THE BASIS OF NORMS FOR VAGUE QUANTIFIERS

Colm O'Muircheartaigh and George Gaskell
Cognitive Survey Laboratory, London School of Economics and Political Science, London, England

Key Words: vague quantifiers, response function, norms

Introduction

The underlying assumption in survey measurement is that the survey question and its associated response alternatives (implicit or explicit) have the same meaning to all respondents. It is only in this case that the basic requirement of successful measurement can be satisfied; that is, that differences between the scores obtained for different individuals reflect differences in the attribute or behaviour being measured for the individuals.

A common type of question in surveys is one in which the respondent is asked to report an estimated amount or frequency. An example from the British Crime Survey of 1988 is given below.

"When dealing with people they suspect of crimes, would you say that the police in this area ever break the rules?

(if YES) Would you say this happens ...

Hardly ever
Sometimes
Fairly often
Very often"

A widely used question in health surveys in the United States is

"All in all how would rate your health on the following scale

Excellent
Very good
Good
Fair
Poor"

The response categories for the second part of the first question are described as *vague quantifiers*, as there is neither a formal or an informal definition given for the meaning of the terms. The response categories for the second question are analogous in that they invite the respondent to make a comparative or relative judgement without giving any benchmarks to anchor the responses. In this paper we investigate whether individuals differ in their interpretation of such terms, and whether any observed differences are systematic in nature.

Absolute Versus Relative Frequencies

Questions that ask respondents for an absolute (numerical) report of a behavioural occurrence may present the respondent with a difficult cognitive task. The question "How many hours of television did you watch in the past week?", for example, demands considerable cognitive effort on the part of the respondent. To answer such a question, respondents must count the number of hours, or episodes, estimate a frequency using some rule based method, or average in some way over the events that are accessible in memory (Blair and Burton, 1987). Bradburn and Miles (1979) suggested that in the case of low salience events this imposes an excessive burden. In such cases, indeed, respondents may simply not store precise information in memory about events that they were asked to recall (Bradburn and Danis, 1984).

When absolute amounts are sought, the designer may choose to offer a closed set of mutually exclusive response categories to the respondent. The choice of alternatives has been shown to affect the responses to the question in many situations (Schwarz and Hippler, 1987, 1991, Gaskell, O'Muircheartaigh and Wright, 1994). For some cases the nature of the scale may make the definition of absolute amounts problematic; the example from the British Crime Survey is a case in point - it is not clear that is possible to rephrase the question to measure absolute frequencies.

For these reasons absolute numbers may be subject to errors and alternative measures may be needed. One option is to present the response alternatives as relative rather than absolute amounts, as recommended by Bradburn and Danis (1984).

The basic problem with *relative* categories as response alternatives is that in order to be useful they need to be *vague* - that is to say they must not **require** the computation of amounts or the enumeration of episodes. As vague quantifiers they have *fuzzy* boundaries (Hersh and Caramazza, 1976). Consequently the measurement process will now include a stage in which a mapping is carried out between the 'true score' and the corresponding position on the scale of vague quantifiers. It may be, of course, that for some behaviours respondents carry an accessible answer to a question about their relative position that does not require on-line computation of the absolute score or an approximation to it. Even if the respondent does not carry out this process on being asked the question, however, there is an implicit mapping for the population between the absolute and vague quantities, and this

mapping is described by means of a *response function* (Saris, 1988).

The usefulness of a scale of vague quantifiers depends on having stable calibration within and between individuals and across situations. There is evidence that even in within-subject studies there can be instability (Hammerton, 1976, Moxey and Sanford, 1993). For between-subject studies and for comparisons across situations, it is difficult to establish any satisfactorily stable calibration for vague quantifiers in general (Goodwin, Thomas and Hartley, 1977). Even more worrying, *different* conclusions may be implied for group comparisons for absolute and relative quantifiers apparently representing the same variable or dimension (Schaeffer, 1991).

Theoretical Framework

We assume that there is a mapping between the scale score and the 'true score' for an individual. The scale is usually presented to the respondent as a set of verbal categories; where there are scores attached they are usually the natural numbers 1, 2, 3, ... We have no way of knowing the precise form of the mapping from the absolute values which would represent the true scores to the scale scores. The mapping, which may be linear - but is much more likely to be non-linear - is called the response function. Previous research has shown that the presentation of the scale conveys information to the respondent. A key issue is the use made of this information by the respondents. Schwarz and Hippler (1991) suggest that respondents assume that the scale reflects a *sensible* distribution; scores in the middle indicate average amounts for the population. Offering a high frequency set rather than a low frequency set of response alternatives, for example, can change either the meaning of the scale (a meaning shift) or the respondent's view of his/her relative position (a comparison shift).

With reference to vague quantifiers, how does the respondent arrive at a mental image of the continuum under consideration: what norms or comparison groups inform the respondent judgment? Kahneman and Miller (1986) argue that the respondent does not have a ready-made set of reference points, but that, following the question, norms are constructed ad hoc by recalling relevant exemplars. Thus thinking, in this situation, flows backward from an experience or question to what that experience activates in memory. What is activated in a particular comparison will be affected by context and background. When a respondent in a survey is asked to make a category centred comparison a key issue concerns the norms used to make the judgment. No exemplar model can account for all variants of category knowledge; thus there is 'inheritance' of knowledge from 'comparable' categories.

Normality has two particular aspects in this context. First it is related to *typicality* - this is modelled as a similarity relation and in some ways mirrors the use of the mode as a measure of location or average. The second feature is *representativeness* - which incorporates the rest of the distribution of values and is related to the range or dispersion of values (the higher order moments of the distribution) in statistical description. Schwarz and Hippler (1991) argue that the norm is thought to be near the middle of the scale (thus the effect of the response alternative set) and the rest of the scale is interpreted relative to this (Saris, 1988).

What Norms?

If the survey question and its associated response alternatives have the same meaning to all respondents, this would imply that the norm generated by an individual should reflect only the individual's own position and general information about the population. This we call the *universal norm* hypothesis. If this holds, then there is comparability of scores between individuals, and also between groups of individuals.

Kahneman and Miller's theory suggests that accessibility is a key element in determining the exemplars activated by a question. We argue that there is a strong presumption that the accessible exemplars will be different for different individuals.

A long tradition of theorising in social psychology supports this contention. For example, relative deprivation theory is based on the notion that people's sense of dissatisfaction arises not from the absolute amount of some desired attribute but from the amount people have in comparison to others. Among others, Stouffer and Festinger elaborated a theory of social comparison based on the premise that people are motivated to obtain an accurate appraisal of their attitudes and opinions. In some circumstances this can be based on objective criteria, such as one's golf handicap or a score on test. However, where an objective standard is not available people turn to social sources; thus social comparison involves contrasting one's performance or opinions with those of other people. Festinger also specified that "given a range of possible persons for comparison, someone close to one's own ability or opinion will be chosen for comparison". Thus if a student of physics wants to assess his progress he is more likely to compare himself to other students rather than to Einstein. Developments of this line of theorising can be seen in Thibaut and Kelly, Pettigrew, and Upshaw.

Two issues however remain unresolved; which comparison groups do people select and are such comparisons explicit or implicit? Most theorists agree that more familiar subgroups rather than less familiar ones will be chosen, but no theory postulates a priori which out of a range of familiar groups will be the basis of comparison. Secondly, while some theories propose that people make explicit comparison against selected others, Upshaw, and Thibaut and Kelly, argue in favour of internalized standards based on experience. Judgments are made without direct reference to social reality. Thus, Thibaut and Kelly suggest that reference is made to "some modal or average value of all the outcomes known to the person by virtue of personal or vicarious experience". These two positions are not mutually exclusive; the internalized standards that people use in everyday life develop out of repeated experiences with relevant others. Kahnemann and Miller can be seen to offer a synthetic and parsimonious solution to this contrast; their distinction between stimulus and category norms corresponds to two positions in social comparison theory; category norms are those which involve accessing exemplars of types or groups, while stimulus norms appear to resemble the internalized personal standards.

This in itself would not be particularly important for statistical analysis if the variability were unsystematic and unrelated to other characteristics of interest of the individuals. Were this the case the impact would be simply to increase the error variance for the overall estimate and for comparisons. To the extent that there is a pattern or a systematic element to the accessibility of exemplars, or to the extent that different groups in the population generate different norms for the continuum, this could invalidate comparisons between groups. Where differences in scores are related not just to differences in the attribute or behaviour being measured, but also to the values of that attribute or behaviour for social or other groups to which the individuals belong, then scores for different individuals may not be comparable, and furthermore group means may also not be comparable.

If respondents use a group identification to generate exemplars, or if members of the same group tend to generate exemplars that are similar to those for others in the group but different from those for respondents outside the group, then the conditions for the *universal norm* are not satisfied. In this case we define the *segmented (stratified) norm* hypothesis as an alternative description of the response framework. To the extent that categories inherit properties from other categories, we might expect more within group consistency across topics than the 'objective' reality might suggest.

Issues

We are concerned with three issues here. The first is **calibration** - the extent to which vague quantifiers are *fuzzy*. The second is the *norm* - the construction and determination of the reference points used by the respondent in locating him/herself on the dimension of interest. The third is the *context* – the stability or otherwise of the norm for the same set of terms (vague quantifiers) across topics. These three issues can be seen also as indicators of different elements of the quality of the measurement in statistical terms. Calibration is related to within category variability; the norm is related to cross-category/within topic validity (expectation/bias); context is related to cross-category/cross-topic validity (bias). An alternative interpretation can be found in terms of the partitioning of the total variance of the observation (true variance, error variance, instrumental variance, method variance etc).

Data/Experimental Designs

All our experiments were embedded in BMRBI omnibus surveys. A national sample of n=1028 adults was interviewed; each respondent was asked two questions, one of which requested a self-report of hours spent watching TV on an average weekday, the other asked for a judgment of the hours spent by the *typical* person watching TV on an average weekday. Thus this first experiment explores the first aspect of normality described above - *typicality*. Three results emerged from the analysis. First, in an experiment integrated with this but not central to our concern here, we confirmed that the choice of the closed set of alternatives can have a substantial effect on the distribution of responses in this case. The order of magnitude of the difference in the distributions (20% for the proportion of respondents stating that they watch TV 2½ hours or more) is in line with the earlier work by Schwarz on TV questions and much stronger than the effect we found (Gaskell, O'Muircheartaigh, and Wright, 1994) for frequencies of vague events. In relation to our concerns here, the main implication is in demonstrating the vulnerability of absolute or numerical quantifiers.

Second, there was a positive relationship between the level of watching reported by individuals for themselves and the estimates by individuals of the amount of TV watched by the typical person. This is entirely predictable in terms of norm theory, as the most accessible exemplars are likely to be one's own experience.

Third, it is possible to carry out an imperfect comparison of the two hypotheses. Under the universal

norm hypothesis the determinants of 'typical' will consist of the individual's own level of viewing together with shared general information. We can identify the socio-demographic variables by which actual TV viewing varies, using as our measure the reported 'self' viewing hours. Next to one's own behaviour, it is plausible to expect that the next most accessible exemplars would be those from one's own socio-demographic group. Assuming full information, the segmented norm hypothesis would predict that the estimate of the 'typical' level of viewing would also vary by that factor/variable, and in the same direction. Thus if the prediction of the 'typical' viewing improves by including these socio-demographic terms in the model, this would support the segmented against the uniform norms hypothesis.

The data show that indeed it does. Controlling (in the analysis) for the individual's own score, individuals in socio-demographic groups that (on average) watch more TV tend to think the typical person watches more. We appreciate that there are approximations involved here. First, we are using socio-demographic groups as proxies for the reference groups of the individuals. We do not imply that this relatively crude classification corresponds exactly to the group with whom the individual would actually identify. We suggest, rather, that the socio-demographic class provides a better approximation to the individual's reference group than does the population as a whole. Second, we cannot assume that an individual **knows** the characteristics of his/her reference group - to a greater or lesser extent there may be a difference between the perceived and actual scores for the individual's reference group.

This experiment addresses one aspect of the *normality* of a scale - typicality, in particular the creation of a 'typical' or average value. In the experiment the self score and the 'typical' rating were provided by the same individuals. The results establish that there is a connection, for this variable, between the level of an individual's score and the level of score he/she perceives to be typical in the population; this is consistent with norm theory as the individual's own behaviour almost certainly provides the most accessible exemplars when constructing an answer to a question about a particular behaviour. By comparing subgroups of the population, we also examined how group membership affects the estimation of 'typical' values. We found that, independently of (controlling for) the individual's own score, group membership has an effect on the assessment of what the individual considers to be a 'typical' score.

In our second and third experiments we shift our focus to the second feature/aspect of *normality* – representativeness, in particular the choice by the respondent of one of a range of vague quantifiers to represent his/her position on a continuum. This choice combines the determination of a 'typical' value and the mapping of the absolute or numerical scale onto the verbal scale.

In experiment 2 a national sample of n=1106 adults was interviewed; each respondent was asked two questions about amount of TV viewing; the first asked the respondent to place him/herself on the five-point vague quantifier scale (none at all, hardly any, a little, quite a bit, a lot); the second asked, in open form, for a numerical value for number of hours watched. Experiment 3 comprised a national sample of n=1999 adults; each respondent was asked four questions; two questions dealt with amount of alcohol consumed, and had the same format exactly as experiment 2; the last two questions dealt with frequency of alcohol consumption, had the similar format to experiment 2 but used vague quantifiers appropriate for a frequency (never, rarely, occasionally, sometimes, frequently). The order of the questions in experiment 3 was randomised both by topic(amount/frequency) and by order within topic (verbal/numerical).

By contrast with experiment 1 the respondents in experiments 2 and 3 were not asked to evaluate the meaning of the verbal quantifiers. Rather, each individual was asked to rate *him/herself* using *both* the numerical and verbal quantifiers. Thus the kinds of analysis possible with these data are different from that carried out on the data from experiment 1. Below we address the three issues of calibration, norms, and contexts in turn.

Calibration

Three conclusions may be reached: (i) none/never very nearly means 0.0 – there are only 6 cases where a non-zero numeric answer corresponded to a verbal none/never; (ii) overall the calibration is respectable; the middle 50% of cases is non-overlapping for neighbouring categories, except for the pair occasionally/sometimes. In retrospect, these may have been poor choices as verbal quantifiers in the same scale, as they may well not convey different impressions. Indeed, the difference between them may be seen as a measure of scale position effect rather than meaning effect; (iii) the same result holds within identifiable socio-demographic groups. In general groups with a higher range of reported behaviour on the numeric/absolute quantifiers show better discrimination between categories – not unreasonable given the crudeness of the classification scale.

Comparisons across groups may be considered an element of norm formulation rather than simply calibration.

Differences Between Groups

First we consider the extent to which different groups of individuals in the population differ with respect to the absolute/numeric scores; here we use their self-reported scores as a measure of their actual scores, though we realise that these scores are subject to error. We use the three standard socio-demographic variables – age, gender, and social class – as proxies for group membership. Reported TV watching differs significantly by age and social class, but *not* by gender. The reports of amount of alcohol consumed differ significantly by gender and (possibly) age, but *not* by social class. The reports of frequency of alcohol consumption differ significantly by *all three* variables.

Consider now the implications of these results in terms of the alternative hypotheses that have been put forward to predict the formulation of norms for vague quantifiers. Under the universal norm hypothesis, the mean absolute or numerical value corresponding a particular vague quantifier should be the same regardless of the **level** of behaviour (amount/frequency etc) of the members of the group. Under the segmented group hypothesis the mean value corresponding to a particular vague quantifier should be higher for members of a group that has itself relatively high levels of this behaviour, and should be lower for members of a group for whom the behaviour is relatively light or infrequent. There are two elements to this; first, respondents base their judgment of the relative values of the positions on the scale on accessible exemplars (which may be inherited from another scale with which the respondent is more familiar); second, respondents will tend to have more readily accessible information and exemplars belonging to their own socio-demographic groups. Therefore we predict that differences in the actual behaviour of a socio-demographic groups will be reflected in the assessments of the meaning of scale points when evaluated by members of the group. This is an extension of the argument put forward, and supported by the data, in the first experiment, where it was demonstrated that the assessment of the *typical* was correlated not just with the individual's absolute/numeric value, but additionally with the absolute/numeric values of the group to which the individual belonged.

The method of analysis we have chosen for this purpose requires a number of simplifying assumptions. In a later paper we intend to relax these assumptions. We have chosen to carry out an analysis of covariance [ANOCOVA] where the dependent variable is the vague quantifier, treated as an interval scale variable scored from 0 to 4 (1 to 5). The socio-demographic variables are included as factors in the model (each treated as ordinal scale, since e do not assume a linear, nor even ordinal, relationship even between age and level of the dependent variable). The covariate is the absolute/numeric value obtained from the absolute quantifier question.

Simply examining whether different groups in the population have a different distribution of the vague quantifiers does not answer the general question as to whether the vague quantifiers have different meanings for different groups. The analysis proceeds by taking into account first the different distribution of absolute levels of behaviour for the different groups. This is done by including the covariate first in the predictive model; this, in effect, corrects the rest of the analysis for differences in levels of behaviour between groups. If the universal norm applies, then no other factors will improve the prediction of the allocation of an individual to a particular value of the vague quantifier.

Subsequently the socio-demographic factors are examined to see whether any or all of them – together or in combination – improve the prediction. To the extent that they do, this suggests an element of segmentation in the construction of norms. To the extent that the influential factors correspond to the group levels on the absolute/numeric measures, the more convincing is the case for accessibility/identification as a basis for the segmentation.

For TV watching, there is a very strong correlation between the answers to the absolute amount and vague quantifier questions. This is encouraging but not surprising. The results of experiment 1 indicate that the individual's own score affects the assessment of the 'typical'; had the individual's own score been the only input to the formulation of norms this could have meant that every individual would consider him/herself 'typical'. This result shows that in addition to the input from their own absolute/numeric score individuals use other information that enables them to position themselves on the continuum.

Having taken into account the individual's own score, we can now test whether the formulation of norms can be explained by the reference group of the individual. Conditioning in the effect of the individual's own absolute score, we added the three factors age, gender, and social class into the model to see whether they improved the prediction of the individual's position on the vague quantifier scale. Gender was not significant; both age and social class were significant, as was the interaction between them. When a factor makes a contribution to the explanatory power of the model, this means, in effect, that having used the individual's

score on the absolute scale to predict his/he position on the vague quantifier scale, knowing the value (score) of an individual on that factor improves our prediction of the position that individual will choose on the vague quantifier scale.

For amount of alcohol consumed, there is once again a strong positive correlation between the absolute score and the vague quantifier score. When the additional factors are added into the model all three factors make a significant contribution to the quality of the prediction. With frequency of alcohol consumption age makes a significant contribution, gender a marginal contribution, and social class is not significant.

There are some technical limitations to the analysis carried out here. First, the analysis assumes that the vague quantifier is measured on an interval scale, whereas it is in fact ordinal. Some checks on the sensitivity of the results to this assumption (using non-linear transformations of the scale) suggest that they are robust. Second, there may be boundary problems for the prediction given that quantities cannot be less than zero; the effect in our case is probably slight. Third, we have not allowed for interaction s between the covariate and the predictors. In this situation such interactions would correspond to a situation in which not only the level of absolute score that corresponds to a vague quantifier score would vary according to the group to which the individual belonged or related but the differences between groups would also be affected by group membership. In technical terms this would mean a difference in slopes for the regression of the dependent variable on the absolute score. In later work we intend to tackle these problems by using ordered probit analysis, possibly stratified.

References

Blair, E., and Burton, S. (1987). Cognitive processes used by survey respondents to answer behavioral frequency questions. *Journal of Consumer Research* **14**, 280-288.

Bradburn, N.M., and Danis, C. (1984). Potential contributions of cognitive research to survey questionnaire design. In *Cognitive Aspects of Survey Methodology: Building a Bridge between Disciplines*, ed. Thomas Jabine, Maron Straf, Judith Tanur, and Roger Tourangeau, 101-129, Washington, DC: National Academy Press.

Bradburn, N.M., and Miles, C. (1979). Vague Quantifiers. *Public Opinion Quarterly*, **43**, 92-101.

Gaskell, G.D., O'Muircheartaigh, C.A., and Wright, D.B. (1994). Survey questions about the frequency of vaguely defined events: The effects of response alternatives. *Public Opinion Quarterly* (in press).

Goodwin, A.R., Thomas, S., and Hartley, J. (1977). Are some parts larger than others? Qualifying Hammerton's quantifiers. *Applied Ergonomics*, **8**, 93-95.

Hammerton, M. (1976). How much is a large part? *Applied Ergonomics*, **7**, 10-12.

Hersh, H.M., and Caramazza, A. (1976). A fuzzy set approach to modifiers and vagueness in natural language. *Journal of Experimental Psychology: General*, **105**, 254-276.

Kahneman, D., and Miller, D.T. (1986). Norm theory: Comparing reality to its alternatives. *Psychological Review*, **93**, 136-153.

Moxey, L.M., and Sanford, A.J. (1993). *Communicating Quantities*. Hove, England: Lawrence Erlbaum Associates.

Saris, W.E. (1988). *Variations in response functions: A source of measurement error in attitude research.* Sociometric Research Foundation: Amsterdam.

Schaeffer, N.C. (1991). Hardly every or constantly? Group comparisons using vague quantifiers. *Public Opinion Quarterly*, **55**, 395-423.

Schwarz, N., and Hippler, H.J. (1987). What response scales may tell your respondents: informative functions or response alternatives. In *Social Information Processing and Survey Methodology*, ed. Hans-J. Hippler, Norbert Schwarz, and Seymour Sudman, 163-178. New York: Spinger-Verlag.

Schwarz, N., and Hippler, H.J. (1991). Response alternatives: The impact of their choice and presentation order. In P.P. Biemer, R.M. Groves, L.E. Lyberg, N.A. Mathiowetz, S. Sudman (Eds.), *Measurement Errors in Surveys*, 41-56, New York: Wiley and Sons.