Sizing up categories: A reply to Morris (1977)

A. J. Wilkins

Morris (1977) claims that Wilkins' (1971) measure of category size may be inappropriate. In this reply to his criticism it will be argued that Wilkins' measure is perfectly appropriate, but that size, as a unitary concept, is inappropriate.

Measures and manipulations of category size

The literature contains a wide variety of methods of measuring or manipulating the size of a taxonomic category, but they may be divided into two broad groups: analytical and subject based. Analytical methods involve the use of nested pairs of categories such as animal-dog (Landauer & Freedman, 1968; Meyer, 1970; Landauer & Meyer, 1972) or counting the exemplars of a category in some compendium of English words (Neisser, 1964; Collins & Quillian, 1970). In addition to the criticisms that have been levied against these measures of size (e.g. Smith et al. 1974) all analytic measures suffer from the shortcoming that they fail to take account of the manner in which the category name may be interpreted. There is no guarantee that the category name is necessarily perceived by subjects as referring to all the objects that logic would dictate. This shortcoming assumed particular importance in the study by Collins & Quillian (1970). Collins & Quillian cite a subject who after classifying as animals ‘instances such as camel, otter, rabbit, kangaroo, etc.... was surprised when she encountered lizard. Apparently she had in mind a sub-class of animals (roughly “mammals”) and had been deciding whether each instance was a member of that subclass.’

Idiosyncratic interpretations of the category label such as the above have less importance in subject-based methods. These may be based on subjects' production of exemplars (e.g. Freedman & Loftus, 1971; Wilkins, 1971) or on subjects’ ratings (Battig & Montague, 1969; Morris, 1977). The essence of Morris' argument is that subjects' ratings of size have a greater face validity than other measures and yet fail to correlate with Wilkins' measure of size.

Wilkins' measure of size was the number of exemplars tabulated in category norms. These norms have been produced by asking each subject in a large group to give a limited number of category exemplars. The exemplars produced by the group are then tabulated in order of the frequency with which each occurs. The number of exemplars tabulated for a given category shows a high degree of consistency from one norm to another as can be seen from Table 1. Evidently Wilkins' measure of category size is reliable. But is it valid? Obviously the number of exemplars tabulated depends not only upon the size of the category but also the consistency in subjects' choice of exemplars. If every subject in the group chose the same few exemplars the category would appear unduly small. It is therefore of interest that Wilkins' measure of category size correlates with a measure of size that is unaffected by inter-subject consistency. The measure is obtained from the Bousfield equation. Bousfield (1944) found that when subjects were required to produce as many exemplars of a category as they could, the total number of exemplars emitted as a function of time was well fitted by a negative exponential. The asymptote of this curve provides an estimate of the number of exemplars that would be produced after infinite time, and can therefore be regarded as a measure of the size of the category. Murray (1975) has shown that estimates of the asymptote based on the performance of individual subjects are correlated (r = 0.796) with the number of exemplars tabulated in Battig & Montague's category norms. Evidently Wilkins' measure is not seriously confounded by inter-subject consistency. (For further evidence supporting this conclusion, see Hermann et al. 1973.)
Table 1. Spearman rank correlations between number of exemplars per category for the 15 categories listed in all the following category norms: Cohen et al. (1957), Battig & Montague (1969), Loess et al. (1969) and Hunt & Hodge (1971)

<table>
<thead>
<tr>
<th></th>
<th>CBW</th>
<th>LBC</th>
<th>HH</th>
</tr>
</thead>
<tbody>
<tr>
<td>BM</td>
<td>0.81*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HH</td>
<td>0.91*</td>
<td>0.84*</td>
<td></td>
</tr>
<tr>
<td>LBC</td>
<td>0.90*</td>
<td></td>
<td>0.74*</td>
</tr>
</tbody>
</table>

* P < 0.01.

The question remains as to why Wilkins obtained the difference between the two groups of categories labelled ‘large’ and ‘small’ on the basis of his measure of size. As with any study of this kind, it is quite possible that some factor other than size was inadvertently manipulated and that this accounted for the effect attributed to size. As Morris points out, ‘the fact that ratings of size differ between the two groups such that large categories are rated as small greatly increases the chances of this being so’. This point is not in dispute. What is at issue is whether, as Morris claims, ratings provide a preferable index of category size, given that the concept of size is poorly defined.

The use of ratings

Morris bases his argument for the use of ratings on their reliability and utility in other areas of research. Unfortunately for this argument, the use of ratings can sometimes leave the experimenter ignorant of the criteria on which subjects perform their judgements, and this ignorance can easily lead to a circular operational definition without external validation. It is of little use to argue that categories with a high rating are ‘large’ if the only definition of size is that ‘large’ categories have a high rating. It is therefore critical to ask with what other measures of size ratings correlate. As Herrmann et al. (1973) showed (and as Morris confirms) ratings fail to correlate with the number of exemplars listed in the Battig & Montague (1969) category norms. On the other hand ratings do correlate significantly (r = 0.696) with another measure of category size: the number of exemplars of a category that subjects can list in 30 s (Battig & Montague, 1969). This measure of size, the initial rate of generation of exemplars, is one to which Morris takes exception. He implies that the rate of exemplar generation reflects not only the size of the category but also the accessibility of its exemplars. This may well be true. But it is rather unsatisfactory to advocate the use of ratings while condemning without further explanation a measure with which they correlate.

It is impossible to know on what basis subjects were performing the rating task but at first glance it appears that they based their rating on the ease with which a few of the most typical exemplars came to mind. And this despite the fact that Battig & Montague specifically instructed their subjects not to think of exemplars when making their ratings.

The concept of category size

As Rosch (1975) has pointed out, exemplars vary in the degree with which they fit a category. Ratings of the degree of category fit correlate highly with the accessibility of exemplars as determined by the frequency with which they are produced in a category production task (Caramazza et al. 1976). Exemplars with a high production frequency are emitted earlier (Murray, 1975) and are categorized more quickly (Wilkins, 1971). Evidently there are exemplars of any category that are rated as better exemplars than others, that are categorized more quickly, and that are emitted earlier and more frequently in a category production task. The number of such ‘good’ exemplars that a category possesses might very well be expected to influence ratings of the size of the category and also the number of exemplars of that category.
that subjects could think of in a limited period of time. But the number of such exemplars would not conventionally be regarded as the size of that category.

What, then, is category size? Obviously as one considers exemplars with a poorer category-fit the size of the category increases: the size of a category and the degree of fit of its exemplars are not independent concepts. The relationship between size and degree of fit can be represented by a Zipf function (Zipf, 1949). A hypothetical example is shown in Fig. 1. For a given category the degree of fit of an exemplar is plotted against the rank of that degree of fit so that the value $x$ represents the number of exemplars having a degree of fit greater than or equal to the corresponding criterion, $y$. The shape of a Zipf function of this kind shows how the size of a category ($x$) fluctuates with changes in the criterion of degree of fit necessary for category membership ($y$). Figure 2 shows empirical functions obtained using Rosch's (1975) ratings of category fit for exemplars of four categories listed in the Battig & Montague (1969) norms. Although the scope of Rosch's data is limited it is sufficient to demonstrate that different categories have differently shaped functions which intersect, indicating that the relative size of categories varies with level of criterion.

The level of criterion is arbitrary. Measures of size which may be assumed to have a high criterion of degree of fit, such as the mean number of exemplars emitted in a limited time (Freedman & Loftus, 1971), should not necessarily correlate with measures of size with a low criterion. The total number of exemplars listed in category norms (Wilkins, 1971) may be assumed to be a measure with a low criterion since many exemplars with poor category fit are included. Herrman et al. (1973) in their factor analysis of the Battig & Montague norms obtained evidence for a separation of measures much along these lines. Their description of the factors I and II as size and accessibility may be rephrased in terms of measures with low and high criteria of fit.

The range in the shapes of the Zipf functions shown in Fig. 2 indicate that category size can have little meaning except in relation to the degree of category fit of exemplars. It follows that attempts to distinguish between exemplar search models of categorization (Landauer & Meyer, 1972) and feature comparison models (Sheaffer & Wallace, 1969; Smith et al. 1974) on the basis of the effects of category size will continue to be fruitless until such time as assumptions about the models' representation of category fit are made sufficiently explicit for any role of size to be determined.
Figure 2. Empirical Zipf functions obtained using Rosch's (1975) ratings of category fit for exemplars of four categories listed in the Battig & Montague (1969) norms.

References


Received 1 November 1977

Requests for reprints should be addressed to Dr A. J. Wilkins, MRC Applied Psychology Unit, 15 Chaucer Road, Cambridge CB2 2EF.