



# STANDARDIZED GENERAL-PURPOSE NOISE REACTION QUESTIONS FOR COMMUNITY NOISE SURVEYS: RESEARCH AND A RECOMMENDATION

COMMUNITY RESPONSE TO NOISE TEAM OF IC BEN

(The International Commission on the Biological Effects of Noise)

J. M. FIELDS

*10407 Royal Road, Silver Spring, MD 20903, U.S.A. E-mail: [jfields@capaccess.org](mailto:jfields@capaccess.org)*

R. G. DE JONG

*TNO Prevention and Health, Section Environment, P.O. Box 2215, 2301 CE Leiden,  
The Netherlands*

T. GJESTLAND

*SINTEF Telecom and Informatics, N-7465 Trondheim, Norway*

I. H. FLINDELL

*Institute of Sound and Vibration Research, University of Southampton, Southampton S017 1BJ, England*

R. F. S. JOB

*Department of Psychology, University of Sydney, NSW 2006 Australia*

S. KURRA

*Istanbul Technical University, Faculty of Architecture, Division of Environment, Taksim 80191,  
Istanbul, Turkey*

P. LERCHER

*Institute of Hygiene and Social Medicine, University of Innsbruck, Austria*

M. VALLET

*INRETS-LTE, Case 24, 69675 Bron Cedex, France*

T. YANO

*Kumamoto University, Kurokami 2-39-1, Kumamoto 860, Japan*

## RESEARCH TEAM AT RUHR UNIVERSITY

R. GUSKI

*Department of Psychology, University of Ruhr, Bochum, Germany*

U. FELSCHER-SUHR

*University of Ruhr, Bochum, Germany*

AND

R. SCHUMER

*FernUniversität Hagen, Germany**(Received 30 August 2000)*

Differences in survey questions' wordings and weakness in some questions used to measure noise annoyance have interfered with accumulating knowledge about the factors that affect different communities' responses to noise. In 1993 an ICBEN team, Community Response to Noise, set the goal of creating high-quality survey questions that would yield internationally comparable measures of overall reactions to noise sources. After 7 years of discussions and research the team has developed and tested a method that attempts to meet those goals. The team recommends the use of a pair of multi-purpose questions in community noise surveys. The wording of the questions is presented for the nine languages for which a standardized empirical study protocol has been followed to select annoyance scale words. The team's protocol can be used to create comparable questions for additional languages in the future.

© 2001 Academic Press

## 1. INTRODUCTION

The development of a useful, widely acceptable, and scientifically strong body of knowledge about reactions to environmental noise is dependent upon the accumulation of knowledge from many studies. Knowledge about community residents' reactions to noise has come primarily from over 300 combined socio-acoustic surveys [1] in which residents' reactions to noise are analyzed in relationship to those residents' objectively determined, acoustical noise environment. It has long been recognized that a major barrier to accumulating a useful body of knowledge from these surveys has been the difficulty in comparing the surveys' results [2-4]. This is due in part to the use of different survey questions.

Comparable questions are of special importance in these socio-acoustical surveys for two reasons: (1) A major survey product is a tabulation of answers to a noise reaction question (percent of residents with high annoyance) that is used to judge whether the reactions at the same exposure in different locations, cultures and countries are sufficiently similar to support uniform national and international noise regulations. (2) Relatively standard procedures are already used to measure the surveys' fundamental acoustical variables such as the total noise exposure, the peak noise levels, the numbers of noise events and the timing of the noise events. Previous studies using diverse noise-reaction questions seem to indicate that reactions in different communities to the same noise exposure may vary as much as would be predicted by a 30-decibel difference within a community [5, p. 238]. Given the diverse social survey questions used, however, this lack of agreement could be due to diverse social survey questions. After more than 35 years and 350 surveys, different studies continue to use their own diverse, non-comparable reaction questions. While it is not clear that any particular noise-reaction question is best, it is clear that the lack of a shared question hinders the accumulation of comparable information. Over the last 35 years the widespread recognition of this problem has generated several analyses of the problem and recommendations for standard noise reaction questions [6-8].

In 1993 the Community Response to Noise Team (Team 6) of the International Commission on the Biological Effects of Noise (ICBEN) developed a program to facilitate comparisons between socio-acoustic surveys. The result of the first part of this program, the development of reporting guidelines for socio-acoustic surveys, was published in 1997 [9]. The team's second major project, the development of scientifically sound, comparable noise-reaction questions, is the subject of this article.

In the years after the 1993 ICBEN meeting, a group of researchers including the ICBEN Team 6 members, a University of Ruhr team and other community noise acousticians reviewed previous studies, conducted workshops and performed research with the goal of developing a noise reaction measure for social surveys that has the following characteristics:

1. permits valid international comparisons of survey results within and between languages;
2. provides a high-quality, reliable measure of a general reaction to a noise experienced in a residential environment;
3. yields transparent results that will be consistently interpreted by survey respondents, policy makers, and report readers;
4. yields an interval-level measurement scale (i.e., the response scale answers are equally spaced) meeting the assumptions for regression and many other analysis techniques;
5. is likely to be widely adopted internationally;
6. is suitable for all questionnaire administration modes (face-to-face, telephone or self-administered).

Designing questions to meet these criteria requires choices on such issues as the type of question (open-ended or closed), the description of the reaction (e.g., annoyance, bother, disturb), the word referring to the acoustical environment (e.g., noise, sound), the specification of the residential conditions (e.g., overall or night/day, inside/outside, open-/closed windows, etc.), the type of answer scale (verbally labelled or only numbered), the exact words on a verbal answer scale, the number of points on that scale (e.g., 4-points, 11-points, etc.), and the general wording of the question. This article reports the steps that were taken to address these issues. The first section reviews relevant research. The next section reports on additional research that the team conducted to select annoyance scale modifiers. Finally, the two recommended noise reaction questions and the rationale supporting them are described. The 5-point verbal scale question is:

“Thinking about the last (. . .12 months or so. . .), when you are here at home, how much does noise from (. . .noise source . . .) bother, disturb, or annoy you; Extremely, Very, Moderately, Slightly or Not at all?”

The (0–10) point numeric scale question is (see section 4.1 for the full wording):

“...what number from zero to ten best shows how much you are bothered, disturbed, or annoyed by (. . .source. . .) noise?”

The rationale for specifying the exact wording for two questions is not to eliminate all other reaction measures. Although these questions are valuable for comparisons and for measures of general reactions, many noise reaction issues require additional noise reaction questions.

## 2. REVIEW OF RESEARCH ON QUESTION WORDING DIFFERENCES

Bibliographic searches reveal a large body of social science literature on the effects of questionnaire design and question wording on answers to social survey questions [10–12]. Although this general literature is valuable, it does not resolve many question wording choices for a noise survey or any specific survey because: (1) contradictory effects can be hypothesized for any particular situation so that it is not clear which effect, if any, may be dominant; (2) the underlying principles that explain some published findings are not sufficiently specific to be applied to a specific situation; (3) the published literature is likely to be biased toward reporting positive effects (i.e., instances in which questionnaire design did not affect results are less likely to be published).

TABLE 1

*Effect of scale points on reporting amount of television viewing*

QUESTION: How much television do you watch a day? ( <i>N</i> = 132)					
Low watching choices scale (%)			High watching choices scale (%)		
Under 2½ hours	Up to ½ h	7%	84%	Up to 2 ½ h	63% 63%
	½-1 h	18%			
	1-1 ½ h	26%			
	1 ½-2 h	15%			
	2-2 ½ h	18%			
Over 2½ hours	More than 2 ½ h	16%	16%	2 ½-3 h	23%
				3-3 ½ h	8%
				3 ½-4 h	5%
				4-4 ½ h	2%
				More than 4 1/4 h	0

The remainder of this section reviews and reports upon findings on five choices that must be made in the design of a noise reaction question. A sixth issue, the choice between a 4- and 5-point scale, is discussed in section 5.2.

## 2.1. RESEARCH ON ANSWER-SCALE LABELS

Research on the effect of variations in question wording has directed the current research effort toward an examination of alternative answer scales. Previous community noise surveys have used a range of answer scales but do not provide a firm basis for comparing alternative labels because the surveys that have differed from one another in their answer scales have also differed from one another in enough other ways that the effects of answer scale differences could not be isolated. As a result our knowledge about the importance of these issues comes from consistent findings from non-noise studies.

One of the most dramatic and most often replicated findings on the effect of answer-scale labelling is a finding from a German experiment on answers to the question "How much television do you watch a day?" [13, p. 391]. In Table 1, it is seen that on the low watching category scale where five of the six scale points are for under 2.5 h of daily television viewing, 84 per cent of the respondents reported low watching (i.e., under 2.5 h). In contrast, on the high watching category scale, where 2.5 h is only one of six points, only 63 per cent reported low watching (chi square = 7.7,  $p < 0.005$ ). This research shows that alternative presentation formats for logically identical questions can affect the answers about even a factual matter.

Although no equally definitive study has been conducted for noise annoyance scales, two studies suggest that similar effects may occur. In a small laboratory study of ratings of 13 nuisances, Rohrman found that when "very" was point #4 on a 5-point scale, an average of 31 per cent reported being at least "very" annoyed (i.e., point #4, "very" or point #5, "extremely"), but when "very" was point #5 on the scale only 14 per cent chose "very" annoyed and thus would be estimated to be at least very annoyed (analysis of data provided by Rohrman, 1998). Another laboratory noise study found that subjects gave slightly lower annoyance responses at low noise levels to the same 5-point annoyance scale when that 5-point question was preceded by a binary annoyance question about any annoyance (Is the sound annoying?) [14]. Neither of these studies provides a firm basis for question

design decisions because the first study included only 30 subjects and the second study's design is likely to have underestimated the effect of a binary question since all subjects knew that they would also answer the following 5-point scale. These noise studies do, however, reinforce the findings from other studies about the importance of using strictly comparable annoyance scales.

Researchers faced with comparing the results from different surveys with different annoyance scales have devised more or less elaborate methods for attempting to calibrate the scales based, for the most part, on intuition and logical considerations [3, 15: p. 3434]. The research reviewed in this section suggests that such calibrations are uncertain and that a firmer basis for between-survey comparisons is identically worded and scaled survey questions.

## 2.2. RESEARCH COMPARING NEGATIVE AND BIPOLAR SCALES

Most noise annoyance questions use a unipolar scale that extends from a negative pole to a neutral point. One study that asked about negative reactions also includes an unusual bipolar scale that extends from a negative to a positive pole. This study about both aircraft and road-traffic noise near Toronto airport [16, Appendix A] first asked a neutral screening question, "Do you ever notice ... sounds.." from "aircraft" and "main road traffic noise ...". If respondents "noticed" a sound they were asked to "... rate each of the sounds ..." on a bipolar 9-point verbal scale that included the four positive points of "Extremely, Considerably, Moderately, Slightly Agreeable", one "Neutral" point, and four negative points of "Extremely, Considerably, Moderately, Slightly Disturbing." Later respondents used a more conventional unipolar 0–10 numeric scale of "How do you rate..(the same).. noise ..." where the end points were labelled "0 Not at all disturbed" and "10 Unbearably disturbed". Precise conclusions about differences between negative and positive questions are limited with this survey for several reasons: (1) respondents not "noticing" a sound were not asked the bipolar question, (2) the scales had different numbers of points, (3) the bipolar road traffic question included a negative descriptor ("road traffic *noise*"), and (4) the unipolar scale had an extreme label for the top point ("Unbearably disturbed").

Our reanalysis of these data found that strictly positive reactions ("agreeable") were given by small proportions of respondents; about 4 per cent for aircraft (at all noise levels) and from 1 to 10 per cent for road traffic (depending on noise level). The variations by noise level for road traffic may be partly due to the "notice" filter question that could have foreclosed some positive reactions at low noise levels and partly due to sampling error (the estimate of 10 per cent is surrounded by a 95 per cent confidence interval of  $\pm 6$  per cent). In Table 2, the aircraft noise responses for unipolar and bipolar questions are compared in alternating columns within three noise classes. In the table the sum of the positive and neutral categories on the bipolar scale are seen to be almost the same as the sum of the two lowest points on the 11-point numeric scale. For road traffic the percentages were the same (69 per cent) for the sample as a whole. The estimates of the percentages with high annoyance ratings ("extremely" for the bipolar scale or point #9 or #10 for the unipolar) are seen in Table 2 to be similar or greater for the bipolar scale than for the unipolar scale. A similar pattern was observed for the road traffic data. Although the differences in the Toronto survey questions' wording and numbers of scale points interfere with a precise comparison of bipolar and unipolar scales, they do suggest that a bipolar scale and unipolar scale will give similar estimates of the proportion of the population that is annoyed by noise. The extent of similarity would be judged to be less from these data if the end points of the numeric scale had not been collapsed before making the comparisons presented above. The results for any type of analysis, however, do suggest that there are some positive reactions, but that surveys that

TABLE 2

*Comparison of reactions to aircraft noise measured by one bipolar scale and one unipolar negative scale in three noise ranges (Toronto aircraft/road survey)*

Definition of scale points		% choosing the annoyance response by scale type at three noise levels					
		<60 dB LAeq(24)		60–69 dB LAeq(24)		≥ 70 dB LAeq(24)	
Bipolar	Unipolar	Bipolar	Unipolar	Bipolar	Unipolar	Bipolar	Unipolar
Not hear or Extremely to slightly agreeable or Neutral	0 <i>Not hear or Not at all disturbed</i>		30%		12%		8%
	1	34%	3%	20%	2%	9%	6%
Slightly disturbing	2		9%		7%		2%
	3	23%	7%	18%	5%	25%	5%
	4		9%		8%		12%
Moderately disturbing	5	15%	12%	17%	15%	16%	13%
	6		7%		13%		9%
Considerably disturbing	7		11%		14%		9%
	8	13%	6%	15%	10%	24%	13%
Extremely disturbing	9		4%		5%		11%
	10 <i>Unbearably disturbing</i>	16%	4%	31%	10%	27%	13%
Total		101% (323)	100% (323)	101% (251)	101% (249)	101% (93)	101% (93)

require information about positive reactions could measure them in separate questions asked of all respondents regardless of whether or not they may “notice” the noise.

### 2.3. RESEARCH ON THE PRESENCE OF A MIDDLE ALTERNATIVE

Research on the effects of middle, neutral alternatives on answers to bipolar opinion questions [10, p. 162] suggests to some investigators that a 5-point scale may distort annoyance responses towards the choice of point # 3 (the middle alternative) on a 5-point noise annoyance scale. Such a pattern would suggest that an even number of scale points might yield a better scale. This possibility has been examined by analyzing the 5-point annoyance questions from six community noise surveys that contain 53 different noise situations in which over 12 000 respondents used a total of seven questions to give 73 noise environment ratings (one survey with 20 environments included two separate 5-point scales). The analysis measured the extent to which any point on these 5-point answer scales exhibited a heaping of responses. “Heaping” is arbitrarily defined as a five percentage point higher choice for one answer than for both of the adjacent answers or, in the case of an endpoint, than for the single adjacent answer.

The first assessment method (simple heaping) determined how many of the 73 ratings of noise environments had a peaking or “heaping” of responses at each scale point. The heaping, defined by the 5 per cent rule, is seen in Table 3 to not be present in 16 per cent of the ratings and to be primarily concentrated at either endpoint for more than half of the rating environments. Nonetheless, there is more heaping at point #3 (the middle) point than at point #2 or point #4 on the scale. Because such simple heaping at a scale point could be due to the noise level itself, a second assessment method is more conclusive.

The second assessment method (multi-modal heaping) tabulates the number of environments that exhibit departures from an expected response pattern. Given the most popular response category for a particular noise environment, the expected pattern is for the proportion of respondents selecting each response category to monotonically decrease away from that most popular response category. Multimodal heaping occurs at the midpoint when the midpoint of the scale interrupts the normal monotonic decrease from another popular scale point. For example, such bimodal heaping occurred in the 40–44 LAeq environments for study USA-220 when 68 per cent of the respondents chose point #1, 8 per cent chose point #2, 13 per cent point #3, 5 per cent point #4, and 5 per cent point #5. When the 73 rating environments were examined from all studies, there was no multimodal heaping 85 per cent of the time. For the remaining 15 per cent (11 environments) such multi-modal heaping involved point #1 in 9 environments, point #5 in 7, point #2 in 1, point #4 in 1, and point #3, the middle point, in 5. In short, multi-modal heaping involved the middle point only 7 per cent of the time and did not occur at all 85 per cent of the time. If some heaping does occur, it appears to occur so infrequently as to not preclude the choice of a 5-point scale. The absence of heaping may be due to fact that the middle alternative on this unipolar scale does not represent the convenient opportunity to express an “undecided” response as does a bipolar scale for which the phenomena has been reported for non-noise topics.

#### 2.4. RESEARCH ON THE WORDING OF NOISE QUESTION STEMS

Social science research has shown that seemingly innocuous differences in the wording of the stems of survey questions can have dramatic effects on respondents’ answers [17]. One classic question wording experiment has been repeated 4 times over 36 years with a similar question wording effect each time. Each time part of the sample was asked “Do you think the United States should forbid public speeches against democracy?” About 25 per cent more respondents opposed such speeches when they were asked an alternative question that had exactly the same policy implications: “Do you think the United States should allow public speeches against democracy?” [10, p. 277]. This is one of many examples in which the wording of survey questions clearly affects respondents’ answers.

Results from noise annoyance questions have shown large effects for some variations in question wording but not for other, seemingly important, differences in question wording. The only noise annoyance survey in which alternate forms of questionnaires were used with closely matched samples of respondents (respondents from adjacent households) found that many variations in question wording or location had no effect on answers [18, p. 250; 19]. The amount of speech interference reported at a specific location in the room where the interview was conducted was no less for the half of the respondents with whom the habitual present tense was used (“... does the noise from the trains make you stop talking or pause or speak louder?”) than for those with whom the word “ever” was used and even marked for emphasis by the interviewer (“... has the noise from trains *ever* made you stop talking or pause or speak louder?”) [19].

TABLE 3  
*Scale points with a simple heaping of responses (six studies providing seven data sets)*

Study (ID Number in catalog [1])	Scale type	Number of interviews	Number of groups with heaping of answers (e.g., 5% more than adjacent categories)						Total number of groups
			Scale point at which heaping occurs						
			No heaping	Point #1 (lowest)	Point #2	Point #3 (middle)	Point #4	Point #5 (highest)	
1967 4-Airport: USA-022	Numeric	3499	0	9	0	1	0	4	10
1969 3-Airport: USA-032	Numeric	2899	0	3	0	1	0	7	8
1970 2-Airport: USA-044	Numeric	1945	0	7	0	2	0	2	9
LAX-Night: USA-082	Verbal	1471	1	0	0	1	0	0	2
Noise change: USA-203	Verbal	953	1	0	0	1	0	2	4
Sonic boom: USA-375	Verbal	1546	6	12	1	0	2	2	20
Sonic boom: USA-375	Numeric	1568	4	9	4	2	0	1	20
Total <sup>†</sup>			12 (16%)	40 (55%)	5 (7%)	8 (11%)	2 (3%)	18 (25%)	73

<sup>†</sup>Note: Percentages and numbers can exceed the totals because some groups had heaping at more than one scale point.



The conclusion drawn from the research reviewed in this section is that it is not possible to predict when question wording may have large effects or no effects at all. The safest course for comparisons is to base comparisons on identical questions that are presented in similar contexts.

## 2.5. RESEARCH ON THE LOCATION OF NOISE REACTION QUESTIONS WITHIN A QUESTIONNAIRE

In community noise reaction surveys of a single noise source, the primary response question is usually placed before more extensive sets of questions about the same source to avoid the possibility of biasing respondents' answers by heightening their awareness of the effects of that noise source. Three field experiments have been conducted, however, where the location of the primary response question was varied in the questionnaires but did not affect answers.

In two of the experiments the noise response question was asked twice in each questionnaire, first in the traditional location early in the questionnaire and second, after 10 to 20 min of detailed questions about the noise source, near the end of the questionnaire. Neither an experiment in a road traffic survey [20] nor an experiment in a railway survey [19] found any evidence that answers at the end of the questionnaire expressed any higher noise annoyance. In the road traffic survey, for example, the percentage of the sample expressing each of seven degrees of dissatisfaction from lowest (1 = definitely satisfactory) to highest (7 = definitely unsatisfactory) was 8, 6, 8, 16, 17, 13 and 32 per cent at the beginning of the questionnaire and 8, 6, 9, 15, 14, 15 and 33 per cent at the end of the questionnaire for the 2881 respondents who answered both questions. A similar lack of an effect was found for the railway survey even for the half of the sample that was informed just before the ending question that one of the sponsors was "... British Rail, the people who run the railways ..." [19, p. 62]. For both the road traffic and railway survey the mean of the respondents' differences in answers between the first and second question (each question was scored from 1 to 7) was not significantly different from zero ( $p > 0.05$ ).

In a third survey in which approximately 743 ratings were given by 275 respondents over four rounds of interviews about home energy usage, the same 7-point numeric noise dissatisfaction scale preceded a 4-point verbal scale by two questions in one version of the questionnaire and followed that 4-point verbal scale in a second version of the questionnaire [21]. When these data were analyzed for this report no difference was found for the 7-point scale when the question appeared in different locations. For the 4-point verbal scale, somewhat greater annoyance was expressed when the 4-point scale came first (about 10 per cent more were "very" annoyed and about 7 per cent fewer were "not at all annoyed"). This result does not, therefore, support the hypothesis that annoyance scores are biased by placing a noise annoyance question in the less conventional location at the end of a questionnaire. In each of the three experiments, the lack of a location effect remained after controls for noise level were introduced into the analyses.

The existing research suggests that departures from the standard beginning-of-questionnaire location do not bias answers to noise-response questions. However, a remaining concern about question placement arises from the finding that answers to noise annoyance questions can be different for different times of day [22, p. 64] and for outside locations [23, p. 185]. Although the first two question experiments reviewed in this section occurred in questionnaires that asked about many locations and times of day, none of the three experiments placed the general noise annoyance question immediately after questions that focused on only a single time of day or only the out-of-doors environment. In the

absence of such tests it seems best to avoid placing a question immediately after a series of questions that specify a single time of day or location in the home.

### 3. METHOD FOR DEVELOPING HIGH-QUALITY, MULTI-PURPOSE, COMPARABLE NOISE REACTION QUESTIONS

The IC BEN team's initial goal to develop comparable questions was broadened to include the additional five question-quality goals described in section 1.0. The present section describes the principles, steps and methods that were followed in developing the recommended noise reaction questions.

#### 3.1. STEP #1: REVIEW PREVIOUS RESEARCH AND TEAM MEMBERS' EXPERIENCES TO SET THE BASIC FORM OF THE QUESTION

Socio-acoustic community surveys have been conducted on a regular basis since the late 1950s. After the first few studies, a broad consensus developed to use direct, close-ended questions as the primary measures of overall reactions to noise in residential areas. When the committee members reviewed questions from previous surveys and considered their own experiences they accepted that consensus. The reasons for accepting the various parts of this consensus are described later in section 5 ("Basis For Choosing These Questions"). The committee also noted that there were many differences in the detailed wording of the questions that might affect respondents' answers and thus concluded that for the purposes of comparisons between surveys it was important to agree upon a standard wording. The approach to developing a standard wording was different for the question stem (the question itself) and the answer scale (the answers that respondents choose to report their reactions).

#### 3.2. STEP #2: RELY ON REVIEWS AND EXPERT JUDGMENT TO REFINE THE WORDING OF THE QUESTION STEM IN ENGLISH

The draft of the wording for the question stem was circulated to all committee members, revised, and then subjected to thorough examination in workshops by general audiences at acoustical conferences and by small groups of community noise study experts. The early draft versions of the questions were also published in conference proceedings [24, 25]. The questions used wordings and concepts that were among those that had been found to be acceptable in previous surveys. Although it was the English version of the question that was subjected to this review, most of the participants in the review process were not native English speakers and thus considered the appropriateness of the question for their own languages.

#### 3.3. STEP #3: TRANSLATE AND BACK-TRANSLATE TO DEVELOP AND ADAPT THE QUESTION STEM FOR LANGUAGES OTHER THAN ENGLISH

After the question stem was developed in English, the participants translated the question into their own languages. This translation process was expected to be routine except for the choice of a replacement for the phrase "bother, annoy, or disturb" that would convey the idea of a general negative reaction as it does in English. The translators did not attempt to

exactly translate these three terms but rather to select from one to three terms that would convey the same general negative reaction in their own languages. As a check on the adequacy of the translation, the question stem was translated back into English by at least one other native speaker who had not previously worked on the project. Any differences were resolved by the native speakers.

#### 3.4. STEP #4: CONDUCT PARALLEL STUDIES FOLLOWING A UNIFORM PROTOCOL TO SELECT VERBAL MODIFIERS FOR THE ANSWER SCALES

When the committee met to review proposed wordings, there was general agreement on the wording of the question stem but not on the selection of verbal modifiers for the answer scales or on the number of points for a verbal answer scale. The committee decided that neither dictionary translations nor expert judgment provided a sufficient basis to select verbal modifiers with good metric qualities for a single language or consistent meanings across different languages. It was decided that empirical studies were needed in each language to choose verbal modifiers that would have the same positions on an underlying scale of intensity of reaction to noise. These empirical studies from the first nine languages also provided information that helped in deciding between a 4- and 5-point verbal answer scale (see section 5.4).

### 4. THE RECOMMENDATION

ICBEN's Community Response to Noise Team (Team 6) recommends that each survey use two questions to measure annoyance reactions for the purpose of making comparisons between social surveys. This section presents the English version of the two questions (including their accompanying answer cards) and the instructions for administering the questions.

#### 4.1. THE RECOMMENDED NOISE-REACTION QUESTIONS

The recommended measurement procedure consists of one verbal answer scale question (*Q.V.*) and one numeric answer scale question (*Q.N.*). In English, the questions are the following (other languages are in Appendix A):

[ASK ALL RESPONDENTS]

Q.V “Thinking about the last (..12 months or so..), when you are here at home, how much does noise from (..noise source..) bother, disturb, or annoy you; Extremely, Very, Moderately, Slightly or Not at all?”

Q.N “Next is a zero to ten opinion scale for how much (..source..) noise bothers, disturbs or annoys you when you are here at home. If you are not at all annoyed choose zero, if you are extremely annoyed choose ten, if you are somewhere in between choose a number between zero and ten. Thinking about the last (..12 months or so..), what number from zero to ten best shows how much you are bothered, disturbed, or annoyed by (..source..) noise?”

The words appearing in parentheses are to be replaced by phrases that are most appropriate for the noise source and time period being studied. When these questions are used in an interviewer-administered questionnaire, respondents choose their answers from the answer cards in Figure 1.

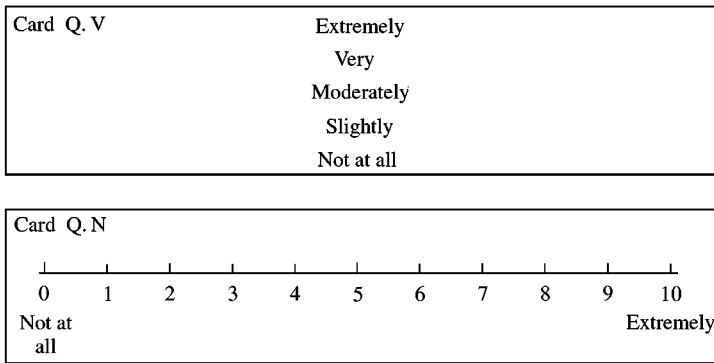


Figure 1. Answer cards for recommended annoyance questions.

On the basis of the findings in section 6 it is recommended that the top two scale points on the 5-point scale (“very” and “extremely”) be combined to measure the proportion of respondents who are “highly” annoyed.

Although these two questions are methodologically sound and preferred for comparisons between surveys, some surveys may require additional general noise reaction questions. The purpose of each particular survey must dictate whether these two questions are to be the primary reaction questions for the survey or would be supplementary questions included for the purposes of making comparisons between surveys.

#### 4.2. INSTRUCTIONS FOR ADMINISTERING THE QUESTIONS WITHIN A SURVEY

To obtain comparable annoyance measures the questions must be presented to respondents in an identical way. The following guidelines are suggested to ensure the uniform administration of these questions (Justifications for the guidelines are given in section 5):

1. *Ask all respondents both questions.* Two questions are needed to obtain a reliable measure of reactions. Each and every respondent is first asked both of these questions. If it is necessary to determine whether some respondents do not hear the noise source, a question about the audibility of the noise can be asked separately.

2. *Present the full scale, exactly as worded, to all respondents.* Interviewers should not rephrase or “explain” questions. The questions should not be divided into parts by, for example, using an initial, screening, binary question about whether the respondent is “annoyed”.

3. *Place the questions early in the questionnaire, unless this conflicts with other survey objectives.* Most noise/annoyance surveys put the first noise/annoyance question early in the questionnaire before it is clear that the focus of the survey is on noise generally or on noise from a particular source. If these proposed questions can be the primary noise/annoyance questions for a survey, then it is recommended that they be placed early in the questionnaire. If other noise annoyance questions are more important for the survey’s purposes, then the recommended questions can be placed later.

4. *If pretests indicate that the questions are perceived as repetitious, include appropriate instructions.* If the questions are not placed early in the questionnaire, potential interviewer or respondent discomfort with apparently repetitious questions can be solved with

introductions to the questions similar to the following:

*(Introduction A)* “Now we return to the noise from (..source..) and take everything into account we have discussed. Thinking about the last ... [insert recommended questions] *(Introduction B)* “People in other countries have answered this next question to tell us how they feel about noise. Now you can use it for the noise here. Thinking about the last ... [insert recommended questions]” *(Introduction C)* “Even though all of the questions are slightly different, I know a few of them can seem similar for people in your situation. If any seem repetitious for you, just give a quick answer and we will move to other questions.”

5. *Prepare written instructions for interviewers.* For telephone or personal interviews the interviewers should be provided with written instructions that: (1) instruct interviewers to ask questions exactly as written, (2) train interviewers to respond to “I don’t understand” with methods that do not require paraphrasing the question, (3) urge respondents to choose between the offered answers, (4) encourage all residents to answer these questions (new residents can be instructed to answer about only their recent period of residence), and, if repetition is expected to be a problem, (5) provide interviewers with instructions for respondents who find the questions to be repetitious.

## 5. BASIS FOR CHOOSING THESE QUESTIONS

Considerable thought and research went into the selection of the two questions recommended in the previous section. This section gives the rationale for the general type of question approach, the types of answer scales, the exact wording for these question stems, and the selection of the response modifier words for the answer scales.

### 5.1. THE TYPE OF QUESTION

The recommended questions are short, direct (closed) rating questions that ask about only neutral or negative reactions and are intended for all respondents. Before recommending these questions careful consideration was given to the sometimes-voiced assumption that such direct questions could give biased estimates of the numbers of people who are “truly” annoyed by noise.

#### 5.1.1. *Direct rating question with closed scale*

These direct rating questions name the noise source, ask for respondents’ feelings about the noise source, and present respondents with choices between a limited number of answers. After some initial research with indirect questions [26] such direct rating questions have been almost universally accepted as the primary measure of the relationship between noise and residents’ reactions. Answers to such direct questions are more explicit and readily interpreted than indirect questions or comparison questions, the two other types of questions that are sometimes used for special purposes in noise surveys.

Indirect questions attempt to ascertain the underlying impact of noise with either open questions in which the noise source is not identified, complaint questions in which respondents report complaint actions rather than feelings, or behavioral questions in which respondents report behavioral adaptations rather than feelings. Although useful for specific purposes, none has supplanted the direct questions as the primary indicator of noise impact because these questions can only be used to indirectly infer how people feel about noise impact. In addition, such indirect questions are less highly related to noise exposure [27, p. 187] Indirect, open questions that allow respondents to volunteer their own answers are expensive to analyze and result in answers that cannot be directly compared.

The other type of question, a comparison question, provides an anchor for a rating by asking respondents to compare their feeling about the specified noise to their feeling about some other object. The overwhelming problem with comparison questions is the absence of a common, shared anchor that could provide a uniform point of comparison across surveys or even across neighborhoods in the same survey. The most obvious anchors, other neighborhood nuisances, vary so greatly from site to site that they cannot be used for comparing noise responses at different sites. Magnitude estimation techniques could, in theory, use other shared reference points to resolve this problem, but previous research has found that such techniques are not sufficiently refined for a question to be recommended for wide usage in noise-reaction surveys (Fields, 1996).

### 5.1.2. *Unipolar scale (Negative neutral)*

The recommended questions ask about the negative problem, “noise”, and use unipolar scales that extend from a negative pole (“extremely annoyed”) to a neutral position (“not at all annoyed”) but not to a positive pole (“extremely enjoyable”). The practice of asking about such negative reactions is almost universal in community noise surveys. The decision to ask about the negative problem was made with the knowledge from previous surveys and the analysis in section 2.3 that reactions to transportation noise are overwhelmingly either negative or neutral and that scales with both positive and negative scale points yield almost the same response distributions as those with only negative scale points. In addition a question about the negative concept of “noise” is expected to be more quickly and easily understood than a question that asked for a positive rating of “sounds”. Of course, positive reactions can still be studied by asking additional questions about positive reactions.

### 5.1.3. *One-part question for every respondent*

The same two questions are asked of every respondent regardless of length of residence, audibility of noise source, or degree of annoyance. This uniform administration procedure simplifies the questionnaire design, increases the comparability of measures in different surveys, increases the accuracy of the data and still provides the flexibility to analyze the responses of only subgroups of respondents. New residents can, for example, be asked the question, but be removed from some analyses. The alternative procedure of designing an additional internationally accepted uniform screening question that would eliminate new residents would almost certainly create additional differences between surveys as the different investigators decided how to define such terms as “new” (one month or one year) or a “change in residence” (e.g., Would a move within a neighborhood be a change?).

The questions are not screened or broken into separate parts primarily because such screening procedures distort the measurement of reactions by interfering with obtaining clear, unambiguous measures of annoyance. The difficulty is that research suggests that both audibility and “any annoyance” screening questions contain measures of degree of annoyance and are affected by the screening question’s use of only two scale points (binary response of “yes” or “no”). An English road-traffic survey, for example, shows that a screening question about “hearing” or “not hearing” a noise source is not interpreted according to a dictionary definition of audibility. In that probability sample of the population of England, automobiles could be heard at all homes (all the respondents lived on a road), but 11 per cent reported that they could not hear “cars, lorries or other road traffic” [28, p. 35]. In accord with the types of “conversational norms” that are found in interviews [29, p. 43], it is likely that when these respondents were asked a patently unreasonable question, they reinterpreted the question to not be about audibility but rather to be about low levels of annoyance (e.g., “Do not REALLY hear the noise”). A similar

reinterpretation could be expected in answers to any binary screening question about being “annoyed” or “not annoyed”. Research on answers scales cited in section 2.1 leads to the expectation that some respondents who would answer “slight” annoyance on a 5-point scale would give a response of “no annoyance” on a binary, 2-point scale. A screening question introduces additional ambiguity in interpreting the responses to a 5-point annoyance scale since the four points on that scale presented after the answer “yes” to a binary “any annoyance” question might be perceived as forming a 4-point scale. Since annoyance is a finely graded response rather than a simple, unambiguous, binary condition, it is to be expected that the number of scale points will affect answers. A more accurate measure of the audibility of noise sources can be obtained from a question that follows the primary annoyance question and, as a result, makes it clear that not annoyed respondents can still report hearing a noise.

#### 5.1.4. *General, non-specific reaction question*

The recommended questions seek to obtain general, persistent reactions that allow respondents to integrate their experiences over different times and locations in their home. The questions are designed to obtain overall assessments from respondents who differ from one another in their sensitivities to noise in different locations and time-of-day conditions. These questions do not specify one particular combination of conditions because an overall response measure necessarily involves an integrated response over a range of different types of experiences and sensitivities. The questions do not explicitly list the range of conditions over which the experiences should be integrated for five reasons: (1) a complete list would involve too many conditions (room of home, location on property, season of year, day of week, hour of day, window-opening conditions, activity during exposure, number of noise events, peak levels of noise events, etc.); (2) a long list may lead respondents toward objective assessments of sound exposure levels and away from subjective feelings about exposures; (3) a long, complex question may confuse some respondents who will resolve the complex task by just answering for one condition, perhaps the first or last condition mentioned, while ignoring their most important, but seemingly insufficiently sophisticated, general subjective response; (4) a long list of conditions is more difficult to adapt to different cultures, languages and types of buildings; (5) a long question is less likely to be included in many surveys.

## 5.2. THE TYPE OF ANSWER SCALE

The initial goal for this project was to choose a single answer scale. After several international workshops, however, both the 5-point verbal scale and the 0–10 numeric scale were chosen.

### 5.2.1. *A verbal and a numeric scale*

Each of the scales has a different strength. The verbal scale is needed for the clearest, most transparent communication between respondents and policy makers or other readers of social survey reports. The simple task of choosing a word is most likely to be easily performed by respondents of any degree of sophistication in any culture. The resulting selected word is, when presented in a report, simply passed on to readers as the respondent’s choice. The protocol for choosing the words (described in section 5.4) attempts to ensure that the commonly understood meaning of the word is consistent with its position on the scale.

The numeric scale is felt to provide greater assurance that the scale points are equally spaced and thus meet the assumptions for linear regression and similar powerful analysis techniques that can represent the continuous range of responses to noise. The numeric scale also reduces the possibility of distortion by idiosyncratic interpretations of the verbal labels for scale points.

### 5.2.2. *Two scales*

Having more than one scale for a cross-survey comparison and, in fact, all analysis purposes is consistent with the most basic principles of increasing the reliability of psychometric measurements [30–32]. In addition a second scale provides some assurance against translation difficulties and provides the previously listed strengths of both verbal and numeric scales.

### 5.2.3. *A 5-point verbal scale*

Verbal scales of six points or more were rejected because such long scales were judged to be too cumbersome for telephone interviews (at least for a unipolar scale). Scales of three points were rejected as not providing a sufficient range of alternatives. On a 3-point scale there would only be two degrees of annoyance for those who were other than “not at all annoyed”. In the absence of empirical data, the standard psychometric criteria of reliability and validity could not be used in selecting between 4- and 5-point scales. Although both scale lengths have been used in previous noise surveys, the effects of length cannot be evaluated with noise-annoyance data because scale length is confounded with other differences between surveys and with wording differences in the question stems. As a result five other criteria were considered, upon which 5-point scales were slightly better on two criteria and equivalent on the remaining three.

5.2.3.1. *Consideration A—resolution of scale.* Logic and some research [33] suggest that five points provide additional resolution by allowing respondents to give a more finely graded, exact response than a 4-point scale.

5.2.3.2. *Consideration B—mitigation of end-of-scale scoring effects.* Respondents who tend to avoid ends of scales are given an additional usable rating point on a 5-point scale. In addition the 5-point scale slightly reduces the differences between those respondents who assume that the top word (“extremely”) represents the endpoint (an intensity of 100 per cent on an underlying 0–100 per cent scale) and those who assume the top word is the mid-point of an interval (an intensity of 90 per cent for the top fifth of the 5-point scale compared to 87.5 per cent for a 4-point scale). The recommendation that the top two points on a 5-point scale be combined to measure “highly annoyed” further mitigates these end-of-scale scoring distortions for the 5-point scale.

5.2.3.3. *Consideration C—rating of best 4- and 5-point scales in modifier study.* The standard annoyance modifier study, to be described in section 5.4, applies three criteria to choose the best verbal modifiers for one 4-point and one 5-point scale for each of nine languages. The three criteria are: (1) extent to which the available words are equally spaced on the annoyance continuum; (2) extent to which respondents agree on the location of the words on the annoyance continuum; (3) percentage of subjects who prefer the usage of the words for a scale. These same three criteria were used to compare the presumed quality of the best 4-point and best 5-point scale that could be identified for each language. These comparisons



did not show a preponderance of evidence for either the 4- or 5-point scale. In eight of the nine languages, some criteria supported one scale and some the other (only in French did all three criteria support a single scale, the 5-point scale).

5.2.3.4. *Consideration D—previous scale usage.* No previous, widely used verbal scales could be accepted on the basis of the present research. The words in the most widely used 4-point verbal scale (“not at all, a little, moderately, or very”) [34, 35] were not the words chosen using the 4-point scale criteria applied for this study (“not at all, somewhat, significantly, extremely”). The intensity scores for the previous scale’s words of 1-13-44-76 were systematically lower than the equidistance criterion of 0-33-67-100. Similarly, no previous widely used, 5-point verbal scale provided a basis for choosing between scales. Three studies have used the same five words in a similar annoyance scale [36–38]. The results from several additional studies using the same words are not comparable, however, because they used a two-part annoyance question in which respondents were first screened with a 2-point scale (“annoyed or not”) and then, if they indicated annoyance, were asked for an annoyance response for the remaining four points [39, 40: pp. 1056–1057].

5.2.3.5. *Consideration E—skewed scale distributions.* Four-point scales are sometimes seen as preferable to 5-point scales that might have middle alternatives that could encourage a heaping of responses on that middle answer. The noise-scale research reviewed in section 2.2 indicates, however, that this is not an important issue for noise reaction questions.

*Eleven-point (0–10) numeric scale.* A 0–10-point scale is likely to be easily understood by peoples of all countries and cultures who are familiar with currency in a base-10 monetary system and other familiar counting situations. The mid-point of a 0–10 scale is readily and correctly assumed to be “5” unlike the mid-point of a 10-point scale (1–10) that is 5.5 not 5.0.

*Scale card visual aids.* Visual displays of answers shown with the recommended question in section 4.1 are used in self-administered interviews and are recommended as an aid for personal interviews even though they could not be used in most telephone interviews. The 0–10 scale is presented as a simple, equally divided line rather than as a more complex graphic such as a thermometer because of concerns that an artistically rendered thermometer with a large bulb or other graphic device might not appear to have equal intervals. The line is arrayed horizontally to reduce the space needed in a self-administered questionnaire. For the 5-point verbal scale, the words are visually presented vertically with the most intense response word at the top. The words are vertically arranged because the varying length of the words would mean that a horizontal placement could not be equally spaced. The words are not identified with numbers because numbers might distort respondents’ judgments away from the words’ intrinsic intensity scores and thus also away from the intensity scores that are assumed to apply by the report readers who base their interpretation on only the words. For example, the choice of 0 or 1 for the lowest scale point might alter the meaning of “not at all annoyed” of this unipolar scale if “0” was interpreted as absolutely no annoyance while “1” was interpreted as 1/5th of the total annoyance scale in some cultures.

### 5.3. THE WORDING OF THE QUESTION STEM

Given the rationale in the previous section, the actual wording of each question in each language was carefully crafted. Back translations were performed by at least one individual for each question for each language before the final wordings were accepted.

To illustrate the details involved in the final decisions on wording, the English, 5-point verbal scale is divided into phrases in the next paragraph in bold italics with explanations in square brackets following each phrase or choice of words.

***“Thinking about the last (.12 months or so.), when*** [#1: The indefinite “thinking” and “12 months or so” encourages a general response to the noise, rather than an exclusive, comparison of the last 12 months with any other period.] ***you*** [#2: The respondent’s own reaction, not that of family members, is requested.] ***are*** [#3: The habitual, present tense of the verb, “are”, encourages the habitual, general response as explained in #1.] ***here at home,*** [#4: This phrase is intended to measure the general evaluation for the respondent’s dwelling environment while excluding the broader neighborhood shopping and recreation areas (as might be suggested by “around here”) and not strictly limiting answers to inside the building (as would be implied by “in your house”).] ***how much*** [#5: This phrase prepares the respondent for choosing an answer of degree of response.] ***does*** [Habitual present tense—see #3.] ***noise*** [#6: The single word “noise” rather than the phrase “the noise” is used to avoid the implication that such noise must be present. “Noise” is used rather than a neutral word for the reasons given in section 5.1.] ***from (... noise source...)*** [#7: The name of the noise source is specified not left unclear.] ***bother, disturb, or annoy*** [#8: These three words were judged to be necessary to convey the general impression of a negative reaction in English.] ***you:*** [Own reaction reinforced—See #2.] ***Extremely, Very, Moderately, Slightly*** [#9: These four words were selected by the protocols contained in the empirical study described in the next section.] ***or Not at all?”*** [#10: This phrase was found to have the lowest annoyance intensity rating in several studies [14, 41].]

The wording for the 0–10 numeric scale was equally carefully considered. The numeric scale question was constructed with about twice as many words because the concept of a numeric, 0–10 scale is more difficult to explain, especially when a visual device cannot be used in a telephone interview. When a visual answer card can be used, careful attention is needed to its construction. For the numeric scale, for example, the scale points are equally separated and the labels for the endpoints are centered on their scale points and are in a sufficiently small font that they do not overlap other scale points.

#### 5.4. RESEARCH TO SELECT THE RESPONSE MODIFIER WORDS FOR THE ANSWER SCALES

This section describes the standardized research project that choose the labels for the answers to the 5-point verbal scale and for the endpoints of the 0–10 numeric scale for each language. The project draws heavily on techniques originally designed for scaling of modifiers generally [42] and then developed for use with noise surveys in 1966 in Germany [43] and later in the United States [41]. This section describes the four steps followed in all languages with examples drawn primarily from the English language study.

##### 5.4.1. *Forming a pool of test words*

The group of participating researchers from each country created a final list of 21 candidate response modifiers for evaluation in their language. In English a longer list was initially developed by brainstorming about possible modifiers and listing modifiers that had been used in noise annoyance surveys or that had been examined in previous response modifier experiments on noise [14, 41] or on any subject matter [44, 45]. To meet minimal standards for clarity of meaning in a questionnaire for oral administration to the general population these long lists of modifying phrases were screened using the following six criteria: (1) the phrase is sufficiently short to be understood when read in a list of five

modifiers (e.g., exclude: “very definitely extreme”); (2) the phrase is used in common speech (e.g., exclude “unfathomly”); (3) the phrase does not mix positive and negative modifiers (i.e., exclude phrases such as “not too strongly” or “not very much” because they might easily be misunderstood); (4) the phrase is a modifier, not an unmodified restatement of the basic response (e.g., exclude “annoyed” or “disturbed” without a modifier); (5) the modifier describes a feeling and not an extreme behavioral reaction or level of adaptation (e.g., Both “unbearable” and “unacceptable” contain the logical contradiction that the resident is living with the noise); (6) the modifier does not use the superlative form of an adverb (e.g., both “the absolute worst” and “the most awful” are not likely to be literally true and thus may measure respondents’ tendencies toward hyperbole rather than their feelings). These criteria excluded many items that had been used in previous adverbial modifier studies. The investigators also conferred with colleagues on the content of the list.

#### 5.4.2. *Collecting data*

Empirical data were collected on the 21 words in each language by having subjects in each language follow the same protocol to evaluate the language’s words. The protocol was initially written in German and English by native speakers of both languages at the Department of Psychology, Ruhr University in Bochum, Germany. The protocol was then sent to each of the language teams and translated into the remaining seven languages. The protocol addressed subject recruitment procedures, subject briefing instructions, oral instructions to be read to the subjects and the wording of the instructions for the questionnaire.

Subjects included a mixture of university students and employees of technical firms. The average age was about 35 years, but varied from 19 to 44 for different study sites. After providing some background information the subjects completed the questionnaire by performing the following four tasks to evaluate the 21 words:

**Task #1: Intensity grouping.** Subjects placed each word in one of nine groups ranked from no annoyance to “... the most annoyance you can imagine” (This introduced respondents to the words, but the results were not used in the analysis.)

**Task #2: Intensity scoring.** Subjects indicated the intensity associated with each word by placing the word on its own 10-cm line that extended from “No/lowest degree of annoyance” to “Highest degree of annoyance.”

**Task #3: 5-point preference question.** Subjects selected one preferred word for each of the scale points by first choosing a word “that you would be most likely to use” for the “greatest amount of bother or annoyance you might feel” and then expressing a preference for the three words that should complete the remaining three points on a 5-point scale. (The lowest point was predetermined.)

**Task #4: 4-point preference question.** Subjects selected one word for each of the 4-point scale points by expressing a preference for the two words that should complete the remaining two points on a 4-point scale. For both the 4- and 5-point preference questions subjects were instructed to choose words that “people would normally use when talking”. Subjects were instructed to select words that were “equally spaced” between “not at all annoyed” and the previously chosen high annoyance word.

The questionnaires were completed by 1754 subjects at over 25 sites in 12 countries in nine languages (Dutch/Flemish = 93, English = 231, French = 45, German = 61, Hungarian = 47, Japanese = 1102, Norwegian = 56, Spanish = 59, Turkish = 60). The data from the experiments were then analyzed at two central locations as well as at most of the participating laboratories. At least 12 reports have been written about the results for specific languages [46–57].

#### 5.4.3. Analyzing data

A separate but identical analysis was conducted for each language. The result of this analysis was a series of eight ratings for each word that could in turn be used to apply the IC BEN protocol for selecting one word for each scale point. The eight ratings are divided between three for the 5-point scale analysis, three for the 4-point scale analysis and two that would be used for both analyses. The eight ratings are presented for the English language 5-point scales in Table 4 and Figures 2 and 3. Results for the remaining languages are given in Appendix A. The eight ratings are the following:

**Intensity score average:** the average of the positions in which subjects marked the word on the 10-cm scale when the marks are scored in millimeters (0–100). The average is computed after the distance is expressed in millimeters from 0 to 100. The intensity score is given in the first row ( $\bar{x}$ ) of Table 4 for each of the 21 English words. Intensity scores for each English study site are given in Figure 2. The 95 per cent confidence interval for each site's score is marked by the line extending from each site's symbol in Figure 2.

**Intensity score standard deviation:** the root mean square of the intensity scores. A large number indicates a lack of agreement on a word's position. The standard deviation ( $\sigma$ ) is given in the second row of Table 4.

**Scale point candidacy: 5-point scale:** the single scale point (#1, #2, #3, #4, or #5) for which the word is a candidate. This is the scale point for which the word was most often chosen as the preferred word on a 5-point scale. Each word's candidacy is shown by its grouping in Table 4 and Figures 2 and 3.

**Scale point candidacy: 4-point scale:** same as preceding definition but 4-point scale. These 4-point scale results are given in the lower portion of Table 4.

**Net preference score: 5-point scale:** percent of subjects preferring the word for that word's "candidate" position decreased by the percent preferring the word for other position(s). The net preference score may be negative if a word is chosen for more than two positions. The net preference score for the complete sample appears in the "Net.Pref. %" row in the middle of Table 4 and, for each site, in Figure 3 together with the 95 percent confidence interval for that score.

**Net preference score: 4-point scale:** same as preceding definition but the 4-point scale is in the last row of Table 4.

**Difference from scale point intensity criterion: 5-point scale:** the difference between the word's intensity score and the intensity criterion for that word's candidate scale point. The intensity criteria for the 5-point scale are 0, 25, 50, 75, 100 for points #1–#5 respectively. Differences are given in the " $\Delta = \text{IC-Inten}$ " (Intensity Criterion—Intensity score) row of data in Table 4.

**Difference from scale point intensity criterion: 4-point scale:** same as preceding definition but 4-point scale with the 4-point criteria of 0, 33.3, 66.7 and 100 with results being given in the next-to-the-last row of Table 4.

#### 5.4.4. Selecting words

The scale point candidacy ranking described above resulted in the pools of candidate words for each scale point that are indicated in Table 4 and Figures 2 and 3. The analysis then proceeded by choosing among the words in each pool. The best candidate was selected from the pool based on three scores: "Net preference score", "Difference from scale point intensity criterion", and "Intensity score standard deviation". In the simplest case, the same word was best on all three scores. For example, the selection of "extremely" from among the three candidates for the fifth point of the 5-point scale is seen to be such a case in Table 4 where "extremely" is closest to the intensity score (" $\Delta = \text{IC-Inten}$ " = - 5),

TABLE 4

Reaction modifier study results for 230 English-language subjects at 5 sites (England, Australia (2 sites) and U.S.A. (2 sites))

Modifier word candidates (Words chosen for the 5-point verbal scale are <u>underlined</u> )																				Means		
<u>Not</u>	Insig	Barel	Hardl	A	<u>Sligh</u>	Parti	Somew	Fairl	<u>Moder</u>	Rathe	Impor	Consi	Subst	Signi	<u>Very</u>	Highl	Stron	Sever	Treme	<u>Extre</u>	best	
<u>at</u>	nific	y	ly	littl	<u>tly</u>	ally	hat	y	<u>ately</u>	r	tantl	derab	antia	fican	<u>_____</u>	y	gly	ely	ndous	<u>mely</u>	scale	
<u>all</u>	antly			e	<u>_____</u>				<u>_____</u>	y	ly	lly	tlly	<u>_____</u>				ly	<u>_____</u>			
Inten	$\bar{x}$	.77	7.57	8.12	10.28	13.19	15.43	29.62	35.73	40.48	43.72	47.89	65.09	62.24	64.50	67.22	75.62	78.69	79.73	90.74	92.25	94.85
sity	$\sigma$	4.96	8.64	8.05	12.43	8.06	9.35	12.97	15.25	14.94	10.93	17.20	14.32	16.99	15.31	14.21	12.05	10.75	9.44	11.40	9.42	8.68
5-point scale																				Best		
Grouping of modifier candidates for 5-Point scale																				4-5-pt:		
Scale Pt.#	Pt.#1	Point #2 of 5					Point #3 of 5					Point #4 of 5					Point #5 of 5	four				
Intensity	0.00	Intensity Criteria (IC) = 25.00					I.C. = 50.00					Intensity Criteria (IC) = 75.00					IC = 100.00	$\bar{\sigma}$ =10.2				
$\Delta$ =IC-Inten.	1	-17	-17	-15	-12	-10	5	11	-10	-6	-2	15	-13	-11	-8	1	4	5	-9	-8	-5	$\Delta$ =5.40
Net Pref. %	100.0	0.87	6.06	7.79	22.94	36.36	0.87	-0.43	11.69	34.20	12.12	0.43	5.63	4.33	7.79	20.35	21.21	10.39	22.08	13.85	59.74	$\bar{x}$ =38
5-point scale																				Best		
Grouping of modifier candidates for 4-Point scale																				4-pt:		
Scale Pt.#	Pt.#1	Point #2 of 4					Point #3 of 4					Point #4 of 4					three					
Intensity	0.00	Intensity Criteria (IC) = 33.33					Intensity Criteria (IC) = 66.67					IC = 100.00					$\bar{\sigma}$ =12.7					
$\Delta$ =IC-Inten.	1	-26	-25	-23	-20	-18	-4	2	7	10	-19	-2	-4	-2	1	9	12	13	-9	-8	-5	$\Delta$ =2.70
Net Pref. %	100.0	1.73	2.16	2.60	9.52	16.02	9.09	15.58	12.55	9.52	.87	1.30	20.35	10.39	15.15	11.69	9.52	4.33	22.08	12.12	60.61	$\bar{x}$ =30

\*For a 4-point scale the protocol selects: Point #1=not at all, Point #2=somewhat, Point #3=significantly, Point #4=extremely.

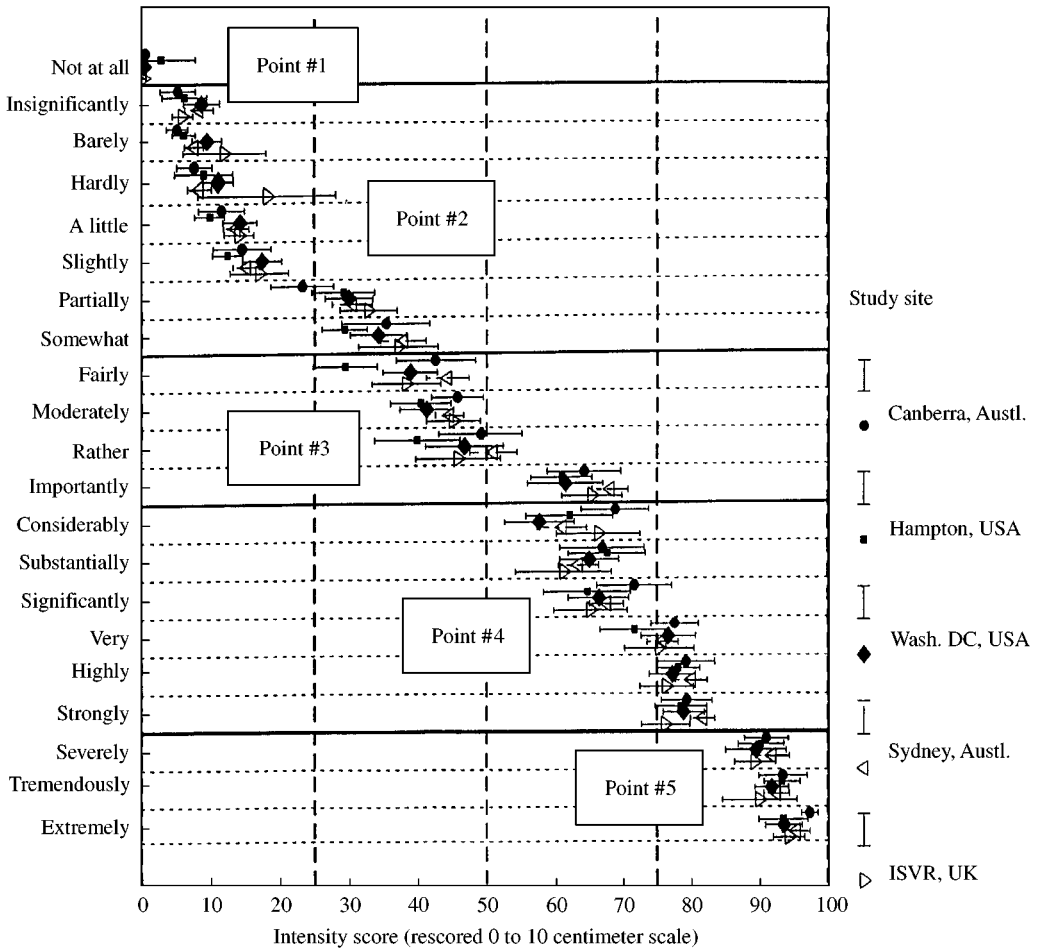


Figure 2. Intensity scores for 21 English words grouped by 5-point scale group (vertical lines are 5-point scale criteria of 0, 25, 50, 75, 100).

has the highest net-preference score at 59.74 per cent, and has the lowest, standard deviation (8.68).

When the different criteria favored different words, however, a more complex scheme was used to select the best word after eliminating weaker candidates from that scale point's eligible pool. The best word was the word remaining after words were eliminated one by one as each failed a criterion at one of the following 13 successive steps: (Step #1) Net preference score  $\geq 5$  per cent; (2) Unsigned difference from scale point intensity criterion  $< 15$  (e.g.,  $\Delta = \text{IC} - \text{Inten.} < 15$ ); (3) Net Preference score within 20 points of most popular remaining candidate word for the scale point (e.g.,  $\Delta\% \text{ Pref} < 20$ ); (4) Standard deviation within 15 points of smallest remaining modifier's standard deviation (e.g.,  $\Delta\sigma = < 15$ ); (5)  $\Delta = \text{IC} - \text{Inten.} < 10$ ; (6)  $\Delta\% \text{ Pref} < 15$ ; (7)  $\Delta\sigma = < 10$ ; (8)  $\Delta = \text{IC} - \text{Inten.} < 5$ ; (9)  $\Delta\% \text{ Pref} < 10$ ; (10)  $\Delta\sigma = < 5$ ; (11) Select the remaining word closest to intensity criterion; or if tied, (12) Select highest remaining preference score; or if tied, (13) Select lowest remaining standard deviation score. The application of these more complex criteria can be illustrated with the selection of "Slightly" for point #2 of the 5-point scale. Of the seven candidate words three are eliminated at step #1 because they all have net preference scores of less

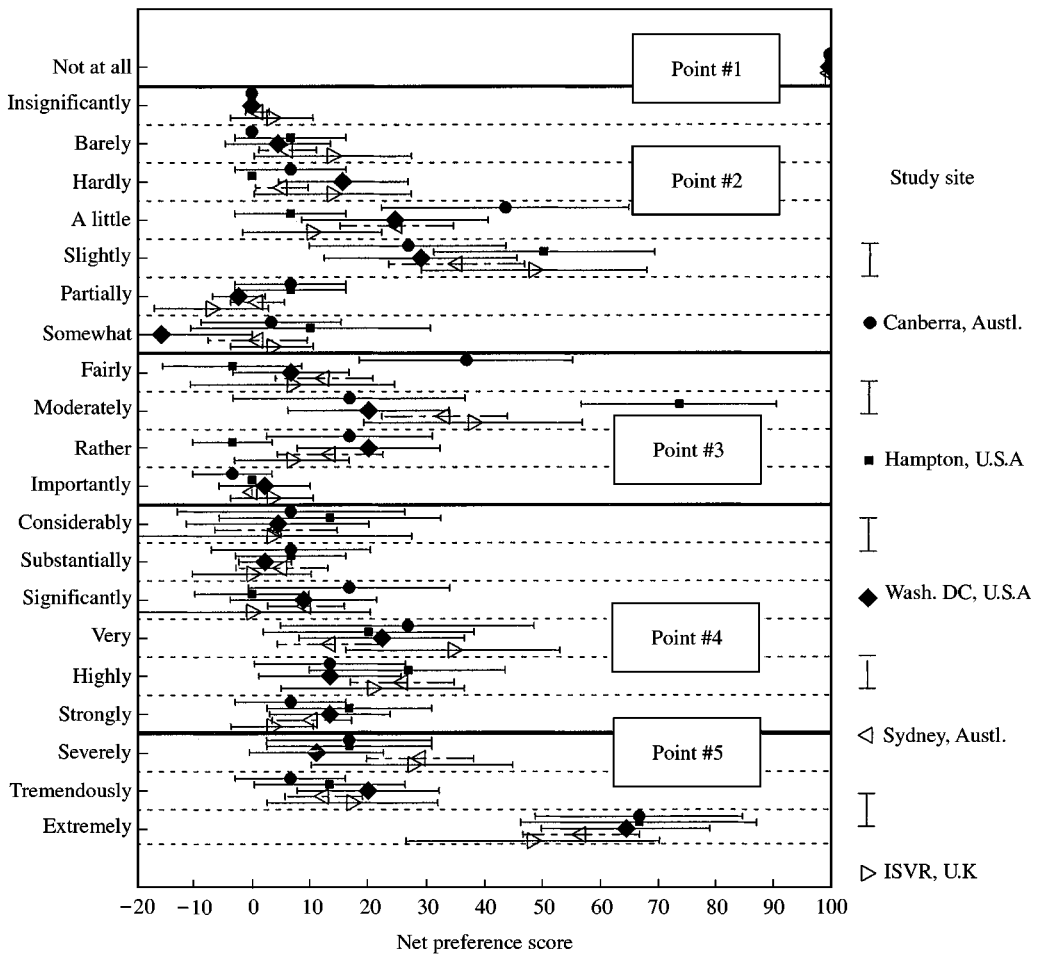


Figure 3. Subjects' preferences for 21 English words for a 5-point scale.

than 5 (Insignificantly, Partially and Somewhat). Next "Barely" is eliminated at Step #2 because its intensity score of 8.12 is more than 15 points ( $\Delta = 17$ ) from the criterion of 25.00. Next "Hardly" is eliminated at Step #3 because its net preference score of 7.79 is more than 20 points below that for "Slightly" (score of 36.36), the most popular remaining word ( $\Delta\% \text{ Pref} = 28.57$ ). Step #4 does not eliminate either of the remaining words but "A little" is eliminated at Step #5 because its intensity score of 13.19 is more than 10 points ( $\Delta = 12$ ) from the criterion of 25.00. This leaves one word, "Slightly", that is therefore selected as the word for point #2 on the 5-point scale.

The rules for the selection procedure also permitted words to be eliminated if the investigators for a language decided, based on empirical data or their professional judgment, that a word either (1) was either extremely awkward linguistically, or (2) had a regulatory or other meaning that could cause the word to be misinterpreted, or (3) would receive different intensity scores from members of different age groups, cultural groups or other groups. In addition, the investigators examined the selected words and considered whether the juxtaposition of the independently selected modifiers might collectively change the meanings of the modifiers so that the subjects' original judgments would no longer be

valid. These more subjective criteria were invoked in only two instances noted in footnotes to the German and Hungarian tables in Appendix A

## 6. DEFINITION OF HIGH ANNOYANCE

The intensity score results from the annoyance modifier study (defined in section 5.4.3) support a recommendation that the top two points (“very” and “extremely”) of the 5-point verbal scale be combined to define the widely used noise impact indicator of “percentage highly annoyed”. The intensity scores for 10 words within 20 units of the “highly” intensity score (79 units) are the following: “considerably” (62), “substantially” (64), “importantly” (65), “significantly” (67), “very” (76), “highly” (79), “strongly” (80), “severely” (91), “tremendously” (92), and “extremely” (95). “Very” is the closest of the five chosen scale words to “highly”. Both words are also close to 75, the intensity criterion for the fourth point on the intensity dimension. These findings are the primary reason for recommending that “very” be used to define “high” annoyance on this 5-point verbal scale. Using two points (“very” and “extremely”), rather than one point to define “high” annoyance also has the advantage of reducing the effects of any tendencies to avoid or select the endpoints of scales. This is also the division point that was used in the 1978 article in which the “percentage highly annoyed” was proposed as a community response measure [3, p. 399]. This is less severe than the 72 per cent cutting point that was proposed for numeric scales in that article since these top two points could be considered to exceed only 60 per cent of the 5-point scale. However, it should be noted that the 60 per cent division encompasses such words as “considerably” (62 per cent), “substantially” (64 per cent), and “importantly” (65 per cent); all of which indicate that the recommended high annoyance division identifies levels of annoyance that are not regarded as being trivial or moderate. No recommendation is offered here for a definition of “high” annoyance using the more abstract numeric scale, as the respondents’ answers do not provide a clear basis for a division. Empirical research comparing these verbal and numeric questions could provide a firm basis in the future.

## 7. CONCLUSIONS AND DISCUSSION

The two questions proposed in this paper provide the most carefully studied basis for comparing results from different surveys and languages. The study team therefore recommends that these two questions be included in future studies of reactions to community noise. In the course of developing the questions, however, the team has become aware of unresolved issues and changes in methods that could be explored if another group considers revisions in the future.

While the modifier-choice study methodology has been carefully designed, was widely discussed before being adopted and is recommended as the basis for establishing comparable scales, some aspects of the methods generated undesired results. Restricting respondents to the choice of only one modifier per scale point on the preference questions yielded poor measures of the extent to which several modifiers were equally acceptable. The results from the intensity grouping question (Task #1 in section 5.4.2) could not be used because of response errors that might have been eliminated by a revised protocol. Greater care might have been taken to ensure that the words, especially those for upper scale points, were ones that public opinion polling experts widely use in the particular language. The upper scale point is of special importance because subjects tended to choose the most extreme word for the upper scale point.



Some issues surrounding the choice of intensity criterion for annoyance words have not been fully resolved. It is not clear whether criteria for scale categories should be specified by their endpoints (0, 25, 50, 75, 100 for a 5-point scale) or some central tendency measure (e.g., mid-points of 10, 30, 50, 70, 90 for a 5-point scale). Although the lowest point has a conceptually unambiguous meaning, the interpretation of the upper point is subject to the extremity of the concept used. The central tendency criterion probably overestimates reactions at the lowest noise levels where answers in the lowest category may represent absolutely no annoyance. On the other hand, the endpoint criterion probably overestimates reactions at other noise levels where respondents to the highest point would not all place themselves at 100 per cent on an underlying intensity scale.

A method for combining the results from the two scales has not yet been suggested because it raises additional issues concerning the intensity score criteria (as discussed in the previous paragraph) and the ways in which respondents choose different scale points. Empirical research using results from these questions should soon provide a firm combination rule. Attempts to combine the results from these different scales on the basis of previous research are weakened by the likelihood that the central tendencies for annoyance levels within the same broad response category will change with noise level, the boundaries between scale points for different scales are not the same, the respondents may consider both the meaning of a word and position of a word (e.g., Table 1), the people in different cultures may differ in their relative sensitivity to a word's meaning and its scale position [58], factors other than intensity of annoyance influence the choice of answers (e.g., the uneven distributions in Table 2), and, for non-identical questions, the wordings of the question stem may affect the responses to the questions. The most sophisticated mid-point scoring approach to such comparisons that is currently available requires assumptions about these features and some investigator judgment [15, p. 3434]. The unresolved complexities in these comparisons are a major reason that this article has advocated the use of both of these standard, comparable questions for survey comparisons.

It is not known whether some of the details of the wording and presentation of the questions create small differences in responses. For example, answers might be affected by such features as: the order of presentation of answers (low ("not at all") to high ("extremely") or high to low), use of "noise" or "sound", visual graphics for scales, or the wording for the location ("here at home" or "around home"). By standardizing the method our approach has enhanced comparisons between surveys. Additional careful, multi-cultural international research on these particular issues could provide a firmer basis for evaluating the present decisions. In the absence of such research, the proposed questions provide the firmest available basis for making comparisons between answers in different surveys.

#### ACKNOWLEDGMENTS

This study was supported and reviewed by other researchers whose contributions matched those of the ICBEN team and University of Ruhr group who appear as joint authors. Other researchers who contributed to the study process by reviewing this article and conducting the study for the following languages included the following: For Dutch/Flemish: D. Botteldooren, and A. Verkeyn, University of Gent, Belgium; R. Derkx, Streekgewest Westelijke Mijnstreek, Geleen, The Netherlands; J. Gutteling, University of Twente, the Netherlands; For English: M. Burgess, Australian Defense Force Academy, Canberra, Australia; J. Hatfield, University of Sydney, Sydney; D. Perkins-Jones, Georgetown University, Washington, D.C.; K. Shepherd, NASA Langley Research Center, Hampton, Virginia. For French: Patricia Champelovier and Jacques Lambert,

INRETS-LEN, Lyon, France. For German: B. Schulte-Fortkamp, University of Oldenburg, Germany; For Spanish: Amando García and A. M. García, University of Valencia; M. Arana and A. Vela, Public University of Navarra. For Hungarian: A. Muntag, IEM Institute for Environmental Protection Budapest, Hungary; F. Hirka, National Institute of Hygiene, Budapest. For Turkish: Z. Karabiber, Yildiz Technical University, Turkey. For Japanese: J. Igarashi, Kobayasi Institute of Physics, Tokyo; K. Kanda, Kumamoto National College of Technology, Kumamoto; T. Kaneko, Kyorin University, Hachioji; S. Kuwano, Osaka University, Osaka; Y. Nii, Osaka City University, Osaka; T. Sato, Hokkai Gakuen University, Sapporo; M. So, Nihon University, Tokyo; I. Yamada, Kobayashi Institute of Physics, Tokyo; Y. Yoshino, Nihon University, Funabashi. B. Rohrmann, University of Melbourne, Australia, provided helpful insights at all stages of the research process.

This work was supported by the National Aeronautics and Space Administration Langley Research Center (contract NAS1-19061).

#### REFERENCES

1. J. M. FIELDS 1991 *NASA TM-187553*, National Aeronautics and Space Administration, Washington, D.C.: An updated catalog of 318 social surveys of residents' reactions to environmental noise (1943-1989).
2. S. FIDELL, D. S. BARBER and T. J. SCHULTZ 1991 *Journal of the Acoustical Society of America* **89**, 221-233. Updating a dosage-effect relationship for the prevalence of annoyance due to general transportation noise.
3. T. J. SCHULTZ 1978 *Journal of the Acoustical Society of America* **64**, 377-405. Synthesis of social surveys on noise annoyance.
4. B. ROHRMANN 1986 in *Contributions to Psychological Acoustics: Results of the Oldenburg Symposium on Psychological Acoustics* (A. Schick, H. Höge and G. Lazarus-Mainka, editors), 285-298. Comparability problems in the design of social-scientific research on environmental stressors.
5. D. M. GREEN and S. FIDELL 1989 *Journal of the Acoustical Society of America* **89**, 234-243. Variability in the criterion for reporting annoyance in community noise surveys.
6. E. JONSSON and P. ATTESLANDER 1963 *SR(63) 27, Organization for Economic Co-operation and Development (OECD), Directorate for Scientific Affairs, Paris*. Social surveys on reactions to airport noise.
7. B. ROHRMANN 1985 *Inter-noise* 85, 145-148. Guidelines for social-scientific noise research?: possibilities and problems.
8. S. NAMBA, J. IGARASHI, S. KUWANO, K. KUMO, M. SASAKI, H. TACHIBANA, A. TAMURA and Y. MISHINA 1996 *Journal of the Acoustical Society of Japan (E)* **17**, 109-113. Report of the committee of the social survey on noise problems.
9. J. M. FIELDS, R. DE JONG, A. L. BROWN, I. H. FLINDELL, T. GJESTLAND, R. F. S. JOB, S. KURRA, P. LERCHER, A. SCHUEMER-KOHRs, M. VALLET and T. YANO 1997 *Journal of Sound and Vibration* **206**, 685-695. Guidelines for reporting core information from community noise reaction surveys.
10. H. SCHUMAN and S. PRESSER 1981 *Questions and Answers in Attitude Surveys*. New York: Academic Press.
11. S. SUDMAN, N. BRADBURN and N. SCHWARZ 1996 *Thinking about Answers: The Application of Cognitive Processes to Survey Methodology*. San Francisco: Jossey-Bass.
12. P. P. BIEMER, R. M. GROVES, L. E. LYBERG, N. A. MATHIOWETZ and S. SEYMOUR 1991 *Measurement Errors in Surveys*. New York: John Wiley and Sons.
13. N. SCHWARZ, H. J. HIPPLER, B. DEUTSCH and F. STRACK 1985 *Public Opinion Quarterly* **49**, 388-395. Response scales: effects of category range on reported behavior and comparative judgments.
14. S. FIDELL and S. TEFFETELLER 1980 *BBN Report 4211*, Bolt Beranek and Newman, Cambridge, Massachusetts: Scaling annoyance for social surveys of community reaction to noise exposure.
15. H. M. E. MIEDEMA and H. VOS 1998 *Journal of the Acoustical Society of America* **104**, 3432-3445. Exposure-response relationships for transportation noise.

16. F. L. HALL, S. E. BIRNIE and S. M. TAYLOR 1979. Comparisons of response to road traffic noise and aircraft noise. Dept. Geography and Dept. Civil Engineering, McMaster Univ., Ontario.
17. N. BRADBURN 1983 in *Handbook of Survey Research* (P. Rossi, J. Wright and A. Anderson, editors), 289–328. New York: Academic Press. Response effects.
18. J. M. FIELDS and J. G. WALKER 1982 *Journal of Sound and Vibration* **85**, 177–255. The response to railway noise in residential areas in Great Britain.
19. J. GARNSWORTHY 1977 *M.Sc. Dissertation, Department of Social Statistics, University of Southampton, England*. A study of question order and wording experiments.
20. F. J. LANGDON 1976 *Journal of Sound and Vibration* **47**, 243–282. Noise nuisance caused by road traffic in residential areas: parts I, ii.
21. I. D. GRIFFITHS, F. J. LANGDON and M. A. SWAN 1980 *Journal of Sound and Vibration* **71**, 227–240. Subjective effects of traffic noise exposure: reliability and seasonal effects.
22. J. M. FIELDS 1986 *DOT/FAA/EE-86/10, Federal Aviation Administration, U.S. Department of Transportation, Washington, D.C.*: Cumulative airport noise exposure metrics: an assessment of evidence for time-of-day weightings.
23. J. M. FIELDS and J. G. WALKER 1980 *10th International Congress on Acoustics, Sydney, C2-9.2*. Community response to railway noise in Great Britain.
24. J. M. FIELDS, R. DE JONG, I. H. FLINDELL, T. GJESTLAND, R. F. S. JOB, S. KURRA, P. LERCHER, A. SCHUEMER-KOHRs, M. VALLET and T. YANO 1998 *Noise Effects '98*, 481–486. Recommendation for shared noise annoyance questions in noise annoyance surveys.
25. J. M. FIELDS 1996 *Inter-noise 96*, 2389–2391. Progress toward the use of shared noise reaction questions.
26. P. N. BORSKY 1954 *NORC Report No. 54, National Opinion Research Center*. Chicago: Community aspects of aircraft annoyance.
27. J. M. FIELDS and J. G. WALKER 1982 *Journal of Sound and Vibration* **85**, 177–255. The response to railway noise in residential areas in Great Britain.
28. J. MORTON-WILLIAMS, B. HEDGES and E. FERNANDO 1978 *Road traffic and the environment*. Social and Community Planning Research, London.
29. N. SCHWARZ and H.-J. HIPLER 1991 in *Measurement Errors in Surveys* (P. Biemer, R. Groves, N. Mathiowetz and S. Sudman, editors), 41–56. Chichester: Wiley. Response alternatives: the impact of their choice and presentation order.
30. J. P. GUILFORD 1954 *Psychometric Methods*. New York: McGraw-Hill.
31. J. C. NUNNALLY 1978 *Psychometric Theory*. New York: McGraw-Hill.
32. R. A. ZELLER and E. G. CARMINES 1980 *Measurement in the Social Sciences: The Link Between Theory and Data*. New York: Cambridge University Press.
33. E. P. COX 1980 *Journal of Marketing Research* **17**, 407–422. The optimal number of response alternatives for a scale: a review.
34. A. C. MCKENNELL 1963 *S.S.337, The Govt. Social Survey, Central Office of Information*, London: Aircraft noise annoyance around London (Heathrow) Airport.
35. MIL RESEARCH 1971 *Second Survey of Aircraft Noise Annoyance Around London (Heathrow) Airport*. London: HMSO.
36. S. FIDELL and G. JONES 1975 *Journal of Sound and Vibration* **42**, 441–427. Effects of cessation of late-night flights on an airport community.
37. S. FIDELL 1978 *Journal of the Acoustical Society of America* **64**, 198–206. Nationwide urban noise survey.
38. S. FIDELL, R. HORONJEFF, S. TEFFETELLER and K. PEARSONS 1981 *NASA CR-3490, National Aeronautics and Space Administration, Washington, D.C.*: Community sensitivity to changes in aircraft noise exposure.
39. P. D. SCHOMER 1983 *Journal of the Acoustical Society of America* **74**, 1773–1781. A survey of community attitudes towards noise near a general aviation airport.
40. S. FIDELL, R. HORONJEFF, J. MILLS, E. BALDWIN, S. TEFFETELLER and K. S. PEARSONS 1985 *Journal of the Acoustical Society of America* **77**, 1054–1068. Aircraft noise annoyance at three joint air carrier and general aviation airports.
41. N. LEVINE 1981 *Journal of Sound and Vibration* **74**, 265–279. The development of an annoyance scale for community noise assessment.
42. L. V. JONES and L. L. THURSTONE 1955 *The Journal of Applied Psychology* **39**, 31–36. The psychophysics of semantics: an experimental investigation.
43. B. ROHRMANN 1978 *Zeitschrift für Sozialpsychologie* 222–245. Empirische studien zur entwicklung von antwortskalen für die sozialwissenschaftliche forschung. (Empirical studies about the development of answer scales for social science research.)

44. B. M. BASS, W. F. CASCIO and E. J. O'CONNOR 1974 *Journal of Applied Psychology* **3**, 313–320. Magnitude estimations of expressions of frequency and amount.
45. P. E. SPECTOR 1976 *Journal of Applied Psychology* **3**, 374–375. Choosing response categories for summated rating scales.
46. A. GARCÍA, A. M. GARCÍA, M. ARANA and A. VELA 1998 *Portugal Congreso Ibérico de Acústica y XXIX Jornadas Tecnicacústica, Lisbon*, 107–110. Propuesta de una escala verbal para evaluar la molestia producida por el ruido ambiental en zonas urbanas. (Proposal of a verbal scale to evaluate the annoyance produced by the environmental noise in urban areas.)
47. T. YANO, K. MASDEN and K. KAWAI 1998 *Noise Effects '98: Noise as a Public Health Problem (Seventh International Congress)*, 519–522. A survey on Japanese and English descriptors of annoyance.
48. U. FELSCHER-SUHR, R. GUSKI and R. SCHUEMER 1998 *Noise Effects '98: Noise as a Public Health Problem (Seventh International Congress)*, 733–736. Some results of an international scaling study and their implications for noise research.
49. P. CHAMPELOVIER 1999, 9908, 1–67. *Conception d'une échelle de gêne due au bruit des transports: contribution française à l' "International scaling study"*. (Design of a transportation noise annoyance scale: French participation in the International Scaling Study). Bron: INRETS.
50. U. FELSCHER-SUHR, R. GUSKI and R. SCHUEMER 1998 *Inter-noise 1998*, 1067–1070. Constructing equidistant annoyance scales — an international study.
51. U. FELSCHER-SUHR, R. GUSKI and R. SCHUEMER 1998 *DAGA 1998, Zürich*, 80–81. Entwicklung einer international vergleichbaren belästigungsskala. (Development of an internationally comparable annoyance scale).
52. U. FELSCHER-SUHR, R. GUSKI, R. SCHUEMER and J. SCHULTE-PELKUM 1999 *Umweltpsychologie* **3**, 34–35. Internationale standardisierungsbestrebungen zur erhebung von lärmbelästigung—eine vorbereitende empirische untersuchung in zehn ländern. (International efforts to standardize the assessment of noise annoyance—a preparatory empirical study in ten countries.)
53. U. FELSCHER-SUHR, R. GUSKI and R. SCHUEMER 2000 *Zeitschrift für Lärmbekämpfung* **47**, 68–70. Internationale standardisierungsbestrebungen zur erhebung der lärmbelästigung. (International efforts to standardize the assessment of noise annoyance).
54. A. GARCÍA, A. M. GARCÍA, M. ARANA and A. VELA 1999 *XXX Jornadas Nacionales de Acústica en Avila, Sociedad Española de Acústica, Madrid*. Evaluacion de la molestia producida por el ruido ambiental. (Evaluation of annoyance produced by environmental noise).
55. M. INOUE, J. IGARASHI, K. KANDA, J. KAKU, T. KANEKO, S. KUWANO, Y. NII, T. SATO, M. SO, I. YAMADA, T. YANO and Y. YOSHINO 2000 *Westprac VII: Seventh Western Pacific Regional Acoustics Conference, Kumamoto, Japan*. Comparison of noise annoyance modifiers between generations and areas.
56. T. YANO, T. SATO, S. KUWANO, I. YAMADA, J. IGARASHI, J. KAKU, M. SO, Y. YOSHINO, T. KANEKO, K. SEKI and K. GOTO 1998 *Technical Report of the Noise and Vibration Committee of the Acoustical Society of Japan, N-98-35*. Sôon no urusasa no kyôtsû shakudo ni kansuru kokusai kyôdô kenkyû—nihongo no urusasa no hyôgengo ni kansuru jikken. (International joint study on a unified scale of noise annoyance: an experiment on the expression of the degree of annoyance.)
57. T. YANO, J. M. FIELDS and K. KANDA 1999 *Technical Report of the Noise and Vibration Committee of the Acoustical Society of Japan, N-99-47*. Sôon no urusasa ni kansuru teido hyôgengo no sedaikan hikaku—kyûshû de no ICBEN no kokusai kyôdô kenkyû. (Comparison of noise annoyance modifiers between generations — ICBEN's joint study in Kyushu.)
58. T. YANO, K. MASDEN, J. M. FIELDS, K. KANDA and K. KAWAI 1999 *Inter-noise 99*, 1331–1336. An experiment on the equivalence of noise annoyance scales in English and Japanese.

#### APPENDIX A: ANNOYANCE QUESTIONS FOR NINE LANGUAGES: RECOMMENDED WORDING AND RESULTS OF MODIFIER EVALUATION STUDIES

This appendix contains two types of information for each of the nine languages that have conducted the standard noise modifier study. First, the agreed-upon wording for the two proposed noise reaction questions is given. Second, the results from the annoyance modifier

study are presented in an identical format for each language. The format and definitions of terms that appear in the tables were explained in section 5.4.4 where the English results in Table 4 were discussed. The words that are chosen by the IC BEN study procedure for the 5-point verbal scale are underlined in the heading of the table. Footnotes indicate that investigator judgment was exercised in the choices of one word in the German scale and one word in the Hungarian scale.

## ENGLISH RESULTS

### RECOMMENDED QUESTIONS:

- Q.V. Thinking about the last (..12 months or so.), when you are here at home, how much does noise from (..noise source..) bother, disturb, or annoy you; Extremely, Very, Moderately, Slightly or Not at all?
- Q.N. Next is a zero to ten opinion scale for how much (..source..) noise bothers, disturbs or annoys you when you are here at home. If you are not at all annoyed choose zero, if you are extremely annoyed choose ten, if you are somewhere in between choose a number between zero and ten. Thinking about the last (..12 months or so.), what number from zero to ten best shows how much you are bothered, disturbed, or annoyed by (..source..) noise?

NOTE: The table of English results is not reproduced here as the results were reported in Table 4 in section 5.4.4.

### DUTCH (FLEMISH)

- Q.V. Wanneer u denkt aan de afgelopen (... 12 maanden of zo ...), in welke mate stoort of hindert het geluid van (... geluidbron ...) u als u hier, bij u thuis bent; extreem, erg, tamelijk, een beetje of helemaal niet?
- Q.N. Hier is een schaal van nul tot tien waarop u kunt aangeven in welke mate geluid u stoort of hindert als u hier thuis bent. Als u helemaal niet gehinderd wordt kiest u de nul, als u extreem gehinderd wordt kiest u de tien. Als u daar ergens tussenin zit, kiest u een getal tussen nul en tien. Als u denkt aan de afgelopen (... 12 maanden of zo ...), welk getal van nul tot tien geeft het beste aan in welke mate u gestoord of gehinderd wordt door geluid van (... geluidbron ...) als u hier thuis bent? (See Table A1).

### FRENCH

- QV: Si vous pensez aux (... douze derniers mois ...), quand vous êtes ici, chez vous, le bruit de (... source ...) vous gêne-t-il: extrêmement, beaucoup, moyennement, légèrement, pas du tout?
- QN: Voici une échelle d'opinion graduée de zero à dix. Vous devez noter sur cette échelle la façon dont le bruit de (... source ...) vous gêne lorsque vous êtes ici, chez vous: notez zero si le bruit ne vous gêne pas du tout et notez dix si le bruit vous gêne extrêmement. Si vous êtes entre ces deux situations, choisissez une note intermédiaire entre zero et dix. Maintenant, si vous pensez aux (... douze derniers mois ...), quand vous êtes ici, chez vous, quelle note comprise entre zero et dix exprime le mieux à la façon dont le bruit de (... source ...) vous gêne? (See Table A2).

TABLE A1

Reaction modifier study results for 93 Dutch\*Flemish-language subjects at three sites (Netherlands (2 sites) and Belgium)

Modifier word candidates (Words chosen for the 5-point verbal scale are <u>underlined</u> )																				Means		
																				for		
<u>helem</u>	niet	nauwe	weini	iets	licht	<u>een</u>	enigz	matig	<u>came</u>	behoo	aanzi	veel	<u>erg</u>	sterk	zeer	ernst	enorm	ontze	uiter	<u>extre</u>	best	
<u>aal</u>		lijks	g		elijk	<u>beet</u>	ins		<u>ijk</u>	rlijk	enlij				ig			ttend	mate	<u>em</u>	scale	
<u>niet</u>					<u>e</u>						k											
Inten	$\bar{x}$	.41	1.42	9.44	12.42	15.65	16.39	16.50	25.93	34.36	39.18	62.10	68.08	69.03	74.21	77.88	80.25	80.50	85.92	87.43	89.08	97.79
sity	$\sigma$	.71	2.61	7.69	6.52	10.34	9.99	9.44	13.50	13.92	14.71	16.98	15.73	11.96	10.81	10.56	8.74	10.18	9.86	9.32	10.30	2.71
5-point scale																				Best		
Grouping of modifier candidates for 5-point scale																				5-pt:		
Scale Pt.#	1st	Point #2 of 5					Point #3 of 5					Point #4 of 5					Pt #5 of 5					four
words																						
Intensity	0.00	Intensity Criteria (IC) = 25.00					I.C. = 50.00					I.C. = 75.00					IC = 100.00					$\bar{0}$ = 9.4
words																						
$ \Delta=IC-Inten.$	0	-24	-16	-13	-9	-9	-8	1	-16	-11	12	18	19	-1	3	5	5	-14	-13	-11	-2	$ \Delta=3.33$
words																						
Net Pref. %	98.92	-1.08	15.05	17.20	6.45	11.83	25.81	3.23	10.75	30.11	5.38	1.08	3.23	11.83	16.13	9.68	15.05	4.30	4.30	.00	73.12	$\bar{x}=36$
words																						
4-point scale																				Best		
Grouping of modifier candidates for 4-point scale*																				4-pt:		
Scale Pt.#	1st	Point #2 of 4					Point #3 of 4					Point #4 of 4					three					
words																						
Intensity	0.00	Intensity Criteria (IC) = 33.33					Intensity Criteria (IC) = 66.67					IC = 100.00					$\bar{0}$ = 10.8					
words																						
$ \Delta=IC-Inten.$	0	-32	-24	-21	-18	-17	-17	-7	1	6	-5	1	2	8	11	14	14	-14	-13	-11	-2	$ \Delta=1.55$
words																						
Net Pref. %	98.92	-1.08	1.08	17.20	4.30	9.68	19.35	16.13	16.13	7.53	17.20	9.68	17.20	4.30	11.83	12.90	5.38	9.68	2.15	1.08	74.19	$\bar{x}=33$
words																						

\*For a 4-point scale the protocol selects: Point#1=helemaal niet, Point#2=matig, Point #3=aanzienlijk, Point #4=extreem.

TABLE A2

Reaction modifier study results for 45 French-language subjects at one site

Modifier word candidates (Words chosen for the 5-point verbal scale are <u>underlined</u> )																		Means				
																		for				
<u>pas</u>	pas	presq	un	guère	un	<u>léger</u>	à	modér	<u>moyen</u>	assez	plutôt	<u>beauc</u>	vraim	très	forte	haute	consi	énorm	terri	<u>extrê</u>	best	
<u>du</u>		ue	petit		peu	<u>ement</u>	peine	ément	<u>nemen</u>		t	<u>oup</u>	ent		ment	ment	dérab	ément	bleme	<u>nemen</u>	scale	
<u>tout</u>		pas	peu															ement		nt	<u>t</u>	
Inten	$\bar{x}$	.22	1.09	9.42	12.67	13.33	16.18	18.11	11.82	42.41	50.29	52.40	53.51	74.40	75.96	79.51	84.11	84.36	92.31	94.09	98.02	99.24
sity	$\sigma$	.52	1.79	8.12	9.13	10.33	11.19	12.42	10.74	13.93	4.60	14.08	16.40	9.82	11.37	11.33	9.53	12.60	9.75	6.65	3.22	1.65
5-point scale																		Best				
Grouping of modifier candidates for 5-point scale																		5-pt:				
Scale Pt.#	1st	Point #2 of 5					Point #3 of 5					Point #4 of 5					Pt #5 of 5	four				
Intensity	0.00	Intensity Criteria (IC) = 25.00					I.C. = 50.00					Intensity Criteria (IC) = 75.00					IC = 100.00	$\bar{\sigma}$ = 7.1				
$\Delta$ =IC-Inten.	0	-24	-16	-12	-12	-9	-7	-38	-8	0	2	-21	-1	1	5	9	9	-8	-6	-2	-1	$\Delta$ =2.13
Net Pref. %	100.0	.00	2.22	17.78	8.89	26.67	37.78	2.22	2.22	73.33	11.11	2.22	20.00	22.22	22.22	24.44	.00	2.22	4.44	20.00	73.33	$\bar{x}$ =51
4-point scale																		Best				
Grouping of modifier candidates for 4-point scale*																		4-pt:				
Scale Pt.#	1st	Point #2 of 4					Point #3 of 4					Point #4 of 4					three					
Intensity	0.00	Intensity Criteria (IC) = 33.33					Intensity Criteria (IC) = 66.67					IC = 100.00					$\bar{\sigma}$ = 8.5					
$\Delta$ =IC-Inten.	0	-32	-24	-21	-20	-17	-15	-22	9	17	19	-13	8	9	13	17	18	-8	-6	-2	-1	$\Delta$ =5.86
Net Pref. %	100.0	.00	2.22	4.44	2.22	31.11	17.78	-2.22	17.78	6.67	.00	11.11	24.44	22.22	11.11	13.33	2.22	2.22	6.67	20.00	71.11	$\bar{x}$ =38

\*For a 4-point scale the protocol selects: Point#1=pas du tout, Point#2=modérément, Point #3=beaucoup, Point #4=extrêmement.

## GERMAN

- Q.V. Wenn Sie einmal an die letzten (... 12 Monate ...) hier bei Ihnen denken, wie stark haben Sie sich durch Lärm von (... Quelle ...) insgesamt gestört oder belästigt gefühlt: Äußerst, stark, mittelmäßig, etwas, oder überhaupt nicht?
- Q.N. Ich habe hier eine Messlatte von Null bis Zehn, auf der Sie angeben können, wie sehr Sie der Lärm von (... Quelle ...) insgesamt gestört oder belästigt hat. Wenn Sie sich äußerst gestört oder belästigt fühlten, wählen Sie die Zehn, wenn Sie sich überhaupt nicht gestört oder belästigt fühlten, geben Sie bitte die Null an, und wenn Sie irgendwo dazwischen liegen, wählen Sie eine Zahl zwischen Null und Zehn. Wenn Sie nun an die letzten (... 12 Monate ...) hier bei Ihnen denken, welche Zahl zwischen Null und Zehn gibt am besten an, wie stark Sie sich durch den Lärm von (... Quelle ...) insgesamt gestört oder belästigt fühlten? (See Table A3).

## HUNGARIAN

- Q.V. Tekintve az utóbbi (... időszakot, 1 évet ...) mennyire zavarja Önt a (... zajforrás ...) zaja, amikor otthon tartózkodik: rettenetesen, nagyon, közepesen, kissé vagy egyáltalán nem?
- Q.N. Képzelsen el egy 0-tól 10-ig terjedő skálát arról, hogy érzése szerint mennyire zavarja Önt a (... zajforrás ...) zaja, amikor otthon tartózkodik. Ha egyáltalán nem zavarja, válassza a 0-t, ha rettenetesen zavarja, válassza a 10-et, ha pedig a kettő közötti mértékben zavarja, válasszon egy megfelelő számot 0 és 10 között. Tekintve az utóbbi (... időszakot, 1 évet ...) milyen 0-10 közötti számmal jellemezné azt, hogy a (... zajforrás ...) zaja mennyire zavarja Önt? (See Table A4).

## JAPANESE

- Q.V. Kako (... 12 ka getsu kurai ...) wo furikaette, anata wa jitaku de (... sôon gen wo ireru ...) karano sôon de dono teido nayamasareru, aruiwa, jamasareru, urusai to kanjiru deshôka: hijôni, daibu, tashô, sorehodonai, mattakunai?
- Q.N. Tsugi wa, anata ga jitaku de (... sôon gen wo ireru ...) karano sôon de dono teido nayamasareru, aruiwa, jamasareru, urusai to kanjiru ka wo shimesu tame no 0 kara 10 made no sûji de arawashita shakudo desu. Moshi, anata ga mattaku urusaku nai to kanjiru nara 0 wo erande kudasai. Hijôni urusai to kanjiru nara 10 wo erande kudasai. Moshi, sono teido ga korera no aida no dokoka ni areba, 0 kara 10 made no sûji no uchi tekîtô na mono wo erande kudasai. Kako (... 12 ka getsu kurai ...) wo furikaette, anata ga (... sôon gen wo ireru ...) karano sôon de nayamasaretari, aruiwa, jamasaretari, urusai to kanjiru teido wo mottomo yoku arawasu no wa 0 kara 10 made no dono sûji deshôka?

(NOTE: The Japanese question is transliterated into an English alphabet using the Hepburn system. This is only one of several possible systems, each of which yields slightly different spellings and diacritical marks. The question can be downloaded in Japanese script from the site at this address: <http://acoust.arch.kumamoto-u.ac.jp/standard-questions.html>) (See Table A5).





TABLE A4

## Reaction modifier study results for 47 Hungarian-language subjects at two sites

Modifier word candidates (Words chosen for the 5-point verbal scale are <u>underlined</u> )																			Means			
<u>egyál</u>	nem	alig	picit	kicsi	némil	<u>kissé</u>	kevés	észre	mérsé	<u>közep</u>	jelen	megle	erőse	külön	<u>nagyo</u>	módfe	túlsá	mérté	végte	<u>rette</u>	best	
<u>talán</u>			it	eg	<u>_____</u>	sé	vehet	kelte	<u>esen</u>	tősen	hetős	n	ösen*	<u>n*</u>	lett	gosan	ktele	lenül	<u>netet</u>	scale		
<u>nem</u>					<u>_____</u>	ően	n	<u>_____</u>	en			<u>_____</u>		nül	<u>esen</u>							
Inten	$\bar{x}$	.74	1.11	7.34	10.38	12.57	12.77	13.47	15.94	28.19	32.21	45.11	62.85	67.17	72.28	74.21	76.32	79.89	88.30	90.21	96.62	96.98
sity	$\sigma$	1.31	1.66	5.75	7.22	8.36	8.46	8.75	7.93	16.42	11.73	8.00	21.04	18.88	18.09	15.75	15.50	12.98	8.64	11.80	5.14	4.83
5-point scale																			Best			
Grouping of modifier candidates for 5-point scale																						
Scale Pt.#	1st	Point #2 of 5					Point #3 of 5					Point #4 of 5					Pt #5 of 5	four				
Intensity	0.00	Intensity Criteria (IC) = 25.00					I.C. = 50.00					Intensity Criteria (IC) = 75.00					IC = 100.00		$\bar{O}$ = 9.3			
$\Delta$ =IC-Inten.	1	-24	-18	-15	-12	-12	-12	-9	-22	-18	-5	-12	-8	-3	-1	1	5	13	-10	-3	-3	$\Delta$ =5.06
Net Pref. %	100.0	.00	4.26	2.13	25.53	4.26	29.79	2.13	.00	6.38	40.43	12.77	2.13	10.64	8.51	14.89	8.51	6.38	8.51	17.02	61.70	$\bar{x}$ =35
4-point scale																			Best			
Grouping of modifier candidates for 4-point scale**																						
Scale Pt.#	1st	Point #2 of 4					Point #3 of 4					Point #4 of 4				three						
Intensity	0.00	Intensity Criteria (IC) = 33.33					Intensity Criteria (IC) = 66.67					IC = 100.00				$\bar{O}$ = 11.8						
$\Delta$ =IC-Inten.	1	-32	-26	-23	-21	-21	-20	-17	-5	-1	12	-4	1	6	8	10	13	22	-10	-3	-3	$\Delta$ =1.55
Net Pref. %	100.0	.00	4.26	.00	10.64	10.64	17.02	6.38	12.77	27.66	4.26	12.77	21.28	14.89	6.38	17.02	2.13	6.38	8.51	17.02	61.70	$\bar{x}$ =37

\* Hungarian-speaking researchers selected the second most highly rated word "nagyon" for point #5 on the scale. "Különösen", the most highly scored word, was rejected because it could have the meaning of "peculiarly" in this context.

\*\*For a 4-point scale the protocol selects: Point#1=egyáltalán nem, Point#2=mérsékelten, Point #3=meglehetősen, Point #4=rettenetesen.



## NORWEGIAN

- Q.V. Tenk etter på støysituasjonen de siste (... 12 månedene ...). Hvor plaget er du av støy fra (... støykilde ...) når du er hjemme? Er du voldsomt plaget, mye plaget, ganske plaget, litt plaget, ikke plaget.?
- Q.N. Angi på en skala fra null til ti hvor plaget du er av støy fra (... støykilde ...) når du er hjemme. Hvis du ikke er plaget, velger du null. Hvis du er voldsomt plaget, velger du ti. Hvis du vurderer støyplogen mellom disse grensene, velger du et tall mellom null og ti. Tenk etter på støysituasjonen de siste (... 12 månedene ...). Hvilket tall mellom null og ti angir hvor plaget du er av støy fra (... støykilde ...) når du er hjemme? (See Table A6).

## SPANISH

- Q.V. Tomando en consideración los últimos (... 12 meses ...), indique Vd. en qué cuantía le molesta o perturba el ruido producido por (... indicar la fuente de ruido ...) cuando se encuentra en su casa: extremadamente, muy, medianamente, ligeramente, absolutamente nada.
- Q.N. A continuación se da una escala de opinión de cero a diez para que Vd. pueda expresar en qué cuantía le molesta o perturba el ruido producido por (... indicar la fuente de ruido ...) cuando se encuentra en su casa. Por ejemplo, si Vd. está "absolutamente nada" molesto por el ruido debería escoger el cero, y si Vd. está "extremadamente" molesto debería escoger el diez. Tomando en consideración los últimos (... 12 meses ...), indique qué número desde el cero al diez expresa mejor la cuantía en que Vd. está molesto o perturbado por el ruido producido por (... indicar la fuente de ruido ...) (See Table A7).

## TURKISH

- Q.V. Yaklaşık son (... 12 ayı ...) düşündüğünüzde, (... gürültü kaynağından ...) gelen gürültü, burada evinizdeyken sizi ne kadar rahatsız etmektedir? Feci şekilde, çok, orta derecede, hafifçe, hiç değil?
- Q.N. Şimdi, burada evinizdeyken (..kaynak..) gürültüsünün sizi ne kadar rahatsız ettiğini "sıfır" ile "on" arasında sayılarla gösteren bir görüş (veya kanaat) ölçeği verilmektedir. Eğer hiç rahatsız değilseniz "sıfır"ı seçiniz, eğer feci şekilde rahatsız iseniz "on"u seçiniz, bunların arasında iseniz "sıfır" ile "on" arasında bir sayı seçiniz. Yaklaşık son (..12 ayı..) düşünerek (..kaynak..) gürültüsünden olan rahatsızlığınızı "sıfır"dan "on"a kadar hangi sayı en iyi gösterir? (See Table A8).

TABLE A6  
Reaction modifier study results for 56 Norwegian-language subjects

Modifier word candidates (Words chosen for the 5-point verbal scale are <u>underlined</u> )																				Means			
<u>ikke</u>	ubet	<u>litt</u>	noe	minim	moder	delvi	endel	nokså	forho	midde	<u>qansk</u>	temme	betyd	<u>mye</u>	meget	veldi	svært	sterk	alvor	<u>volds</u>	best		
_____	ydeli	_____		alt	at	s			ldsvi	ls	<u>e</u>	lig	elig	_____		g		t	lig	<u>omt</u>	scale		
_____	g	_____							s		_____		_____							_____			
Inten	x	.57	6.75	15.66	25.61	5.48	34.82	35.86	40.98	44.68	47.02	47.59	50.52	66.41	72.11	72.45	79.18	83.80	86.63	87.64	93.59	96.50	
sity	σ	1.39	<u>4.41</u>	<u>7.88</u>	11.80	4.15	13.39	11.75	14.67	14.49	14.18	6.34	14.95	12.03	16.29	9.24	9.49	11.75	9.04	7.48	7.24	5.77	
5-point scale																				Best			
Grouping of modifier candidates for 5-point scale																				+5-pt:			
Scale Pt.#	1st	Point #2 of 5	Point #3 of 5				Point #4 of 5				Pt #5 of 5	four											
words																							
Intensity	0.00	(IC) = 25.00	Intensity Criteria = 50.00				Intensity Criteria (IC) = 75.00				IC = 100.00	σ = 9.5											
Δ=IC-Inten.	1	-18	-9	1	-45	-15	-14	-9	-5	-3	-2	1	-9	-3	-3	4	9	12	13	-6	-4	Δ=3.98	
Net Pref. %	100.0	3.57	44.64	28.57	1.79	1.79	1.79	14.29	5.36	1.79	19.64	10.71	1.79	12.50	26.79	1.79	1.79	.00	3.57	32.14	35.71	x=29	
4-point scale																				Best			
Grouping of modifier candidates for 4-point scale*																				+4-pt:			
Scale Pt.#	1st	Point #2 of 4				Point #3 of 4				Pt #4 of 4	three												
words																							
Intensity	0.00	Intensity Criteria (IC) = 33.33				Intensity Criteria (IC) = 66.67				IC = 100.00	σ=11.3												
Δ=IC-Inten.	1	-27	-18	-8	-28	1	3	8	11	-20	-19	-16	0	5	6	13	-16	-13	-12	-6	-4	Δ=5.05	
Net Pref. %	100.0	1.79	23.21	32.14	.00	7.14	23.21	-19.6	1.79	1.79	1.79	7.14	7.14	19.64	32.14	5.36	1.79	7.14	8.93	33.93	35.71	x=26	

STANDARDIZED NOISE REACTION QUESTIONS

\*For a 4-point scale the protocol selects: Point#1=ikke, Point#2=delvis, Point #3=betydelig, Point #4=voldsomt.



