Can Ordinal Utility Exist?

S. Kemp[1], and R. C. Grace[1]

[1]Psychology Department, University of Canterbury, Christchurch, New Zealand.

Keywords: *Ordinal utility, measurement, sorting procedures*.

EXTENDED ABSTRACT

It is often assumed that the measurement of utility attains the status of an ordinal but not of an interval scale. If utility arises from integrating information from different dimensions or attributes and trade-offs are permitted, such utility satisfies either interval scale status or only weak (often very weak) ordering can be attained. If, on the other hand, utility is regarded as determined behavioural from preference orders, it is very difficult to rank the goods without resorting to ratings based on interval scales unless the number of items is small. The combination of these two considerations should lead us to question seriously whether in practice ordinal utility is attainable unless interval scale status is also attainable.

1. INTRODUCTION

A common view among economists and psychologists alike is that constructs such as utility which depend on multiple attributes or dimensions can only be measured on ordinal scales. We take issue with this view, and demonstrate that, in most situations it seems unlikely that utility could satisfy the assumptions necessary for ordinal scaling without also satisfying those for interval scaling. If so, the use of statistical methods based on the more restrictive assumptions of ordinal scaling may be questioned. The concerns we raise are common to the disciplines of psychology and economics, but as psychologists we use the (slightly different) terminology of psychology.

2. SCALE TYPES.

The measurement of variables that are essentially subjective in nature has long been a concern of both psychologists and economists. Stevens (1946, 1955) identified four different scale types: nominal, ordinal, interval and ratio, and subsequent work in representational measurement theory (e.g. Luce, 1996; Narens, 2002) has formally elaborated these distinctions, with a particular focus on the difference between ordinal and interval scales. An ordinal scale is one in which it is possible to rank order items on some attribute, and to make such statements as "*a* is greater than *b*" (i.e. *a* has more of the attribute than *b*), but it is not generally possible to say how much greater *a* is than *b*. However, if items are measured on an interval scale, it is possible also to compute the difference between *a* and *b*, and, for example, compare this difference to the difference between *b* and *c*. Thus interval scales may be viewed as ordinal scales with an extra element of precision. Conventional examples of ordinal scales include preferences for goods or bundles of goods. Interval scales include measures of distance, temperature, and money. Many interval scales, like distance, also satisfy the requirements to be ratio (cardinal) scales. It is often believed that many scales that attempt to measure subjective dimensions attain ordinal but not interval scale status. Hence, for example, the frequent recommendation in both psychology and economics to use ordinal scale statistical methods when, say, the dependent measures are rating scales.

An important aspect of ordinal scaling is the proportion of element pairs that may be ordered. Some feature total ordering. That is, for all the pairs of elements it is possible to say that one is greater than the other. In weak ordering, some pairs cannot be ordered. In some definitions (e.g. Narens, 2002, ch. 5), it is possible for a set to be weakly ordered even if no pair of elements can be ordered. Of course, a definition of ordinal utility which allowed that no preference order could be established between any pairs of goods, while theoretically feasible, would be useless in practice. To be useful, an ordinal utility scale would require that a nontrivial proportion of element pairs can be ordered.

3. UTILITY AS AN INTERNAL CONSTRUCT

Broadly speaking there are two ways we can view utility, either as an internal construct or as a measurable variable. We consider first utility as an internal construct.

In general, individual or household utility is taken to be a function of a number of input variables. For example,

$$U = U(Z_1, Z_2, \ldots. Z_m)$$

where the Z_i may be taken as different commodities or bundles of commodities or attributes (whether internal or external). Typically, it is allowed that deficiencies in, say, Z_p can be compensated for by corresponding increases in the other Z_i.

Now, given this relationship, what measurement scale does U satisfy?

It is fairly clear that if the Z_i are measured on interval scales, then, given the nature of the functional relationship, then U can also be measurable on an interval scale

But what happens if at least some of the Z_i are measured on ordinal scales with total ordering? The general answer is that U is then only weakly ordered. Moreover, the number or proportion of pairs of elements that can be ordered will decrease with the number of Z_i that are measured ordinally, and increase as the correlation between the Z_i increases. To give a single example, if m = 2, and the correlation between the Z_i is zero, one half of the element pairs can be ordered.

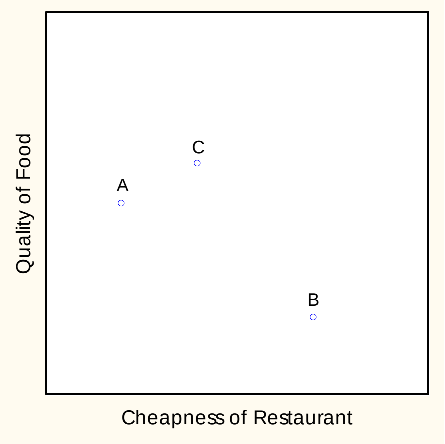

Figure 1. Three points on the two-dimensional space of Quality of Food and Cheapness of Restaurant. A choice or comparison between A and B on overall restaurant value requires a trade-off, but not that between A and C. An unambiguous trade-off is not possible if the dimensions only satisfy ordinal scale assumptions.

The reason is not hard to see. Consider we are trying to determine the utility of different restaurants to dine at this evening (see Figure 1), which are characterized in terms of quality and cheapness. Generally, regardless of the scale type, pairs of restaurants can always be ordered if one is superior to the other on both quality and cheapness. If both quality and cheapness (or, indeed, only quality) are measured on ordinal scales, no ordering is possible when, say, Andrew's restaurant is cheaper than Brian's, but Brian's has the better food (Wakker, 1989).

As the correlation between the variables increases (so for example if there was a tendency for cheaper restaurants to also have better food), so does the number of orderable pairs. For example, if the correlation were 1, all the pairs could be ordered (because the cheaper restaurant always has the better food). If $\rho = -1$, none of them can be unambiguously ordered, because we are always faced with a trade off between cheapness and quality. Inclusion of additional dimensions (e.g. how close the restaurant is to home) will generally increase the proportion of pairs that cannot be ordered.

The overall conclusion, then, for utility regarded as an internal functional construct is that if the Z_i satisfy interval scale assumptions then U will too. However, if the Z_i only satisfy (total) ordinal assumptions, U will not satisfy (total) ordinal assumptions. Thus, in general either U is interval or it is only weakly ordered. Moreover, in most realistic situations in which utility is determined by several dimensions, the dimensions will not all positively correlated – for example, restaurants with great food are often not particularly cheap - and thus U is likely to be weakly ordered to the point where very little ordering information is available.

One important caveat should be noted. It is often quite feasible to establish ordinal scales with total ordering when the possibility of trade-offs is eliminated. For example, choices may be made by considering the different Zi dimensions sequentially. Ranking proceeds on, say, Z_p. Ties on Z_p are then resolved by considering Z_q, and so on. It is easy to see that this process may produce total ordering. However, trade-offs are not permitted. For example, if one obtains the lowest rank on Z_p, no amount of excellence on the other Z_i can compensate. Incidentally, such an ordering system, particularly when the Z_i can be ordered in terms of discrimination power, often produces "good" choices, and is probably frequently used in practice (see, e.g., Brandstätter, Gigerenzer, & Hertwig, 2006; Gigerenzer & Goldstein, 1996; Gigerenzer & Selten, 2001; Hutchinson & Gigerenzer, 2005). A variety of other strategies are available that also discount trade-offs. For example, one could simply count the number of dimensions in which one choice is preferable to another.

4. UTILITY AS MEASURED FROM OBSERVED PREFERENCES.

A quite different approach to utility is to consider it as defined from observed preferences. Clearly, if a total (or near total) ordering of goods can be achieved and confirmed, then we are justified in assuming a behaviourally-derived measure of ordinal utility. Furthermore there would be neither need nor justification for assuming that this utility measure has interval scale status.

We do not deny that such scales may be obtainable. Rather we contend that they are difficult to obtain and, in consequence, rather rare.

Some indication of the difficulties can be gained from considering the way in which players are ranked in different sports. For example, the United States Tennis Association (2006) recommends a rating scheme ranging from 1.0 (“This player is just starting to play tennis”) to 7.0 (international tournament player) in 0.5 steps for tennis players. Ratings may be made on six specific characteristics (forehand, backhand, serve/return of serve, volley, special shots, and playing style), each scored on the same 1 to 7 scale. Both the overall and the specific ratings are probably better described as ordinal rather than interval scales: Certainly there is no presumption in the Guidebook that, say, the interval between 1.0 and 2.0 is equal to that between 6.0 and 7.0. It is also clearly recognised that a 3.0 level player will not necessarily be at a 3.0 level on all six specific characteristics. How then should the player who scores a mixture of 2.5s and 3.5s be rated overall against a player who scores all 3.0s? No attempt at constructing a function is suggested (cf Section 2 above). Instead, the emphasis is on what actually happens when the players meet: “A player’s competitive record is the best test of his or her rating.” (United States Tennis Association, 2006, p. 11).

However, a number of games must be played to establish this ranking. Within a small club this is unlikely to be problematic, because the players enjoy playing tennis! But when the number of elements becomes large, the number of comparisons increases rapidly. Research on sorting algorithms in computer science (for a standard review, see Knuth, 1998) has shown that depending on the type of sort used, the number of comparisons required to order n objects varies according to a proportion of between n^2 and n log n.

There is an additional complication that may be more important to us than to the computer scientist. Different comparisons are not normally carried out at the same time. Thus, for example, if when they last played tennis, Gail beat Paula, and Paula beat Sally, we would not necessarily expect Gail to now beat Sally if, in the meantime, Sally has been practicing a great deal while Gail has been lying on a couch recovering from a broken leg. Similarly, if an individual is making repeated choices or comparisons, time and memory play an important role.

Consider that Bruce is seeking to buy an artistic photograph to hang on his living room wall. He goes around various galleries in town looking at different options. In this case, the comparisons will generally involve at least some degree of memorising the different attributes of the photographs. It is not easy to establish a preference order under such conditions.

In general, psychological research indicates that people find ranking of any reasonable number of items (more than 10, say) quite difficult. Ranking items is reported to be harder and requires more time than rating, a process which involves assigning numbers to the items and then treating these numbers as having interval scale properties (Alwin & Krosnick, 1985; Rankin & Grube, 1980; Russell & Gray, 1994).

Kemp, Grace & Clark (2007) carried out two experiments to investigate how people would go about assessing the beauty of artistic photographs. In the first experiment, 26 respondents were asked to arrive at an eventual rank order of the artistic merit of 20 photographs. The photographs could only be viewed one by one and repeat viewings were not permitted. The majority of the participants (16) chose to do this after first using either grades or categories of excellence of the photographs. Eight used ranking measures from the start and two used a mixed strategy.

In a second experiment, 40 respondents were asked simply to assess the merit of the same 20 artistic photographs. Thus, it was open to them whether they attempted a ranking or used a category or grading (or indeed any other) system. Half of the respondents viewed the photographs serially (as in the first experiment). The other half could simultaneously view small versions of all the photographs as an array (of thumbnail sketches) and enlarge individual photographs whenever they chose. In both serial and array conditions, 15 (of the 20) participants graded the photographs out of either 10 or 20. The remaining 5 participants in the serial condition used a category rating scheme, as did 2 of those in the array condition. One participant in the array condition used letter mark grades (A+, A, etc), and the remaining 2 produced a rank order.

The results of these experiments underline the point that rank ordering stimuli is not a very natural thing to do, and moreover it is often accomplished, particularly when the stimuli are not all present, through a prior rating procedure that uses a conventional numerical interval scale. Overall, the results suggest that ranking is often a less basic psychological process than interval scale assessment.

In sum, although we note that in some circumstances it is possible for people to establish preference orders in a reasonably natural way, this is not usually an easy or automatic process. In fact, people frequently go about it by first assigning numbers to the items. These numbers inevitably assume interval scale properties (and indeed are often arrived at by the processes outlined in Section 2). This behaviour calls into question how often in practice ordinal scales of preference orders are established without making use of interval scale measurement.

5. IMPLICATIONS.

The reasoning above suggests that ordinal utility is difficult to attain unless interval-scaled utility is also attainable. However, it does not necessarily follow that we can simply assume interval-scale utility. One reason for caution is that there is a good deal of research showing that different preference orders are not always stable, and may sometimes be especially constructed to meet the demands of a particular task set by the researcher (e.g. Hsee, 2000; Slovic, 1975, 1991).

It seems to us that in the past psychologists and economists alike have frequently questioned whether measures such as utility really meet the assumptions necessary for interval or cardinal scale status. On finding reasonable grounds for doubting that these assumptions are met, they have then fallen back to the apparently safer position of assuming ordinal scale status. However, our reasoning suggests that this position too may often need questioning. It is actually quite difficult to find grounds for assuming that utility measures are ordinal if they are not also interval. A practical consequence is that researchers might consider moderating their enthusiasm for using statistical methods, for example regression models (e.g. Long, 1997), that assume dependent variables such as utility are measured on ordinal but not interval scales.

6. REFERENCES

Alwin, D. F., and J. A. Krosnick. (1985), The measurement of values in surveys: A comparison of ratings and rankings. *Public Opinion Quarterly,* 49, 535-552.

Brandstätter, E., G. Gigerenzer, and R. Hertwig. (2006), The priority heuristic: Making choices without tradeoffs. *Psychological Review*, 113, 409-432.

Gigerenzer, G., and D. G. Goldstein. (1996), Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, 103, 650-669.

Gigerenzer, G., and R. Selten. (Eds.) (2001), Bounded rationality: The adaptive toolbox. MIT Press, Cambridge, MA.

Hsee, C. K. (2000). Attribute evaluability: Its implications for joint-separate evaluation reversals and beyond. In D. Kahneman and A. Tversky (Eds.), Choices, values, and frames. Cambridge University Press, Cambridge UK (pp. 543-563).

Hutchinson, J. M. C., and G. Gigerenzer. (2005), Simple heuristics and rules of thumb, where psychologists and behavioural biologists might meet. *Behavioural Processes*, 69, 97-124.

Kemp, S., R. C. Grace, and A. Clark. (2007), How do people assess the beauty of photographs? Unpublished manuscript.

Knuth, D. E. (1998), The art of computer programming. Vol. 3. Sorting and Searching. 2nd Edn. Addison-Wesley, Reading, MA.

Long, J. S. (1997), Regression models for categorical and limited dependent variables. Sage, London.

Luce, R. D. (1996), The ongoing dialog between empirical science and measurement theory, *Journal of Mathematical Psychology,* 40, 78-98.

Narens, L. (2002), Theories of meaningfulness. Lawrence Erlbaum, Mahwah, NJ.

Rankin, W. L., and J. W. Grube. (1980), A comparison of ranking and rating procedures for value system measurement. *European Journal of Social Psychology*, 10, 233-246.

Russell, P. A., and C. D. Gray. (1994), Ranking or rating? Some data and their implications for the measurement of evaluative response. *British Journal of Psychology,* 85, 79-92.

Slovic, P. (1975), Choice between equally valued alternatives. *Journal of Experimental Psychology: Human Perception and Performance,* 1, 280-287.

Slovic, P. (1991), The construction of preference. *American Psychologist*, 50, 364-371.

Stevens, S. S. (1946), On the theory of scales of measurement. *Science*, 103, 677-680.

Stevens, S. S. (1955), On the averaging of data. *Science*, 121, 113-116.

United States Tennis Association. (2006), National Tennis Rating Program (NTRP) Guidebook. http://www.usta.com/leagues.

Wakker, P. P. (1989), Additive representations of preferences: A new foundation of decision analysis. Kluwer, Dordrecht.