APPLIED COGNITIVE PSYCHOLOGY, VOL. 5, 37–49 (1991)

Response Scales as Frames of Reference: The Impact of Frequency Range on Diagnostic Judgements

NORBERT SCHWARZ
Zentrum für Umfragen, Methoden und Analysen, ZUMA, Mannheim, FRG
HERBERT BLESS, GERD BOHNER
Universität Mannheim, FRG
UWE HARLACHER and MARGIT KELLENBENZ
Universität Heidelberg, FRG

SUMMARY

In social and psychological research, respondents are often asked to report the frequency of a behaviour by checking the appropriate alternative from a list of response categories provided to them. Previous research indicated that respondents extract comparison information from the range of the response alternatives, assuming that the average respondent is represented by values in the middle range of the scale, and that the extremes of the scale represent the extremes of the distribution. Extending this line of research, the present studies demonstrate that the users of a respondent's report are also likely to use the range of the response alternatives as a frame of reference in evaluating the implications of the report. Specifically, subjects are found to draw different conclusions about the respondent's personality (Experiment 1), or the severity of his or her medical condition (Experiment 2), from the same absolute frequency report, depending upon the range of the response scale on which the frequency was checked. Moreover, experienced medical doctors were as likely to be influenced by scale range as first-year medical students, suggesting that the phenomenon is of considerable applied importance. Implications for the use of response alternatives in psychological research and diagnostic judgement are discussed.

In psychological testing, as well as in laboratory experiments and survey research, respondents are often asked to report the frequency with which they engage in a certain beaviour or make a certain experience. To obtain the desired behavioural information, respondents are typically asked to check the appropriate alternative from a set of response categories provided to them. The selected alternative is assumed to inform the researcher about the respondent's behaviour. It is frequently overlooked, however, that a given set of response alternatives may be far more than a simple 'measurement device'. Rather, it may also constitute a source of information for the respondent, because respondents assume that the range of the response alternatives reflects the researcher's knowledge of, or expectations about, the distribution of the behaviour in the 'real world'. Specifically, they assume that the average behaviour is represented by response alternatives in the middle range of the scale and that the extremes of the scale reflect the extremes of the distribution (see Schwarz and Hippler, 1987; Schwarz, 1988, in press for reviews).

Accordingly, respondents were found to extract comparison information from the range of the response alternatives provided to them (Schwarz, Hippler,

0888–4080/91/010037–13$06.50

Received 4 April 1989

Deutsch and Strack, 1985; Scharz and Scheuring, 1988). Given the above assumptions, checking one from an ordered set of response alternatives may be considered as determining one's own location in a distribution, as the following example illustrates. Assume that some respondents are asked to report their average daily TV consumption on a scale ranging, in ½-hour steps, from 'up to ½ hour' to '2½ hours and more', while others receive a scale ranging from 'up to 2½ hours' to '4½ hours and more' (see Figure 1 for a similar example). Given an average TV consumption of 2 hours in the Federal Republic of Germany, West German respondents are likely to check a response category in the upper range of the low frequency scale which suggests to them that they watch *more* TV than is 'typical'. In contrast respondents who receive the high frequency range scale are likely to check a category in the lower range of that scale, suggesting to them that they watch *less* TV than is 'typical'. Accordingly, respondents who were given the low frequency scale evaluated TV to be more important in their own life (Schwarz *et al.*, 1985, Experiment 1), and reported lower satisfaction with the variety of things they do in their leisure time (Schwarz *et al.*, 1985, Experiment 2), than respondents who were given the high-frequency scale.

This and related research (Schwarz and Scheuring, 1988) illustrates that respondents use their own location on the scale to determine their location in the distribution. Thus, the range of response alternatives serves as a frame of reference that may affect respondents' subsequent judgements, either because respondents use the inferred 'average' behavioural frequency as a standard of comparison, as suggested above, or because they use the frequency range of the response scale to anchor subsequent rating scales, as suggested by Ostrom and Upshaw (1968).

However, the use of scale range as a frame of reference may not be restricted to respondents. Rather, the recipient of a respondent's behavioural report may also evaluate this report within the frame of reference suggested by the scale. If so, the conclusions drawn by a diagnostician, for example, may not only reflect the reported absolute frequency of the behaviour under study, but also the frequency range of the scale on which this report was provided.

The studies reported in the present paper were designed to explore this possibility in the domain of personality inferences (Experiment 1) and medical diagnosis (Experiment 2). In general, we expect that the recipients of a respondent's behavioural report will use the frequency range of the response scale as a frame of reference in evaluating the implications of the reported behaviour. Accordingly, they may be likely to draw different conclusions from the *same* behavioural frequency report as a function of the range of the scale on which this report is provided. The major goal of the present paper is to provide experimental tests of this hypothesis and to elaborate its applied implications.

However, much as the impact of response alternatives on respondents' own inferences was found to decrease as other relevant information becomes more accessible (Schwarz and Bienias, in press), we may expect that the impact of scale range on recipients' inferences decreases as the availability of other relevant information increases. Given that a number of different comparison standards may be used for any judgement (Schwarz and Scheuring, 1988; Schwarz and Strack, in press), the influence of the comparison information provided by the response scale should be attenuated when other potentially applicable comparison standards are temporarily or chronically highly accessible (Higgins, Strauman and Klein, 1986).

This additional hypothesis is tested in two ways. In Experiment 1 the cognitive accessibility of subjects' own behaviour is temporarily increased, and it is assumed

that subjects are less likely to use the response alternatives as a frame of reference under this condition. Experiment 2 extends this line of reasoning to the applied domain of medical decision-making, based on the assumption that relevant information is chronically more accessible to experts, who can draw on a rich base of experience, than to novices. If so, experienced diagnosticians should be less likely to rely on the frame of reference provided by the scale than novices.

EXPERIMENT 1

Experiment 1 was conducted as a modified replication of a study reported by Schwarz *et al.* (1985). As described above, subjects of the previous study were asked to report their own TV consumption on a high- or a low-frequency response scale, and the frequency range of the response alternatives was found to affect subsequent comparative judgements. In the present study, subjects were given a behavioural report provided by a target person on either a *high* or a *low* frequency range scale, and were asked to estimate how satisfied the target person is with the variety of things she does in her leisure time.

To provide a test of the hypothesis that the impact of scale range decreases as other potentially applicable comparison information becomes more accessible, subjects were asked to report their own TV consumption in an open-answer format either *before* or *after* they evaluated the target's leisure time satisfaction. These manipulations resulted in a 2 (low vs. high frequency range scale) × 2 (high vs. low accessibility of own behaviour) factorial between-subjects design.

It was expected that subjects would estimate the target's satisfaction with the variety of her leisure time activities to be higher when the report was given on the high frequency scale, suggesting that the target watches less TV than 'typical', than when it was given on the low frequency scale, suggesting that the target watches more TV than 'typical'. Moreover, the impact of scale range was expected to decrease when other comparison information was easily accessible. Accordingly, the impact of scale range was expected to be attenuated when subjects had previously reported their own TV consumption, thus increasing the accessibility of their own behaviour as a standard of comparison.

Method

Fifty-nine students (27 males and 32 females) of the University of Heidelberg, Federal Republic of Germany, were recruited individually in a university cafeteria for a study on 'impression formation', and were randomly assigned to conditions. They received a self-administered questionnaire in which a target person reported a daily TV consumption of '2 to 2½ hours', checked either on the high or the low frequency range scale shown in Figure 1.

In all experimental conditions the target person was described as a 28-year-old student. Before receiving the target person's behavioural report, subjects assigned to the 'high accessibility of own behaviour' condition reported their own TV consumption in an open-answer format. Subsequently, they estimated the target's leisure time satisfaction along an 11-point rating scale, with the end-points labelled 1 = 'very dissatisfied', 11 = 'very satisfied'. Subjects assigned to the 'low

accessibility of own behaviour' condition reported their own TV consumption after they had estimated the target person's leisure time satisfaction.

Low frequency scale	*High frequency scale*
( ) not at all	( ) up to 2 hours
( ) up to ½ hour	(X) 2 to 2½ hours
( ) ½ to 1 hour	( ) 2½ to 3 hours
( ) 1 to 1½ hours	( ) 3 to 3½ hours
( ) 1½ to 2 hours	( ) 3½ to 4 hours
(X) 2 to 2½ hours	( ) 4 to 4½ hours
( ) more than 2½ hours	( ) more than 4½ hours

Note. The target person's reported TV consumption is marked X.

Figure 1. Response alternatives for daily TV consumption

Results

Subjects' estimates of the target's satisfaction with her leisure time variety were analysed by a 2 (scale range) × 2 (accessibility of own behaviour) × 2 (sex) ANOVA. Because no effects of sex emerged (all $p > .30$), the reported data are pooled over this variable. As predicted, this analysis revealed a significant interaction effect of scale range and accessibilty of own behaviour, $F(1,55)= 8.83, p < .004$. Specifically, subjects who had *not* previously reported their own TV consumption estimated the target's leisure time satisfaction to be higher when her report was given on the high ($M = 5.3$) rather than the low ($M = 3.9$) frequency scale, $p < .05$, Duncan test. This effect replicates the previously obtained results (Schwarz *et al.*, 1985), indicating that the subjects used the frequency range of the scale as a frame of reference in making inferences about the target person, as was previously shown for respondents themselves.

In contrast, subjects who had previously reported their own TV consumption estimated the target's satisfaction with the variety of her leisure time activities to be higher when she gave her report on the low ($M = 6.4$) rather than high ($M = 4.5$) frequency scale, $p < .05$, Duncan test. This finding apparently contradicts our expectation that subjects would use their own behaviour as a standard of comparison under these conditions, which should eliminate — rather than reverse — the impact of scale range. An analysis of subjects' own behavioural reports, provided in an open-answer format *before* subjects were exposed to the report of the target person, reveals, however, that randomization was not successful under these conditions. While subjects who were assigned to the low frequency scale condition reported watching TV for an average of 1½ hours per day, subjects assigned to the high frequency scale condition reported an average of ½ hour, $p < .05$, Duncan test. Thus, the pattern of data suggests that subjects who reported their own TV consumption may indeed have used their own behaviour rather than the comparison information provided by the scale to evaluate the target's satisfaction. This, however, resulted in different judgements due to unexpected behavioural differences between both experimental conditions. In line with this interpretation of the unexpected result, subjects' reported own TV consumption is positively correlated with their evaluations of the target's leisure time satisfaction, $r(30) = .41, p < .02$, in these experimental conditions.

Discussion

In summary, the present findings demonstrate that the recipients of a behavioural report that is provided on a precoded scale use the range of the response alternatives as a frame of reference in making subsequent judgements, at least if their attention is not drawn to alternative standards of comparison, such as their own behaviour. This finding extends previous research by indicating that the use of response alternatives as a frame of reference is not limited to the respondent himself or herself, who may have paid particular attention to the response alternatives to determine his or her own behavioural frequency. However, the impact of scale range is apparently attenuated when other sources of comparison information are highly accessible, as was presumably the case when respondents were asked to report their own behaviour before they were exposed to information about the target person. Unfortunately, the data are not as conclusive as we would like on this point, due to the failure in random assignment described above.

EXPERIMENT 2

The results of Experiment 1 have potentially important applied implications. In many areas of clinical research and practice, self-report intsruments are commonly used to assess the frequency of patients' behaviours. An analysis of these scales indicates that they use either vague quantifiers, such as 'rarely', 'sometimes', 'frequently', and so on (e.g. Kassielke and Hänsgen, 1982; von Zerssen and Koeller, 1975, 1976) or numeric response alternatives (e.g. Fahrenberg, 1975; Kury, 1977), such as the ones explored in the present research programme. As a large body of research indicates, the use of vague quantifiers is highly problematic because respondents' understanding of terms such as 'rarely' or 'sometimes' shows considerable variation, and different respondents use different terms for the same absolute frequency (cf. Pepper, 1981 for a comprehensive review). Accordingly, the use of numeric response alternatives has been strongly recommended (cf. Pepper, 1981). Some scales follow this recommendation. For example, the best-known German symptoms checklist, the 'Freiburger Beschwerdeliste (FBL)' (Fahrenberg, 1975; Kury, 1977), asks respondents to report the frequency of 78 symptoms (such as headaches, or lack of energy) by checking numeric response alternatives, such as 'about twice a year', 'about twice a month', and so on. While numeric response alternatives avoid the problems associated with vague quantifiers, it is conceivable that they elicit response range effects of the type identified in Experiment 1.

To the extent that professional diagnosticians use the same strategies as lay-persons, the conclusions that they draw from a behavioural report on a symptoms checklist may not only depend on the absolute frequency of the reported behaviour but may also reflect the nature of the response scale on which this report was provided. Assume, for example, that a patient reports on a symptoms checklist that he or she suffers of lack of energy 'about twice a week'. According to the present research, we may assume that a health-care professional will consider this a more severe medical condition if reported on a scale that ranges from 'less than once a month' to 'more than twice a week', than if reported on a scale that ranges from 'less than twice a week' to 'daily'. Accordingly, the health-care professional may

also be more likely to recommend that the patient sees a doctor for a detailed examination in the former case than in the latter. Such a finding would clearly contradict normative models that hold that medical judgements should be based on a comparison of the absolute frequency of a symptom with a standard provided by medical knowledge and experience, rather than a standard suggested by the scale at hand (cf. Elstein, Shulman and Sprafka, 1978).

While this consideration may be quite discomforting, the findings of Experiment 1 also suggest that the impact of the frequency range of the symptoms checklist may perhaps not be very pronounced for experienced professionals. To the degree that experts can draw upon a wide range of other information that is well-organized and highly accessible (cf. Lesgold, 1988; Chi, Glaser and Farr, in press), they may use other applicable standards to evaluate the severity of the reported symptoms. If so, the hypothesized impact of the response scale may be limited to inexperienced novices, for whom the chronic accessibility of alternative standards of comparison is low.

To explore these considerations, we asked practising medical doctors and first-year students of medicine to evaluate the severity of several symptom reports that were presented to them in the context of high or low frequency scales, resulting in a 2 (level of expertise) × 2 (frequency range) factorial between subjects design.

Method

Subjects

Sixty-seven experienced medical doctors (32 female, 35 male), employed in hospitals at Lund, Kristianstad, Ängelholm, and Helsingborg (Sweden), and eighty first-year students of medicine at the University of Lund, Sweden (36 female, 40 male (four subjects did not indicate their sex)) participated in this study, and were randomly assigned to conditions. The doctors' mean age was 36.0 years and their average professional experience was 8.5 years. They represented different medical specializations, with 'general medicine' being the most frequent (31.3 per cent). The mean age of the first-year students was 22.8 years.

Procedure

Subjects were informed that the study investigated whether a standard health survey could be shortened without a decrease in usefulness and reliability. They received a questionnaire that presented nine frequency reports of different physical symptoms (six target items and three fillers), provided by nine different stimulus persons who had ostensibly participated in the health survey. Student subjects answered the self-administered questionnaire in a group setting during regular class hours, whereas the doctors answered it in their offices, where it was later picked up by the experimenter. Subjects had as much time as they wanted to complete the task.

Frequency range

For the six target items, the target person's response was presented in the context of either a high or a low frequency response scale, following a between-subjects design. That is, each subject was only exposed to reports given either on high or on low frequency response scales, thus providing a conservative test of the hypothesis.

Each symptom report was attributed to a different fictitious target person, described by initials, sex, and age. Because the presented symptoms have different objective frequencies, three different response scales were used, as shown in Figure 2.

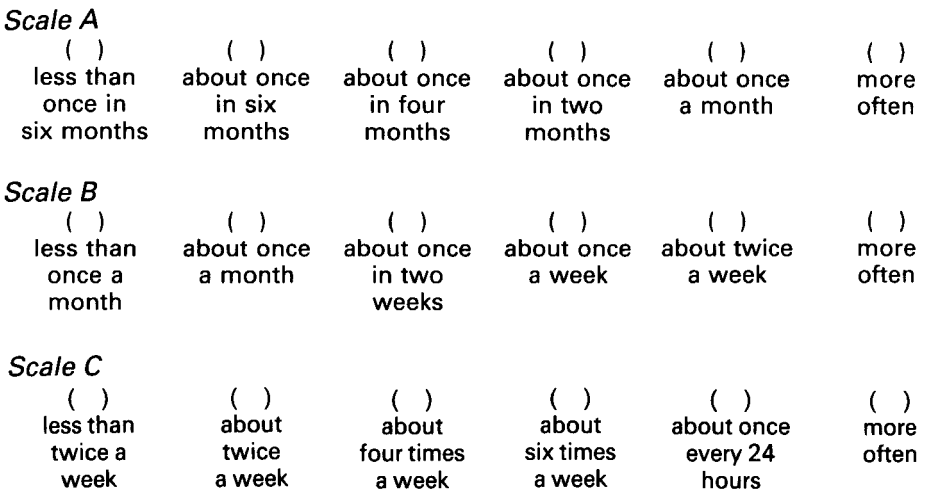

Figure 2. Response scales for medical symptom reports

For two target items ('stitches in the chest'; 'vomiting', attributed to Mr K., 43 years old; and Mr S., 39 years old, respectively), scale A constituted the 'low', and scale B the 'high frequency scale' condition. In both cases the response alternative 'about once a month' had ostensibly been chosen by the target person. For the remaining four target items ('aching loins or back', attributed to Mr Z., 25 years old; 'lack of energy', Mrs K., 41 years old; 'trouble in falling asleep', Mr S., 59 years old; 'lack of concentration', Mrs B., 35 years old), scales B and C represented the 'low' and 'high' conditions, respectively. In these cases the chosen response alternative was 'about twice a week'.

In addition, three filler items ('aching joints'; 'blood in stool'; 'lack of appetite') were presented, using the same scales but different frequency reports, to decrease overall response similarity that may have caused suspicion.

Dependent variables

For each item, subjects rated the *severity* of the symptom along 11-point scales (with the end-points labelled 0 = 'not at all severe', and 10 = 'very severe'), and the *necessity to consult a doctor* (with the end-points labelled 0 = 'not at all necessary to consult a doctor', and 10 = absolutely necessary to consult a doctor'), before they moved on to the next item.

After completion of all ratings they answered an open-ended question about the disease(s) and disorder(s) that may have caused the reported symptoms for each of the nine stimulus persons. These reports were evaluated by five expert judges, who were blind to conditions, as described below.

Results

Symptom evaluation

The mean ratings pertaining to the six target items are shown in Table 1. The six severity ratings provided by each subject, as well as the six consultation recommendations, were entered into two separate multivariate 2 (frequency range) × 2 (subject's level of expertise) × 2 (sex of subject) analyses of variance (MANOVAs), with the multivariate F-statistic based on Wilk's lambda. Because neither a main, nor an interaction, effect of sex emerged, all $F < 1$, the data presented in Table 1 were pooled over this variable.

As expected, these analyses revealed main effects of the frequency range of the response scale on subjects' ratings of the severity of the reported symptoms, multivariate $F(6,128) = 4.52$, $p < .0005$, as well as on their recommendations to see a doctor, multivariate $F(6,128) = 2.85$, $p < .02$.

Table 1. Mean severity and consultation necessity ratings as a function of scale range and expertise

Expertise:	Doctors		Students	
Frequency range of scale:	High	Low	High	Low
A. Rated severity of symptoms				
1 'Aching loins or back'	3.09	4.72	4.94	5.95
3 'Stitches in the chest'	4.39	4.50	5.88	6.17
5 'Vomiting'	4.94	5.38	3.75	4.90
7 'Lack of energy'	2.30	4.13	2.92	5.35
8 'Trouble in falling asleep'	1.56	2.59	2.53	3.07
9 'Lack of concentration'	1.73	3.34	2.22	2.98
B. Rated necessity to consult doctor				
1 'Aching loins or back'	4.48	6.25	6.00	7.07
3 'Stitches in the chest'	6.33	5.78	6.78	6.58
5 'Vomiting'	6.24	6.47	4.00	5.23
7 'Lack of energy'	3.42	4.62	3.06	5.15
8 'Trouble in falling asleep'	2.18	2.75	2.64	2.92
9 'Lack of concentration'	2.00	3.56	1.97	2.95

Note. Range of values is 0 to 10; higher values indicate higher severity and higher necessity to consult a doctor.

Specifically, all symptoms were evaluated as *more severe* when the same absolute frequency report was presented on a low rather than a high frequency response scale. Separate univariate analyses indicated that this pattern is reliable at $p < .05$ for all symptoms, except 'stitches in the chest'. Similarly, subjects were significantly more likely to recommend the *consultation* of a doctor when the symptom was presented on a low rather than a high frequency scale for three ('aching loins or back', 'lack of energy', 'lack of concentration') of the six reported symptoms. Thus, the same absolute frequency of experiencing a physical symptom was evaluated differently depending on the frame of reference provided by the response alternatives.

In addition, a main effect of subjects' expertise emerged on both measures; multivariate $F(6,128) = 6.20$ and 3.55, $p < .005$ and .01, for severity ratings and consultation recommendations, respectively. Specifically, the inexperienced first-year students rated five of the six symptoms as significantly more severe, and were more likely to recommend the consultation of a doctor in response to three of the six symptoms, than the experienced practitioners. This is likely to reflect a risk-avoidance strategy of the student subjects: if uncertain about a medical diagnosis the safe option is to assume that the symptom is severe and to recommend consultation.

Contrary to expectations, however, *no* interaction effect of level of expertise and frequency range of the scale was obtained for *any* of the items, all $F < 1$. Thus, the predicted impact of frequency range on subjects' severity ratings and consultation recommendations was independent of their level of expertise. Most importantly, it was obtained from experienced practitioners as well as from novices.

Perceived causes

After completion of all ratings, subjects had indicated possible underlying causes for the targets' symptoms. It was intended to further analyse the impact of the response scale by rank-ordering the perceived causes according to their severity. Five independent expert judges (medical doctors), blind to experimental conditions, who were asked to rank-order the causes along the severity dimension failed to do so, because the listed causes were too heterogeneous in themselves or represented disorders that may vary considerably in severity (e.g. 'depression', 'scoliosis', 'vertebral compression'). The only classification that seemed practicable was a distinction between *organic* causes on the one hand, and *psychological or psycho-social* causes on the other hand. The first cause that each subject had listed was categorized in this way to explore the impact of scale range and professional experience on subjects' most accessible hypotheses about the underlying causes.

The proportion of psychological/psycho-social causes was analysed for each item as a function of frequency range of the response alternatives and subjects' level of expertise, using a procedure described by Rosenthal and Rosnow (1985, pp. 47 ff.). The relevant percentages are shown in Table 2. Interestingly, the first causal hypothesis put forward by experienced practitioners for each symptom report was not affected by scale range for any of the symptoms, all $p > .15$, whereas the students' hypotheses differed as a function of scale range for three of the six symptoms. The students listed a significantly greater number of psychological causes for 'stitches in the chest', 'vomiting', and 'lack of concentration', when the symptom was reported on a high rather than a low frequency range scale, $z = 2.72$, 2.59, and 3.52, respectively, p's $<.01$. That is, the likelihood that a psychological cause was assumed increased as the perceived severity of the symptom decreased. This result may reflect a subjective theory held by the student subjects, that presumably light symptoms are more likely to be psychologically caused, whereas presumably severe symptoms are more likely to have an organic origin. Thus, an inference may be made from the perceived severity of a symptom to its underlying cause, resulting in an impact of scale range on the hypothesized causes that is mediated by its impact on perceived severity. The practitioners, on the other hand, may have learned from experience that severity is not a valid indicator of causation.

Table 2. Percentage of psychological or psycho-social causes and focused comparisons between scale conditions

Expertise:	Doctors		Students	
Frequency range of scale:	High	Low	High	Low
1 'Aching loins or back'				
%	13	17	51	65
z	−0.57		−1.25	
3 'Stitches in the chest'				
%	31	19	17	0
z	1.04		2.72*	
5 'Vomiting'				
%	29	28	47	19
z	0.15		2.59*	
7 'Lack of energy'				
%	31	27	53	39
z	0.35		1.18	
8 'Trouble in falling asleep'				
%	100	100	72	85
z	0.50		−1.24	
9 'Lack of concentration'				
%	66	54	73	34
z	0.93		3.52*	

Note: Percentages are given in the first row of each entry, and *z*-scores in the second; *z*-scores with an asterisk indicate a significant difference at the .005 level, one-tailed. All other $p > .10$.

Discussion

The results of Experiment 2 indicate that the use of response alternatives as a frame of reference is not restricted to lay-persons. Rather, professional users of a behavioural frequency report were also found to be influenced by the frequency range of the scale on which the report was provided in evaluating its implications. For example, experienced physicians as well as first-year students of medicine evaluated vomiting once a month as indicating a more severe medical condition, and were more likely to recommend consultation, when it was reported on a scale ranging from 'less than once in six months' to 'more often than once a month', than when it was reported on a scale ranging from 'less than once a month' to 'more than twice a week'.

Contrary to expectations, experienced physicians were found to rely on the implicit standards communicated by the response alternatives to the same degree as inexperienced novices. This finding suggests that it may not be sufficient to have relevant knowledge stored 'somewhere' in long-term memory. Rather, it may be necessary that this knowledge is highly accessible at the time of judgement to attenuate the impact of the resposne scale, as was suggested by Experiment 1. In fairness to our expert subjects, we have to add, however, that the only relevant information they had about each fictitious patient was a frequency report pertaining to one single symptom. It seems likely that the impact of this piece of information would be less pronounced if presented in the context of additional medical information, allowing the application of medical kowledge pertaining to symptom configurations (e.g. Lesgold *et al.*, in press). Moreover, it is conceivable

that the doctors would be less affected by the range of the response alternatives if they used the symptoms checklist routinely in their practice, thus acquiring considerable knowledge about the distribution of responses on the scale.

GENERAL DISCUSSION

The present findings, in combination with previous research (see Schwarz, 1988, in press; Schwarz and Hippler, 1987, for reviews) suggest that researchers and diagnosticians who use numeric response alternatives to obtain behavioural information from respondents should be aware of the potential impact of the information provided by the range of the response scale, at the level of data collection as well as interpretation.

At the level of data collection the frequency range of the response alternatives has been found to influence respondents' behavioural reports, in particular if the behaviour is frequent and mundane (Schwarz *et al.*, 1985; Schwarz and Bienias, in press). Because respondents are unlikely to have detailed episodic memories of mundane behaviours (see Bradburn, Rips and Shevell, 1987; Strube, 1987; Schwarz, in press, for reviews), they have to use estimation strategies to determine behavioural frequencies. In doing so they are likely to use the frequency range of the response alternatives as a salient frame of reference, resulting in higher behavioural reports on high rather than low frequency scales. This effect is the more pronounced, the less relevant episodic information is easily available in memory (Schwarz and Bienias, in press).

If the behaviour under study is ill-defined, as is frequently the case when subjective experiences are assessed, the frequency range of te scale is also likely to influence respondents' definition of the target behaviour (Schwarz, Strack, Müller and Chassein, 1988). For example, respondents who were asked how frequently they feel 'really irritated' assumed more severe cases of irritation to be the target of the question when presented a low rather than a high frequency response scale. Apparently, they used their knowledge about the relative frequency of mild and severe irritations, in combination with the response scale provided to them, to determine the meaning of the question.

At the level of data interpretation the users of a respondent's report should be aware of the potential impact of scale range on their *own* conclusions. As Experiment 2 indicated, even experienced experts seem to be highly susceptible to the impact of the response alternatives, and seem to use them as a frame of reference in making diagnostic judgements. While this reliance on the scale at hand may be adequate if the scale is carefully tailored to reflect the diagnostically relevant frequencies, the current findings suggest that a consideration of the scale's adequacy may not be part of the routine procedure used. Accordingly, the resulting decisions may, in part, be based on fortuitous standards that are highly accessible at the time of judgement, rather than on sound knowledge and experience.

ACKNOWLEDGEMENTS

Experiment 1 was conducted by Margit Kellenbenz, and Experiment 2 by Uwe Harlacher, as part of their diploma theses at the University of Heidelberg, under

the direction of the co-authors. The reported research was supported by grant Str 264/2 from the Deutsche Forschungsgemeinschaft to Fritz Strack and Norbert Schwarz. The comments of Fritz Strack on a previous draft are appreciated. Address correspondence to: Norbert Schwarz, ZUMA, P.O. Box 12 21 55, D-6800 Mannheim, W. Germany.

REFERENCES

Bradburn, N. M., Rips, L. J. and Shevell, S. K. (1987). Answering autobiographical questions: the impact of memory and inference on surveys. *Science*, **236,** 157–161.

Chi, M. T. H., Glaser, R. and Farr, M. (Eds) (in press). *The nature of expertise*. Hillsdale, NJ: Erlbaum.

Darschin, W. and Frank, B. (1982). Tendenzen im Zuschauerverhalten. Teleskopie-Ergebnise zur Fernsehnutzung im Jahre 1981. *Media Perspektiven*, **4,** 276–284.

Elstein, A. S., Shulman, L. S. and Sprafka, S. A. (1978). *Medical problem solving: an analysis of clinical reasoning*. Cambridge, MA: Harvard University Press.

Fahrenberg, J. (1975). Die Freiburger Beschwerdenliste FBL. *Zeitschrift für Klinische Psychologie*, **4,** 79–100.

Higgins, E. T., Strauman, T. and Klein, R. (1986). Standards and the process of self-evaluations: multiple affects from multiple stages. In R. M. Sorrentino and E. T. Higgins (Eds), *Handbook of motivation and cognition: foundations of social behavior*, pp. 23–63. New York: Guilford Press.

Kassielke, E. and Hänsgen, K.-P. (1982). *Beschwerdenerfassungsbogen*. Berlin: Psychodiagnostisches Zentrum.

Kury, H. (1977). Kreuzvalidierung der Freiburger Beschwerdenliste (FBL-W). *Zeitschrift für Klinische Psychologie*, **6,** 203–217.

Lesgold, A. (1988). Problem solving. In R. J. Sternberg and E. E. Smith (Eds), *The psychology of human thought*, pp. 188–213. Cambridge: Cambridge University Press.

Lesgold, A., Rubinson, H., Feltovich, P., Glaser, R., Klopfer, D. and Wang, Y. (in press). Expertise in complex skill: diagnosing x-ray pictures. In M. T. H. Chi, R. Glaser and M. Farr (Eds), *The nature of expertise*. Hillsdale, NJ: Erlbaum.

Ostrom, T. M. and Upshaw, H. S. (1968). Psychological perspective and attitude change. In A. Greenwald, T. Brock and T. Ostrom (Eds), *Psychological foundations of attitudes*, pp. 217–242, New York: Academic Press.

Pepper, S. C. (1981). Problems in the quantification of frequency expressions. In D.W. Fiske (Ed.), *Problems with language imprecision* (New Directions for Methodology of Social and Behavioral Science, Vol. 9), San Francisco: Jossey-Bass.

Rosenthal, R. and Rosnow, R. L. (1985). *Contrast analysis*. Cambridge: Cambridge University Press.

Schwarz, N. (1988). Was Befragte aus Antwortalternativen lernen. *Planung und Analyse*, **15,** 103–107.

Schwarz, N. (in press). Assessing frequency reports of mundane behaviors: contributions of cognitive psychology to questionnaire construction. In C. Hendrick and M. S. Clark (Eds), Review of personality and social psychology (Vol. 11). Beverly Hills, CA: Sage.

Schwarz, N. and Bienias, J. (in press). What mediates the impact of response alternatives on frequency reports of mundane behaviors? *Applied Cognitive Psychology*.

Schwarz, N. and Hippler, H. J. (1987). What response scales may tell your respondents. In H. J. Hippler, N. Schwarz and S. Sudman (Eds), *Social information processing and survey methodology*, pp. 163–178. New York: Springer Verlag.

Schwarz, N. and Scheuring, B. (1988). Judgements of relationship satisfaction: inter- and intraindividual comparisons as a function of questionnaire structure. *European Journal of Social Psychology*, **18,** 485–496.

Schwarz, N. and Strack, F. (in press). Evaluating one's life: a judgment model of subjective well-being. In F. Strack, M. Argyle and N. Schwarz (Eds), *Subjective well-being*, London: Pergamon.

Schwarz, N., Hippler, H. J., Deutsch, B. and Strack, F. (1985). Response categories: effects on behavioral reports and comparative judgments. *Public Opinion Quarterly*, **49,** 388–395.
Schwarz, N., Strack, F., Müller, G. and Deutsch, B. (1988). The range of response alternatives may determine the meaning of the question. *Social Cognition*, **6,** 107–117.
Strube, G. (1987). Answering survey questions: the role of memory. In H. J. Hippler, N. Schwarz and S. Sudman (Eds), *Social information processing and survey methodology*, pp. 86–101. New York; Springer Verlag.
Von Zerssen, D. and Koeller, D. M. (1975). *Die Beschwerdenliste.* Weinheim: Beltz.
Von Zerssen, D. and Koeller, D. M. (1976). *Die Befindlichkeitsskala.* Weinheim: Beltz.