants that did paid or unpaid work versus those that didn't work or were retired formed two distinct clusters, and the categories for this variable could be collapsed accordingly. Finally, the response categories for two of the four migration year groups coincided at the centre of the display and could be collapsed together.

Third analysis: Nominal & ordinal scaling

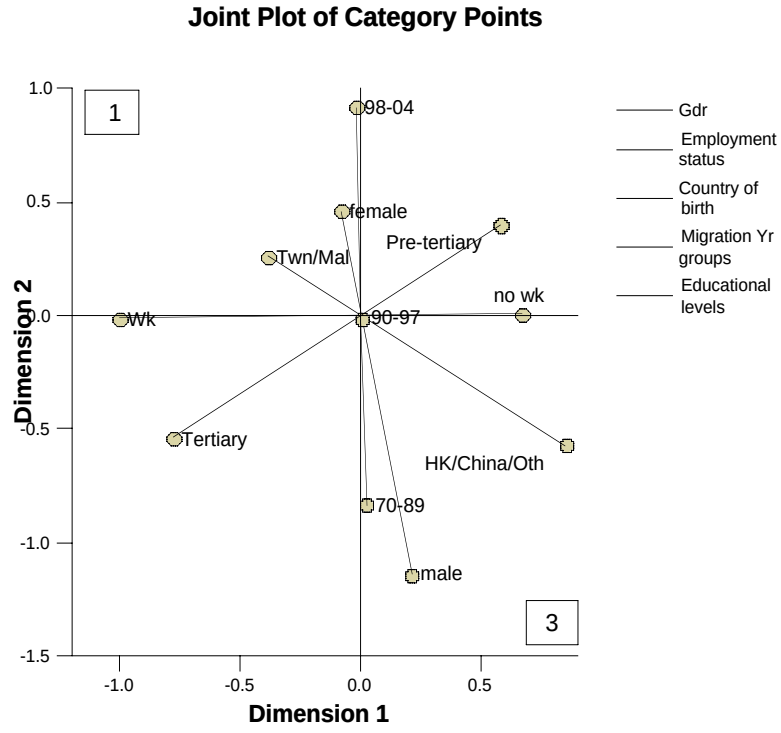

Figure 4. Additional collapsed categories, analysis with measurement properties taken into account

An accompanying examination of associations between these variables (Spearman's rho, a nonparametric measure of correlation) indicated age group to be significantly positively associated with employment status (i.e., being unemployed), and country of birth (i.e., being from China, Hong Kong, or other countries) and significantly negatively associated with gender (i.e., being male), migration year (having arrived more recently). That is, age was significantly associated with each of the other four demographic variables under consideration. Because it was, in a sense, redundant, it was excluded as an indicator.

In Figure 4, the vertical axis was closely aligned with year of migration and the horizontal axis with employment status. It follows that, in terms of the horizontal axis, one might expect participants, regardless of migration year, but without work to have migrated from China or Hong Kong or other countries, to be male, and to have pretertiary qualifications. Alternatively, one might expect participants with work to have migrated from Taiwan or Malaysia, to be female, and to have tertiary qualifications.

If one takes migration year into account, then one might expect male participants who migrated in the years 1970-1989 to be from China, Hong Kong, and other countries. Conversely, one might expect female participants who migrated in the years 1989-2004 to be from Taiwan and Malaysia.

Putting these alignments together and considering the six demographic measures in terms of a vertical split, an emerging generalisation is that participants who migrated prior to 1989, were likely to be male, from China or Hong Kong, aged 60 or more (at time of survey), possessed of TAFE qualifications at best, and likely to be unemployed or retired. In contrast, those who have migrated from 1998 onwards were likely to be female, tertiary educated, with work, and less than 50 years of age (at time of survey).

Companion chi-square analyses indicated that work status was not significantly associated with either gender or migration year but was significantly associated with country of birth ($\chi^2(1) = 31.141$, $p < .001$) such that those without work were more likely to come from Hong Kong, China, or other countries in the region. Likewise, work status was significantly associated with educational qualifications ($\chi^2(1) = 62.1035$, $p < .001$), such that those with pretertiary qualifications were more likely to be without work. Finally, migration year as such was not significantly associated with any of the other four demographic variables.

Selection of variables for subsequent analyses

One way to select variables for subsequent analysis is to select variables with orthogonal (independent) characteristics and to exclude nonorthogonal (significantly associated) variables. If one considered work status to be a useful indicator of attitudinal and behavioural outcomes, then variables with characteristics orthogonal to work status would include gender and migration year but would exclude country of birth and educational qualification.

Discussion

The preceding section demonstrated two approaches that approximate and extend the style of correspondence analysis first espoused by Pierre Bourdieu. As with his spatial representations, it becomes possible to talk about demographic indicators in terms of how participants group in various ways. That is, it becomes possible to discuss how they occupy common social space. It also becomes feasible to select variables for subsequent statistical analyses.

It should be clear that there is no single preferred approach to Optimal Scaling. These examples provide both simpler and more complex analyses of the same variable set. As usual, there is more than one way of representing the underlying demographic realities. Constraints in this process include both the decision to collapse variables for the sake of subsequent analyses and the decision to enter variables into analysis with due recognition of their measurement properties.

The outcome of this work can and should influence the selection of demographic variables into subsequent analyses. Clearly, Optimal scaling should be supplemented with other forms of analysis, including chi-square analysis, examination of correlation coefficients, and other like statistical procedures.

Footnotes

[1]Taken from SPSS Help files

[2]Variables treated as single set in every analysis reported here.

References

Bourdieu, P. (1979). Algeria 1960: The disenchantment of the world, the sense of honour, the Kabyle House or the world reversed. Cambridge: Cambridge University Press.

Bourdieu, P. (1984). Distinction: A social critique of the judgement of taste. London: Routledge & Kegan Paul.

Coupland, D. (1991). Generation X: Tales for an accelerated culture. London: Abacus.

Popper, K. (1963). Conjectures and refutations. London: Routledge and Keagan Paul.

SPSS for Windows. (2003). Chicago, USA: SPSS Inc.

Wang, H., Davies, M., Grimbeek, P., & Loke, K. (2004) The life satisfaction and well-being of a sample of Chinese Australians as indicators of participation in community educational activities. Paper presented at the 1st annual international conference on post-compulsory education and training, Gold Coast, Qld, Australia.