BERND ROHRMANN

Verbal qualifiers for rating scales:
Sociolinguistic considerations and psychometric data

Project Report
University of Melbourne/Australia - Jan 2007
Available at http://www.rohrmannresearch.net/pdfs/rohrmann-vqs-report.pdf

Abstract

Questionnaires are the dominant data collection method in psychology and the social sciences in general, and most use rating scales as response mode. Within category scaling, verbal labelling of rating scales has become the primary approach to enhancing useability. The labels are used as "qualifiers", either for endpoints or for each single scale point. Verbal labelling provides practical advantages, such as ease-of-explanation and familiarity, and facilitates capturing normative judgments. The main disadvantages are inferior measurement quality and proneness to cultural biases. It is thus essential to design verbalized scales carefully if equi-distant and unambiguous instruments are to be achieved - yet only few publications provide pertinent information.

The principal idea underlying the current research is: to create rating scales using verbal labels which reflect the cognitions of respondents and for which psycholinguistic and psychometric data are available. Therefore a series of studies was conducted to clarify the measurement features of relevant words and to develop methodologically sound response scales which are useful for basic and applied research. Altogether 100 words or expressions were tested within five qualifier dimensions: Intensity, frequency, probability, quality, and responses to statements. Their properties were investigated with several categorical scaling and magnitude estimation methods in a variety of contexts. The samples for 6 sub-studies (N=229) were recruited from students and the general population.

The results provide a comprehensive body of quantitative information about common scale labels and enable the systematic design of response formats. The recommended format is multi-modal to enhance both psychometric quality and user-friendliness. To widen the validity scope, further research is underway, including cross-national extensions.

Contact Address:

Professor B. ROHRMANN
Roman Research Road, 94 Fenwick St, Calrton-Nth, Vic 3054, AUSTRALIA
E-Mail mail@rohrmannresearch.net

Research Issue

Rating scales in the social sciences

Various types of questionnaires are by far the most-used method of data collection in psychology and other social sciences, and about all of them use rating scales as primary response mode. Countless articles have followed the seminal work of authors such as Freyd (1923), Thurstone (1928), Likert (1932). A response scale should fulfil psychometric standards of measurement quality as well as practicality criteria, such as comprehensibility for respondents and ease of use. Rating scales are so popular because of their convenience - they are easy to explain and produce straightforward data; but they are also questionable because of serious shortcomings in their measurement features.

Presenting and labelling scale points

Usually rating scales (category scales in psychometric terms) offer between 4 and 11 response alternatives, i.e., ordinal scale points which are supposed to be equidistant (for overviews of response scales and scaling in general see, e.g., Cox 1980, Krosnick & Fabrigar 1997, Dawes & Smith 1985, Foddy 1992, Haertel 1993, McIver & Carmines 1993, Spector 1993). Numbers or words or graphic symbols (or a combination thereof) can be used to denote the categories, but verbal labelling has become the dominant approach to facilitate communication. Either words or short expressions are used, e.g., "never/seldom/sometimes/ often/always", "not/slightly/fairly/quite/very", "bad/poor/fair/good/excellent","strongly-disagree/ disagree/ undecided/agree/ strongly-agree". Instead of labelling every point, only the scale endpoints may be verbalized, e.g., "not-at-all"..."extremely" or "never"..."always" for a 0..10 scale. A common mode of rating scales is based on the combination of words describing a substantive attribute or behavior and various levels of that dimension, e.g.: never/ sometimes/often/always successful (in linguistic terms: combining an adjective with adverbs). How scale points are denoted is very likely to affect response behavior (cf. e.g. Christian & Dillman 2004, Dixon et al. 1984, French-Lazovik et al. 1984, Freyd 1923, Hartley et al. 1984, Klockars & Yamagishi 1988, LeBlanc et al. 1998, Lehto et al. 2000, Moxey & Sanford 1991, Traenkle 1987, Wildt 1978).

The psychometric function of verbal labels can be understood as "qualifier" (cf., e.g., Spector 1976), but various other terms have been used as well, including anchor (Jones & Thurstone 1955), quantifier (e.g., Newstead & Collins 1987, Zimmer 1988) or vague quantifier (e.g., Bradburn & Miles 1979), grader or modifier (Rohrmann 1978), intensifier (e.g., O'Muircheartaigh et al. 1993), multiplier (e.g., Cliff 1959). In the present text, the neutral term *verbal scale point label* will be used, abbreviated by "VSPL". In spite of their ubiquitous use, scientific knowledge about the subjective understanding and metric properties of verbal labels is rather restricted. This is unfortunate as the wording is the main reason for measurement deficiencies (for a discussion of problems see, e.g., Andrews 1984, Hippler et al. 1991, Moxey & Sanford 1991, 1993, Nakao & Prytulak 1983, Newstead & Collins 1987, Parducci 1983, Pepper & Prytulak 1974, Poulton 1989, Presser & Blair 1994, Schwarz et al. 1993, Wegener et al. 1982). A core criticism is that rating scales are more prone to biasing context effects than other scaling techniques.

While quite a few studies investigated adverbs denoting extent or frequency and particularly probability phrases (Budescu & Wallsten 1994, Clark 1990, Clarke et al. 1992 <the only Australian study so far>, Cliff 1972, Diefenbach et al. 1993, Hammerton 1976, Jones & Thurstone 1955, Reagan et al. 1989, Rohrmann 1978, Theil 2002, Windschitl & Wells 1996, Wright et al. 1994), such findings were rarely systematically applied to *scale construction* (see however Rohrmann 1978, for verbally labelled rating scales in German language; Levine 1981, for an English noise annoyance scale.)

Because of the obvious measurement quality problems, around 1980 scientific attention shifted from category-based scaling to magnitude estimation (Krebs & Schmidt 1993, Lodge & Tursky 1979, Orth 1982, Wegener 1982, 1983). Category rating and magnitude estimation differ fundamentally, as they are based on different cognitive operations, that is, thinking in differences or in ratios (Bolanowski & Geischer 1991, Dunn-Rankin 1983, Montgomery 1975, Wegener 1983). The application of magnitude scaling to social science research has been induced by Stevens (1975) and the possibility of "cross-modality matching" (see Cross 1982), i.e. using two out of various available scaling modalities (such as numbers, line length, hand pressure, sound level).

Theoretical and empirical comparisons (e.g., Levine 1994, Lodge & Tursky 1979, Orth 1982, Purdy & Pavlovic 1992, Rohrmann 1985, Schaeffer & Bradburn 1989, Wegener 1983, Wills & Moore 1994) showed that magnitude scaling is principally superior in terms of measurement theory and data quality but is more demanding (both for the respondents and the researcher), requires more time and tends to be less liked by the majority of respondents. In fact, magnitude approaches have not become mainstream scaling methodology; conventional category-based rating scales are still dominating, certainly in applied and field research with non-academic populations, as textbooks for research methods and especially questionnaire design illustrate (e.g., Aiken 1997, Babbie 1989, Czaja & Blair 2005, Dillman 2000, Foddy 1992, Kerlinger & Lee 2000, Krosnick 1999, Krosnick & Fabrigar 1998, Miller 1991, Oppenheim 1992, Sapsford 2007, Schuman 1996, Vaus 1991). Thus the need for methodologically satisfactory category-based rating scales has to be acknowledged.

Obviously verbal labelling provides many advantages, such as ease-of-explanation and familiarity (in fact most people prefer verbal responses when replying to rating tasks, Moxey & Sanford 2000). It also facilitates capturing normative judgments. This is offset (as outlined above) by inferior measurement quality; that cultural factors might confound the data is a further disadvantage (cf. e.g. Auer et al. 2000, Chen et al. 1995, Reid 1995, Schaefer 1991, Tourangeau & Rasinski 1988, Van de Vijver 2001, Van de Vijver & Leung 1997, Weinfurt &

Moghaddam 2001). Furthermore, cross-national comparability of ratings is difficult (cf. e.g., Harzig 2005), as the equivalence of expressions in different languages is usually not known. Only for one topic, the intensity of noise annoyance, has this complex matter been researched systematically (cf. Fields et al. 2001, Felscher-Suhr et al. 1998, Guski et al. 1998; see also Rohrmann 1998). Pertinent knowledge is vital for cross-cultural survey research though.

A further issue is whether the interpretation of qualifiers is stable over time. Research into this matter is extremely rare (Rohrmann 1978, Simpson 1963).

In sum, it is essential to design verbalized scales very carefully if equi-distant and unambiguous instruments are to be achieved - if possible based on psychometric data for scale labels. However, only very few studies are available to provide such information.

Objectives of the project

The principal idea underlying this research project is: to design rating scales using verbal labels which reflect the cognitions of respondents and for which psycholinguistic and psychometric data are available. Therefore a series of studies was planned to clarify the measurement features of relevant verbal scale point labels and to develop methodologically sound response scales which are useful for both basic and applied research.

Research questions to be addressed include:

(1) Which are the best verbal labels for rating scales with 5 to 9 points in terms of equidistance, linguistic distinctiveness and comprehensibility?

(2) Is the modifying function of a VSPL influenced by the content and context of the scaling task at hand?

(3) To what extent is the perception of VSPLs homologous for people of different educational background?

(4) Do category scaling and magnitude estimation provide coherent information about VSPLs?

(5) Is it possible to create ratings scales in different languages which are mutually equivalent in terms of their VSPLs?

(6) Has the subjective interpretation of frequency and intensity expression shifted over time?

Topics (1) to (4) are investigated for VSPLs in English language. Topic (5) is aimed at German and Chinese language. Topic (6) is linked to the author's prior German scaling studies conducted in 1966 and 1976 (cf. Rohrmann 1967, 1978).

Research plan

Principal approach

If rating scales are to be constructed which approximate interval scale quality, it is essential to use equi-distant scale points. While numbers and/or layout features enhance perceived equidistance, words do not necessarily convey this. Consequently, VSPL are to be identified which have the 'right' position on the judgment scale to be constructed (depending on the number of points) and high linguistic distinctiveness (i.e., low variance in their perceived meaning). The principle is to calibrate the response scales.

To gain the necessary information, a combined lab and field study was designed, employing procedures of direct scaling (Anderson et al. 1983, McIver & Carmines 1993). The research plan involved to collect all verbal scale point labels (words or expressions) used or usable in rating scales; to identify principal dimensions of ratings - such as frequency, or intensity - and sort the VSPLs into these categories; and then to examine the quantitative meaning of sets of VSPLs. To increase cross-method validity, several psychometric procedures were chosen as quantification tools, based on either category scaling or magnitude estimation. Furthermore, context effects were to be controlled by using several linguistic 'frames' for the qualifiers under study.

The project was organized into five phases:

- Documentation of verbal scale point labels (VSPL) used in research
- Study <A> = Category scale rating of 100 VSPLs (expressions/words)
- Study <B> = Comparison of category and magnitude scaling outcomes
- Study <C> = Cross-national extensions (non-english data collection)
- Application of findings to scale construction for questionnaires.

Results from studies <A> and <B> are available and presented in this text, as well as implications for rating scale design. For study <C>, several experiments addressing research questions (5) and (6) have been conceptualized, and data collections (in Germany and Hong Kong) are in preparation; the results will be reported separately.

Selection of words/expressions

As a first step, words or expressions which have been used as VSPLs in rating scales and/or studied previously in psychometric research were searched and documented (restricted to English and German-speaking countries).

Qualifiers are used to grade the degree to which a particular attribute is given. There are three fundamental judgment dimensions:

- *Intensity [Q]*, e.g., not, a little, rather, very, extremely;
- *Frequency [F]*, e.g., never, sometimes, often, always;
- *Probability [P]*, e.g., unlikely, hardly, possibly, for sure.

They can be used in manifold combinations with substantive attributes, usually expressed as either verb phrases (e.g., I agree, I am happy, I use trams) or adjectives (e.g. satisfactory, annoyed).

Two further types of judgments are frequently used in social science research and therefore deserve attention:

- *Quality [Q],* e.g., bad, acceptable, satisfactory, excellent;
- *Agreement with statements [S]*, e.g., don't accept, agree, true for me.

All collected words/expressions were allocated to these 5 categories, and further ones were created by combining single modifiers into combined ones, e.g., very often ('I'+'F'), not likely ('I'+'P'), rather good ('I'+'Q'), often true for me ('F'+'S').

The psycholinguistic status and understanding of these qualifiers (cf. Hoermann 1983) and suitability as VSPLs was pretested as follows: each word/expression was inserted into a set of test sentences (e.g., "I am {.....} worried about the risk of an accident"), and 3 raters assessed whether it is linguistically suitable or not.

For each dimension, about 20 items were then selected according to two criteria: suitability for constructing rating scales with 5 to 9 points, and comparability with previous research (including the applicants' studies in Germany, i.e., Rohrmann 1978, 1985, and an Australian study by Clarke 1992). They are listed in *Table 1.*

== Table 1 ==

Scaling tasks

In order to quantify the meaning of the VSPLs, the following scaling tasks where used*:*

<NW> *Category scaling ("numbers for words"):*

Each VSPL, presented on a card, had to be placed on a 11-point "equal appearing interval scale" (Thurstone 1929) in which "0" was presented as lowest and "10" as highest level of the respective dimension/attribute.

<WN> *Category scaling ("words for numbers"):*

Respondents were presented with a set of VSPLs (printed on cards) and asked to choose their preferred verbal label for each level of a numerical five-point scale (presented as a scaling frame, numbered by -2/-1/0/+1/+2), i.e., the had to identify one best-suitable word/expression for each of the five scale points.

<MN/ML> *Magnitude estimation:*

The 'strength' of each VSPL was to be expressed in two magnitude modalities, numbers and lines (to be drawn on a sheet of paper), these being the best-established modes. In each dimension an item at the lower end of the range (e.g., seldom, little, unlikely) was used as baseline; then numbers or lines, respectively, were to be allocated which indicate the perceived ratio between each VSPL and that reference stimulus.

<FR> *Ratings of the familiarity of expressions:*

On a 0-to-10 scale, for each VSPL it was judged how common and familiar it is in everyday language.

Examples for some of these scaling tasks can be found in the appendix.

Furthermore, the VSPLs were presented in three different contexts:

- (N) Noise (e.g.: I am {.....} annoyed by loud aircrafts);

- (J) Job satisfaction (e.g.: I am {....} happy with my workplace);
- (C) 'pure', i.e., without context.

If necessary, different phrases were used for the 5 VSPL categories.

Experimental set-up and data collection

Because of the very small project budget, not all combinations of 5 VSPL types, 3 contexts and 4 scaling tasks could be realized, and only small sample sizes were feasible. In *Table2*, a summary is provided. The participants were recruited from psychology students and the general population; for each sub-group, the target N was 40.

== *Table 2* ==

The experiments were conducted in small groups. The instructions for the various tasks were read out by the experimenter but also presented in a scaling booklet, and participants recorded their responses in the appropriate sections. The sessions started with a 'warm-up' task to familiarize the participants with the unusual task of using scales to scale scale labels.

Propositions

The project was conceived as descriptive rather than hypothesis-testing research. However, the following propositions were stated, to be checked empirically:

- VSPLs at the ends of a continuum are perceived as less ambiguous than those in the middle range;
- the ordinal structure within a set of VSPLs is stable across contexts;
- for items which are prone to context effects, the impact is smaller for magnitude estimates than for category scaling results;
- the variance of ratings is lower for students than non-academic respondents;
- short and commonly used words are preferred as VSPLs.

It is obvious that pertinent results would be relevant for scale construction principles.

Results

The project produced a very large set of data; thus only a selection can be covered in this text. The results are presented in seven sections: sample description; VSPL data from category scaling; results from the magnitude scaling tasks; familiarity of words/expressions; preferred VSPLs for scale positions; effects of content/context; and differences between student and non-academic groups.

Data sets and sample description

Altogether N=229 respondents participated in the sub-studies conducted so far (cf. *Table 2*). For each experiment separate data set were created; these were then merged for task which were identical across sub-groups (e.g., the familiarity ratings).

The mean age of the participants is around 20 for the student and around 40 for the non-student groups; about 2/3 of the participants were female.

Mean scale positions: category scaling

The main results for the category scaling task "Numbers for words" (NW) are presented in *Table 3*. Mean scores and standard deviations are given for one of the three scaling contexts, i.e., noise, as well as results for merged context conditions.

== Tables 3-I, 3-F, 3-P, 3-Q, 3-S ==

The data show that the chosen VSPLs cover the whole range from very low to very high levels, as the mean scores in the 5 modalities range from 0.0 or 0.1 (e.g., "not at all", "never", "no chance", "fully disagree") to 9.9 or 10.0 (e.g., "completely", "always", "for sure", "outstanding", "fully agree").

For some words the quantitative scaling results deviate from qualitative anticipations. Examples include "rather" and "quite", which have been used on level four of 5-point-scales and were expected to score around 6.5 (i.e., placed in the middle between "medium" and "very") - however, here they were rated as 5.8 and 5.9. Another example: the 'quality' qualifier "poor" (rated 1.5) is almost as bad as "bad" (rated 1.0).

For most of the tested VSPLs the inter-individual variability is low (i.e., sd < 1.0). Even some very vague expressions, such as "so-so" or "not too bad" get reasonably definite scale positions. However, for some items people differ considerably in their allocation of quantitative equivalents, e.g., "quite a bit", "rather", "somewhat", "under some circumstances". This variation is higher for mid-range labels, as the meaning of extreme labels such as "not at all" or "always" has almost no ambiguousness. The graph in *Figure 1* illustrates the relationship between M and sd for the intensity labels.

== Figure 1 ==

Altogether the results indicate that most of the words and expressions under study are well understood as qualifiers of particular degrees of intensity, frequency, probability, quality and agreement.

Are the findings of this research in line with data from other studies (e.g. Jones & Thurstone 1955, Windschitl & Wells 1996)? This is difficult to assess, as the scaling approaches differ quite a bit *(sic)*; furthermore, many of the items in this study have never been scaled before. It seems though that the rank order of comparable items is reasonably similar.

It is tempting to check whether existing rating scales have equi-distant VSPLs. For example, using "rarely" *and* "seldom" (here scaled at 1.3 and 1.7) or "often" *and* "frequently" (here scaled as 6.6 and 7.4) in the same rating scale doesn't make much sense (cf. *Table 4*). Probably the most-often used rating scale in the social sciences is "strongly-disagree/disagree/neither-agree-nor-disagree/agree/strongly-agree"; these VSPLs were scored as 0.4, 1.6, 4.9, 8.2, 9.6 and are obviously not fulfilling the equidistance principle. (In fact, "mainly disagree" and "mainly agree" would be better VSPLs for levels 2 and 4 of this 5-

point scale).

The application of the scaling results to rating scale construction will be discussed in the final section.

Results from the magnitude scaling tasks

For the magnitude scaling data, several types of mean scores were computed, with either untreated or standardized individual scores (using 1.0 as reference value for all ratios) or the log of raw scores as input: (a) arithmetic means, (b) geometric means, and (c) the log of the arithmetic mean. Furthermore the CMM ('cross-modality matching') approach was applied, i.e., merged number/line responses were created, using geometric item means; these scores were then transformed onto a 0..100 scale.

The second block of columns in *Tables 3, 6 and 7* contains two of the magnitude scaling results: means and sd's for the 'number' response modality; and the GM for the merged scale scores. Only results for combined context conditions are given.

The results for the 'number' modality show the enormous range of ratios used by the respondents; these ranges are different for intensity, quality and agreement VSPLs. For example, "completely" is scaled as 80.8 times as strong as "not at all"; for quality, the highest item, "very good", gets 14.5, in comparison to 1.4 for bad; for agreement VSPLs, the extremes are 1.5 and 29.9 for "fully disagree" and "fully agree".

== Figure 2 ==

However, it seems questionable to take these data literally (*sensu,* "very good" is 10 times as good as "bad"), because many respondents expressed that they perceived this scaling task as unfamiliar, difficult and unnatural.

It is important to note though that the rank order of the items resulting from the various magnitude scalings is more or less the same as that for category scaling results; only VSPLs in the middle range (such as "fairly", "moderately") are likely to have inversions. *Figure 2* shows an example, i.e., category and magnitude results for intensity items. In fact, the relative position of main scale labels comes out quite similarly in both scaling approaches (the correlations are 0.98, 0.99 and 0.99 for intensity, quality and agreement).

Familiarity of words/expressions

Data on the perceived familiarity of the VSPLs (rated on a 0..10 scale) are presented in the last two columns of *Tables 3I to 3S*. All words/expressions are rated as at least moderately familiar (mean > 5.0). However, while all items were known, only few are seen as completely common (e.g., "not", "never", "always"). Most of the items rated as less common are either expressions composed of several words, such as "moderately often", "under some circumstances", "mostly dissatisfied", "fairly true for me"; or infrequently-used adverbial forms of adjectives, such as "considerably", "moderately", "fairly (even though all these are linguistically correct words).

Interestingly, the standard deviations are considerably higher than those for task NW, assessing the scale position of VSPLs. It seems that people are quite certain about the meaning of these words as qualifiers, even if they don't perceive them as 'household' expressions.

Allocation of labels to scale positions

The results for the "words for numbers" task (WN) can be found in the third block of results in *Tables 3-I to 3-S*. The tables show for each VSPL which percentage of respondents proposed it for a particular scale position. This task - to ask people to create verbalized 5-point rating scales - provides unique results as it has not yet been used in pertinent research. The data demonstrate clear preferences for most allocations (up to 90%, e.g., "never" and "always" for levels '1' and '5' of a frequency scale). It is also obvious that respondents generally prefer extreme labels at the end (e.g., "not at all" rather than "not" at level '1' and "extremely" rather than "very" for 'intensity' level '5'). As can be expected, the choices for levels '2' and '4' are more diverse than those for mid- and end-points. Generally, short labels are preferred.

Effects of content/context differences

Whether the VSPLs were presented context-free or embedded into a particular context (noise, job satisfaction) had very little influence on the “NW” scaling results - most of the respective differences are small and statistically insignificant. In *Tables 3-I to 3-S*, the results for one context are listed (cf. the column "noise" beside "all"). For the magnitude estimation results (restricted to two contexts) a similar pattern evolved. It seems that the quantitative meaning of verbal qualifiers is stable and on the whole independent of the judgmental dimension for which they are used.

A further type of context effects, the influence of the range of items presented to respondents, was not explicitly tested in this project. There is some informal evidence available though: Various pretests were conducted with smaller VSPL sets, and for "quality" and "agreement" items, the magnitude scaling tasks were run for a sub-sets of items only; the respective results seem to indicate that the position of a VSPL on the min-max continuum is not much affected and at least rank order information is stable.

Differences between student vs. population samples

It could be that students and non-students differ in their understanding of VSPLs, induced by, e.g., effects of age, education, and language preferences in sub-cultures. Given the small 'general population' sample (this part of the project is not yet complete), only exploratory analyses could be run. The data show no substantial and systematic differences for the main

VSPLs, i.e., those which are frequently used in rating scales; however the variance of judgments tends to be higher. Altogether the results seem to indicate that the understanding of the VSPLs scaled in this project is consistent and not specific for a student population.

Considerations

Validity constraints

Obviously the external validity of these findings must be restricted, as small non-random samples were employed, and the non-student groups are certainly too small. Furthermore, not all variations of context conditions could be realized. On the other hand, the results are remarkably consistent across sub-samples and converge reasonably well with the (few) comparable studies, so they can be seen as valid, at least for the context of the 100 VSPLs studied in this project.

Regarding internal validity, some participants 'struggled' to understand the instructions, especially for the magnitude scaling tasks, and the explanation of the familiarity task may have been phrased too indistinct; both is likely to have increased unintended response dispersion.

Finally, there are epistemological issues to be considered. From a cognitive psychology or psycholinguistic perspective one may question whether a 'universal' (context-free and timeless) meaning of the words/expressions examined here can be measured and utilized for the construction of equi-distant scales, in spite of the many contexts in which language is used and develops over time. Yet the author's earlier studies (in Germany, 1966, and repeated a decade later, cf. Rohrmann 1976) encouraged a view that people have a good idea of the relative position and 'strength' of a word meant to express a certain level of intensity or probability and so on, and that these cognitions on average didn't change much over 10 years.

To conclude, of course the results have to be interpreted with care; however, they offer a rich potential for informed choices when designing scaling instruments.

Implications for designing rating scales

The outcomes of this research can be utilized for the systematic construction of scales measuring psychological variables and approximating interval scale level. Main considerations for choosing a word/expression for a scale point level are:

(1) appropriate position on the dimension to be measured;
(2) low ambiguity (i.e., low standard deviation in the scaling results);
(3) linguistic compatibility with the other VSPLs chosen for designing a scale;
(4) sufficient familiarity of the expression;
(5) reasonable likelihood of utilization when used in substantive research.

The scale at whole needs to be linguistically coherent and easy to communicate to research participants.

As the results (cf. *Tables 3)* show, for both 5-point and 7-point scales fitting words/ expressions can be found. Possible solutions for a 5-point scale include:

Frequency: "never/seldom/sometimes/often/always".

Intensity: "not/a-little/moderately/quite-a-bit/very".

Probability: "certainly-not/unlikely/about-50:50/likely/for-sure".

Quality: "bad/inadequate/fair/good/excellent".

Agreement: "fully-disagree/mainly-disagree/neutral/mainly-agree/fully-agree".

However, suitable words are not available for all tasks (e.g., there seems to be no good word for level 2 of a 5-point quality scale). Also, for several positions there are equally good alternatives available (cf. e.g., "a-little" and "slightly"; "fair" and "medium" and so on). Therefore in a small add-on study (not reported here) a dozen psychologists were presented with several alternatives of verbalized 5-point scales and asked for their appraisal; the responses were considered in the suggestions outlined above.

A difficult decision in designing scales is how extreme an endpoint to choose. In principal, the target values for items calibrated on a 0--10 scale would be either 0/2.5/5/7.5/10 or 1/3/5/7/9. In the "words-for-numbers" task, participants tended to propose extreme labels; in the case of an intensity scale, this would lead to "not-at-all/slightly/moderately/ considerably/extremely". There is a risk though: extreme endpoints may not be used very often (e.g., in questions such "how satisfied are you with …", "how angry are you about …" etc), by that effectively reducing a 5-point scale to a 3-point one. Pretests can help to decide whether it is better to avoid the top-end VSPL.

In addition to the labeling issue, the use of further scale level indicators is to be decided. The recommended format is multi-modal, i.e., the scale points should be depicted by a combination of numbers, words perceived as equidistant, and graphical means, in order to enhance both psychometric quality and user-friendliness.

For a multi-modal scale design approach, non-verbal scale point labels can be integrated. Examples for a 5-point scale include:

Numerical: 1/2/3/4/5 or -2/-1/0/+1/+2 or --/-/0/+/++;

Graphical means: equidistant frames, or scaled lines, or !/!!/!!/!!!/!!!!, etc.

Examples are shown in *Figure 3.*

== Figure 3 ==

The layout needs to be adapted to the questionnaire mode, e.g., in printed/mailed questionnaires respondents will be asked to circle their chosen response; in personal/face-to-face or telephone interviews they may be asked to verbally indicate the chosen scale point (in this case, numbers 1-to-5 are the easiest mode); in web-based surveys participants need to tick a box. Of course any newly constructed rating scale should be pretested for comprehensibility and acceptability with relevant target groups.

Project continuation

As outlined in the description of the project design, sub-study <C> will expand and conclude this research.

Research question (5) "*Is it possible to create ratings scales in different languages which are mutually equivalent in terms of their VSPLs?"* is planned to be addressed by investigating verbal qualifiers in two languages, German and Chinese. The intended experiments are designed for bi-lingual respondents. Currently, data collections in Hong Kong (several experiments) are completed; the results seem to be of considerable significance, and a pertinent publication is in preparation; The experiments planned for Germany are yet to be realized.

Research question (6) *"Has the subjective interpretation of frequency and intensity expression shifted over time?"* will be addressed in a replication of the author's prior German scaling studies conducted in 1966 and 1976 (cf. Rohrmann 1967, 1978). However, this experiment will be focussed on VSPLs which have been regularly utilized in rating scales (in preparation for 2007 or 2008).

The findings will help to conduct cross-national comparisons much more carefully. A standard approach in such research is to compare the percentages of respondents who replied with "very" or equivalent expressions to questions of interest (e.g., to identify the degree of noise annoyance or fear of crime or residential satisfaction in a community). However, comparisons can only be valid if the quantitative meaning of the utilized response scale and especially the top-end item - e.g, "very", "sehr", "tres", is sufficiently similar.

Directions for further research

To widen the validity scope, further research is indispensable. Firstly, whether the interpretation of qualifiers is consistent across different levels of age and education needs to be investigated with much larger samples. Secondly, within multi-cultural societies it is an issue whether findings for natural English speakers are valid for people with English as second language. Thirdly, different national types of English could be compared, such as, English, American and Australian English.

Such research would enable researchers to identify words and phrases which have a 'cross-culturally stable' qualifier effect. If such qualifiers exist, psychometrically valid response scales for surveys and experiments can be designed which can be employed across the whole population of a country.

References

Aiken, L. R. (1997). *Questionnaires and inventories: Surveying opinions and assessing personality*. Chichester: Wiley.

Anderson, A. B., Basilevski, A., & Hum, D. P. J. (1983). Measurement: Theory and techniques. In P. H. Rossi, J. D. Wright, & A. B. Anderson (Eds.), *Handbook of survey research* (pp. 231-287). New York: Academic Press.

Andrews, F. M. (1984). Construct validity and rerror components of survey measures: a structural modelling approach. *Public Opinion Quarterly, 48*, 409-449.

Auer, S., Hampel, H., Moeller, H.-J., & Reisberg, B. (2000). Translations of measurements and scales: Opportunities and diversities. *International Psychogeriatrics, 12*, 391-394.

Babbie, E. (1989). *The practice of social research*. Belmont, CA: Wadsworth.

Bolanowski, S. J., & Geischer, G. A. (Eds.). (1991). *Ratio scaling of psychological magnitude: In honor of the memory of S.S. Stevens*. Hillsdale New Jersey: Lawrence Erlbaum.

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

Budescu, D. V., & Wallsten, T. S. (1994). Processing linguistic probabilities: General principles and empirical evidence. In J. R. Busemeyer, R. Hastie, & D. L. Medin (Eds.), *Decision making from the perspective of cognitive psychology*. New York: Academic Press.

Chen, C., Lee, S., & Stevenson, H. W. (1995). Response style and cross-cultural comparisons of rating scales among East Asian and North American students. *Psychological Science, 6*, 170-175.

Christian, L. M., & Dillman, D. A. (2004). The influence of graphical and symbolic language manipulations on response to self-administered questions. *Public Opinion Quarterly, 68*, 57-79.

Clark, D. A. (1990). Verbal uncertainty expressions: A critical review of two decades of research. *Current Psychology: Research Reviews*, 203-235.

Clarke, V. A., Ruffin, C. L., Hill, D. J., & Beamen, A. L. (1992). Ratings of orally presented verbal expressions of probability by a heterogeneous sample. *Journal of Applied Social Psychology, 22*, 638-657.

Cliff, N. (1959). Adverbs as multipliers. *Psychological Review, 66*, 27-44.

Cliff, N. (1972). Adverbs multiply adjectives. In J. M. Tanur (Ed.), *Statistics: A guide to the unknown* (pp. 176-184).

Cox, E. P. (1980). The optimal number of response alternatives for a scale: a review. *Journal of Marketing Research, 17*, 402-422.

Cross, D. V. (1982). On judgments of magnitude. In B. Wegener (Ed.), *Social attitudes and psychophysical measurement* (pp. 73-88). Hillsdale: Erlbaum.

Czaja, R., & Blair, J. (2005). *Designing surveys: A guide to decisions and procedures*. Thousand Oaks, CA: Pine Forge Press.

Dawes, R. M., & Smith, T. (1985). Attitude and opinion measurement. In G. Lindzey & E. Aronson (Eds.), *Handbook of Social Psychology* (pp. 509-566). New York: Random House.

Diefenbach, Weinstein, & O'Reilly. (1993). Scale for assessing perceptions of health hazard susceptibility. *Health Education Research, 8*(2), 181-192.Dillman, D. A. (2000). *Mail and internet surveys: The tailored design method*. New York: Wiley.

Dillman, D. A. (2000). *Mail and internet surveys: The tailored design method*. New York: Wiley.

Dixon, P. N., Bobo, M., & Stevick, R. A. (1984). Response differences and preferences for all-category-defined and end-defined Likert formats. *Educational and Psychological Measurement, 44*, 61-66.

Dunn-Rankin, P. (1983). *Scaling methods*. Hillsdale/NJ: Erlbaum.

Felscher-Suhr, U., Guski, R., & Schuemer, R. (1998). Some results of an international scaling study and their impications for noise research. In N. Carter & R. F. S. Job (Eds.), *7th International Congress on Noise as a Public Health Problem* (pp. 733-737). Sydney: Noise Effects '98.

Fields, J. M., De Jong, R. G., Gjestland, T., Flindell, I. H., Job, R. F. S., Kurra, S., Lercher, P., Vallet, M., Yano, T., Guski, R., & Felscher-Suhr, U. (2001). Standardized general-purpose noise reaction questions for community noise surveys: Research and recommendation. *Journal of Sound and Vibration, 242*, 641-679.

Foddy, W. (1992). *Constructing questions for interviews and questionnaires: theory and practice in social research*. Cambridge: Cambridge University Press.

French-Lazovik, G., & Gibson, C. L. (1984). Effects of verbally labeled anchor points on the distributional parameters of rating measures. *Applied Psychological Measurement, 8*, 49-57.

Freyd, M. (1923). The graphic rating scale. *Journal of Educational Psychology, 13*, 51-57.

Guski, R., Felscher-Suhr, U., & Schuemer, R. (1998). *Entwicklung einer international vergleichbaren verbalen Belaestigungsskala.* Paper presented at the DAGA Conference, Zuerich.

Haertel, C. E. J. (1993). Rating format research revisited: Format effectiveness and acceptability depend on rater characteristics. *Journal of Applied Psychology, 78*, 212-217.

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

Hartley, J., Trueman, M., & Rodgers, A. (1984). The effects of verbal and numerical quantifiers on questionnaire responses. *Applied Ergonomics, 15*, 149-155.

Harzing, A. W. (2005). Does the use of english-language questionnaires in cross-national research obscure national differences. *International Journal of Cross Cultural Management, 5*, 213-224.

Hippler, H.-J., Schwarz, N., Noelle-Neumann, E., Knauper, B., & Clark, L. (1991). Der Einfluss numerischer Werte auf die Bedeutung verbaler Skalenendpunkte. *ZUMA-Nachrichten, 28*, 54-65.

Hoermann, H. (1983). The calculating listener, or how many are einige, mehrere and ein paar (some, several, and a few). In R. Bauerle, C. Schwarze, & A. von Stechow (Eds.), *Meaning, use and interpretation of language*. Berlin: De Gruyter.

Jones, L., & Thurstone, L. L. (1955). The psychophysics of semantics: An experimental investigation. *The Journal of Applied Psychology, 39*, 31-36.

Kerlinger, F. N., & Lee, H. B. (2000). *Foundations of behavioural research*. (Fourth Edition ed.). Fort Worth: Harcourt College.

Klockars, A. J., & Yamagishi, M. (1988). The influence of labels and positions in rating scales. *Journal of Educational Measurement, 25*, 85-96.

Krebs, D., & Schmidt, P. (1993). *New directions in attitude measurement*. New York: Gruyter.

Krosnick, J. A. (1999). Survey Research. *Annual Review of Psychology, 50*, 537-567.

Krosnick, J. A., & Fabrigar, L. R. (1997). Designing rating scales for effective measurement in surveys. In L. Lyberg, P. Biemer, M. Collins, E. deLeeuw, C. Dippo, N. Schwarz, & D. Trewin (Eds.), *Survey Measurement and Process Quality* (pp. 141-164). New York: Wiley.

Krosnick, J. A., & Fabrigar, L. R. (1998). *Designing good questionnaires: Insights from psychology*. New York: Oxford University Press.

LeBlanc, A., Chang-Jin, Y., Simpson, C. S., Stamou, L., & McCrary, J. (1998). Pictorial versus verbal rating scales in music preference measurement. *Journal of Research in Music Education, 46*, 425-435.

Lehto, M. R., House, T., & Papastavrou, J. D. (2000). Interpretation of fuzzy qualifiers by chemical workers. *International Journal of Cognitive Ergonomics, 4*, 73-86.

Levine, N. (1981). The development of an annoyance scale for community noise assessment. *Journal of Sound and Vibration, 74*, 265-279.

Levine, T. R. (1994). Do individuals make interval/ratio level responses to magnitude scaled items? *Journal of Social Behavior and Personality*, 377-386.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 22*, 1-55.

Lodge, M., & Tursky, B. (1979). Comparison between category and magnitude scaling of political opinion employing SRC/CPS items. *American Political Rewview, 73*, 50-66.

McIver, J. P., & Carmines, E. G. (1993). Unidimensional scaling. In M. S. Lewis-Beck (Ed.), *Basic measurement* . Beverly Hills: Sage.

Miller, D. C. (1991). *Handbook of research design and social measurement*. London: Sage.

Montgomery, H. (1975). Direct scaling: Category scales, magnitude scales and their relation. *Goeteborg Psychological Reports*.

Moxey, L. M., & Sanford, A. J. (1991). Context effects and the communicative functions of quantifiers: Implications for their use in attitude research. In N. Schwarz & S. Sudman (Eds.), *Context effects in social and psychological research* . New York: Springer.

Moxey, L. M., & Sanford, A. J. (1993). *Communicating quantities: A psychological perspective*. Hillsdale: Erlbaum.

Moxey, L. M., & Sanford, A. J. (2000). Communicating quantities: A review of psycholinguistic evidence of how expressions determine perspectives. *Applied Cognitive Psychology, 14,* 237-255.

Nakao, M. A., & Prytulak, L. S. (1983). Numbers are better than words. *The American Journal of Medicine, 74*, 1061-1065.

Newstead, S. E., & Collis, J. M. (1987). Context and the interpretation of quantifiers of frequency. *Ergonomics, 30*, 1447-1462.

O'Muircheartaigh, C. A., Gaskell, G. D., & Wright, D. B. (1993). Intensifiers in behavioral frequency questions. *Public Opinion Quarterly, 57*, 552-565.

Oppenheim, A. N. (1992). *Questionnaire design, interviewing and attitude measurement*. New York: Basic Books.

Orth, B. A. (1982). A theoretical and empirical study properties of magitude-estimation and category-rating scales. In B. Wegener (Ed.), *Social attitudes and psychophysical measurement* (pp. #-#). Hillsdale: Erlbaum.

Parducci, A. (1983). Perceptual and judgmental relativity. In V. Sarris & A. Parducci (Eds.), *Perspectives in psychological experimentation: Towards the year 2000* (pp. 135-148). London: Erlbaum.

Pepper, S., & Prytulak, L. S. (1974). Sometimes frequently means seldom: Context effects in the interpretation of quantitative expressions. *Journal of Research in Personality*, 95-101.

Poulton, E. C. (1989). *Bias in quantifying judgements*. New Jersey: Erlbaum.

Presser, S., & Blair, J. (1994). Do different methods produce different results? In P. V. Marsden (Ed.), *Sociological Methodology* (pp. 73-104). Cambridge: Blackwell.

Purdy, S. C., & Pavlovic, C. V. (1992). Reliability, sensitivity and validity of magnitude estimation, category scaling and paired-comparison judgments of speech intelligibility by older listeners. *Audiology, 31*, 254-271.

Reagan, R. T., Mosteller, F., & Youtz, C. (1989). Quantitative meanings of verbal probability expressions. *Journal of Applied Psychology, 74*, 433-442.

Reid, R. (1995). Assessment of ADHD with culturally different groups: The use of behavioral rating scales. *School Psychology Review, 24*, 537-560.

Rohrmann, B. (1967). *Zwischenbericht ueber die Hamburger Voruntersuchung zum DFG-Projekt Fluglaermforschung (Sozialpsychologische Sektion)*. Mannheim: Universitaet Mannheim.

Rohrmann, B. (1978). Empirische Studien zur Entwicklung von Antwortskalen fuer die sozialwissenschaftliche Forschung. *Zeitschrift fuer Sozialpsychologie*, 222-245.

Rohrmann, B. (1985). Categorical scaling versus magnitude scaling - A practical comparison. In E. E. Roskamp (Ed.), *Measurement and personality assessment* (pp. 155-164). Amsterdam: North-Holland.

Rohrmann, B. (1998). *The use of verbal scale point labels in annoyance scales.* In C. Norman & S. R. F. Job (Eds.), *7th International Congress on Noise as a Public Health Problem* (Vol. 2, pp. 523-527). Sydney: Noise Effects Pty Ltd.

Sapsford, R. (2007). *Survey research*. London: Sage.

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

Schaeffer, N. C., & Bradburn, N. M. (1989). Respondent behavior in magnitude estimation. *Journal of the American Statistical Association, 84*, 402-413.

Schuman, H. (1996). *Questions and answers in attitude surveys: Experiments on question form, wording, and context*. Newbury Park: Sage.

Schwarz, N., Knuper, B., Hippler, H.-J., Noelle-Neumann, E., & Clark, L. (1993). Rating scales: numeric values may change the meaning of scale labels. *Public Opinion Quarterly, 55*, 570-582.

Simpson, R. H. (1963). Stability in meanings for quantitative terms: A comparison over 20 years. *Quarterly Journal of Speech, 49*, 146-151.

Spector, P. E. (1976). Choosing response categories for summated rating scales. *Journal of Applied Psychology, 61*, 347-375.

Spector, P. E. (1993). *Summated rating scale construction: An introduction*. Florida: Sage.

Stevens, S. S. (1975). *Psychophysics: Introduction to its perceptual, neaural, and social prospects*. New York: Wiley.

Theil, M. (2002). The role of translations of verbal into numerical probability expressions in risk management: a meta analysis. *Journal of Risk Research, 5*, 177-186.

Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology, 33*, 529-554.

Tourangeau, R., & Rasinski, K. A. (1988). Cognitive processes underlying context effects in attitude measurement. *Psychological Bulletin, 103*, 299-314.

Traenkle, X. (1987). Auswirkungen der Gestaltung der Antwortskala auf quantitative Urteile. *Zeitschrift fuer Sozialpsychologie*, 88-99.

Van de Vijver, A., & Leung, K. (1997). *Methods and data analysis for cross-cultural research*. Thousand Oaks: Sage.

Van de Vijver, F. (2001). The evolution of cross-cultural research methods. In D. Matsumoto (Ed.), *The handbook of culture and psychology* (pp. 77-97). New York: Oxford University Press.

Vaus, D. A. d. (1991). *Surveys in social research*. (3 ed.). London: Unwin.

Wegener, B. (1983). Category-rating and magnitude estimation scaling techniques: An empirical comparison. *Social methods and Research, 12*, 31-75.

Wegener, B., Faulbaum, F., & Maag, G. (1982). Die Wirkung von Antwortvorgaben bei Kategorialskalen. *ZUMA-Nachrichten, 10*, 3-20.

Weinfurt, K. P., & Moghaddam, F. M. (2001). Culture and social distance: A case study of methodological cautions. *Journal of Social Psychology, 141*, 101-110.

Wildt, A. R. (1978). Determinants of scale response: Label versus position. *Journal of Marketing Research, 15*, 261-267.

Wills, C. E., & Moore, C. F. (1994). A controversy in scaling of subjective states: Magnitude estimation versus category rating methods. *Research in Nursing and Health, 17*, 231-237.

Windschitl, P. D., & Wells, G. L. (1996). Measuring psychological uncertainty: Verbal versus numeric methods. *Journal of Experimental Psychology: Applied, 2*, 343-364.

Wright, D. B., Gaskell, G. D., & O'Muircheartaigh, C. A. (1994). How much is 'quite a bit'? Mapping between numerical values and vague quantifiers. *Applied Cognitive Psychology*, 479-496.

Zimmer, A. C. (1988). A common framework for colloquial quantifiers and probablility terms. In T. Zetenyi (Ed.), *Fuzzy sets in psychology* (pp. 73-89). Amsterdam: North Holland.

Rating scale design using verbal qualifiers: TABLES & FIGURES

Table 1*:* List of All Items Used in Project VQS

<F> FREQUENCY

always
fairly often
frequently
mostly
never
occasionally
often
moderately often
rarely
seldom
sometimes
very often

<I> INTENSITY

a little
average
completely
considerably
extremely
fairly
*fully
hardly
highly
* in-between
*mainly
medium
moderately
not
not at all
partly
quite
*quite a bit
rather
slightly
somewhat
very
very much

<P> PROBABILITY

about 50-50
a very. good chance
certainly
certainly not
for sure
likely
no chance at all
perhaps
possibly
probably
probably not
quite likely
unlikely
under some circumstances
under most circumstances.
with certainty

<Q> QUALITY

adequate
*average
bad
dissatisfied
excellent
fair
good
inadequate
medium
mostly dissatisfied
mostly satisfied
not too bad
outstanding
poor
satisfactory
satisfied
so so
unsatisfactory
very good
very satisfied
very dissatisfied

<S> (DIS-) AGREEMENT WITH STATEMENTS

agree
disagree
*don’t agree
fairly true for me
*fully agree
*fully disagree
*half-half
in-between
*mainly agree
*mainly disagree
mostly true for me
neither agree/disag
neutral
not true for me
right
somewhat agree
somewhat disagree
s/what true for me
strongly agree
strongly disagree
true for me
undecided

*Note: Items labelled with * were not used in all sub-studies.*

Table 2: Data Collection - Studies <A> and <B>

Subgroup	*Scaling tasks*	*Dimensions*	*Condition*	*Respondents*
<A-C>	Category scaling: WN, NW, FR	F I P Q S	Context-free	44 Students
<A-N>	Category scaling: WN, NW, FR	F I P Q S	Noise context	39 Students
<A-J>	Category scaling: WN, NW, FR	F I P Q S	Job satisf. context	37 Students
<A-P>	Category scaling: WN, NW, FR	F I P Q S	Mixed contexts	44 Gen. population
<B-C>	Magnitude scaling: MN, ML; Cat.: NW	I Q S	Context-free	38 Students
<B-N>	Magnitude scaling: MN, ML; Cat.: NW	I Q S	Noise context	38 Students

Notes:
"NW" = "numbers for words", "WN" = "words for numbers"; "MN" = magnitude scaling in number modality, "ML" = magniude scaling in lines modality, "FR" = ratings of the familiarity of expressions.
Further sections included in each experiment were: A scaling test exercise; respondent's viewpoints regarding category and magnitude scaling; and demographic questions.
For the magnitude scaling tasks in study <B>, a reduced set of VSPLs was used.
Sub-studies <B-J> (Job context) and <B-P> (mixed context, general population) were postponed.

Table 3-I*:* Main Results for "Intensity" Qualifiers

Scaling task	CATEGORY				MAGNITUDE			PREFERED LABEL					FAMILIA-	
	(0...10 scale)				<NM>		<X_{nl}>	for levels (%)					RITY	
Context:	all		noise		all			all					all	
	M	*sd*	*M*	*sd*	*M*	*sd*	*GM*	*1*	*2*	*3*	*4*	*5*	*M*	*sd*
Verbal label														
a little	2.5	1.2	2.3	1.2	10.5	17.5	16		13				7.0	2.5
average	4.8	0.8	4.7	0.7	--	--				28			7.8	2.0
completely	9.8	0.6	9.9	0.5	80.8	161.4	97					40	8.2	1.9
considerably	7.6	1.1	7.6	1.0	57.1	128.7	65				21		6.3	1.9
extremely	9.6	0.5	9.7	0.5	76.3	145.3	96					47	8.3	1.6
fairly	5.3	1.3	5.2	1.5	46.0	112.7	45						6.5	2.1
fully	9.4	1.1	9.4	1.1	77.5	161.0	87						--	--
hardly	1.5	0.8	1.5	0.9	8.8	16.7	10		18				7.1	2.1
highly	8.6	0.7	8.7	0.7	67.8	130.5	81						7.4	2.0
in-between	4.8	0.8	4.7	0.6										
mainly	6.8	1.1	--	--	58.1	128.6	59				18		7.4	1.7
medium	4.9	0.8	4.8	0.8	--	--				25			7.2	2.2
moderately	5.0	1.1	5.0	1.4	43.5	112.5	43			37			6.3	1.9
not	0.4	0.6	0.3	0.5	2.3	3.5	03	17					9.0	1.6
not at all	0.0	0.2	0.0	0.0	1.0	0.0	02	70					9.2	1.3
partly	3.5	1.3	3.6	1.4	21.4	48.6	25		14				6.8	1.9
quite	5.9	1.4	6.4	1.2	38.4	81.2	41						--	--
quite a bit	6.5	1.5	6.7	1.3	45.1	96.6	48						6.5	2.4
rather	5.8	1.5	6.0	1.5	45.9	113.4	44						5.6	2.3
slightly	2.5	1.3	2.5	1.3	11.6	17.2	18		27				6.4	2.1
somewhat	4.5	1.6	4.5	1.6	27.1	49.0	32						5.2	2.3
very	7.9	0.9	8.1	0.8	62.7	129.3	72				16		8.8	1.3
very much	8.7	0.8	8.7	0.6	70.7	145.3	84						8.6	1.5

Notes:

“Magnitude” data: GM= geometric mean; Nm= number modality, standardized raw scores; X_{nl}= scores based on merged number/lines responses.

“Prefered label” : respondents had to suggest one verbal label for each of the levels “1” to “5”.

"--": No data collected.

Table 3-F. Main Results for "Frequency" Qualifiers

Scaling task	CATEGORIAL (0...10 scale)				PREFERED LABEL for levels (%)					FAMILIA- RITY	
Context:	all		noise				all			all	
	M	*sd*	*M*	*sd*	*1*	*2*	*3*	*4*	*5*	*M*	*sd*
Verbal label											
always	10.0	0.2	10.0	0.2					90	9.4	1.0
fairly often	6.1	1.1	6.0	1.3						6.5	2.0
frequently	7.4	1.2	7.5	1.3				21		7.1	1.6
moderately often	5.7	1.2	5.8	1.3						4.6	2.2
mostly	8.0	1.3	7.8	1.3				18		7.6	1.7
never	0.0	0.1	0.0	0.2	92					9.5	1.0
occasionally	3.2	1.1	3.2	1.1		11	20			7.5	1.8
often	6.6	1.2	6.7	1.1				32		7.6	1.8
rarely	1.3	0.6	1.3	0.6		49				7.4	2.1
seldom	1.7	0.7	1.8	0.7		24				5.4	2.5
sometimes	3.6	1.0	3.7	1.1			50			8.4	1.8
very often	8.3	0.9	8.5	0.9				16		7.8	1.7

Notes:
“Prefered label”: respondents had to suggest one verbal label for each of the levels “1” to “5”.
"--": No data collected.

Table 5-P. Main Results for "Probability" Qualifiers

Scaling task	CATEGORIAL (0...10 scale)				PREFERED LABEL for levels (%)					FAMILIA-RITY	
Context:	all		noise				all			all	
	M	*sd*	*M*	*sd*	*1*	*2*	*3*	*4*	*5*	*M*	*sd*
Verbal label											
about 50 : 50	4.8	0.6	4.7	0.7			65			7.2	2.4
a very good chance	8.2	0.8	8.3	0.7						7.2	2.0
certainly	9.6	0.7	9.7	0.6					62	8.3	1.6
certainly not	0.2	0.4	0.1	0.3	47					8.2	1.7
for sure	9.8	0.6	9.9	0.3						7.8	2.1
likely	6.9	1.0	6.9	0.9				32		7.7	1.6
no chance at all	0.0	0.2	0.0	0.2	38					7.8	2.5
perhaps	4.5	1.4	4.8	1.5						7.2	1.9
possibly	5.0	1.4	4.9	1.5			10			7.4	1.9
probably	6.8	1.2	6.8	1.4				24		8.1	1.7
probably not	1.9	0.7	1.9	0.8		20				7.8	1.9
quite likely	7.4	1.1	7.4	1.0				18		6.6	2.1
unlikely	1.7	0.8	1.6	0.7		49				7.8	1.8
under most circumstances	7.5	1.5	8.2	0.8						5.9	2.7
under some circumstances	4.6	1.7	4.3	1.5						5.8	2.6
with certainty	9.8	0.5	9.9	0.4					18	6.5	2.6

Notes:
“Prefered label”: respondents had to suggest one verbal label for each of the levels “1” to “5”.
"--": No data collected.

Table 3-Q. Main Results for "Quality" Qualifiers

Scaling task	CATEGORIAL (0...10 scale)				MAGNITUDE <Nm>		$<X_{nl}>$	PREFERED LABEL for levels (%)					FAMILIA- RITY	
Context:	all		noise		all			all					all	
	M	*sd*	*M*	*sd*	*M*	*sd*	*GM*	*1*	*2*	*3*	*4*	*5*	*M*	*sd*
Verbal label														
adequate	5.6	1.2	6.0	1.2	--	--	--						6.3	1.9
average	4.9	0.5	--	--	7.9	8.7	38			48			--	--
bad	1.0	1.0	0.9	1.0	1.6	1.2	10	31					8.5	2.0
dissatisfied	1.9	1.1	1.5	1.0	--	--	--						7.1	2.2
excellent	9.7	0.6	9.7	0.4	--	--	88					45	9.3	1.0
fair	5.2	1.1	5.3	1.2	7.2	8.4	38		14	12			7.5	1.9
good	7.2	0.8	7.2	0.8	12.2	12.0	63				43		8.9	1.7
inadequate	1.9	1.2	2.0	1.2	2.2	1.6	15		11				6.7	2.0
medium	5.0	0.6	4.9	0.4	8.2	10.0	39			21			7.2	2.1
mostly dissatisfied	1.9	1.1	1.6	1.1	--	--	--						5.8	2.4
mostly satisfied	7.2	1.2	7.3	1.2	--	--	--						6.1	2.9
not too bad	4.6	1.3	4.5	1.1	--	--	--						7.3	2.2
outstanding	9.9	0.4	9.9	0.3	--	--	98					35	8.0	1.7
poor	1.5	1.1	1.4	1.1	2.0	1.8	12	24	26				8.2	2.0
satisfactory	5.9	1.2	6.4	1.2	7.4	7.1	40			14			7.9	1.8
satisfied	7.0	1.2	7.2	1.1	--	--	--						7.3	1.8
so so	4.5	0.7	4.7	0.6	--	--	--						5.9	2.6
unsatisfactory	1.8	1.3	2.1	0.9	3.4	5.0	15	16	13				7.7	1.9
very dissatisfied	0.5	0.7	0.3	0.5	--	--	--	32					6.4	2.7
very good	8.5	0.7	8.7	0.7	14.5	14.7	73				22		8.8	1.6
very satisfied	8.9	0.9	9.0	0.7	--	--	--					14	7.1	2.2

Notes:

"Magnitude" data: GM= geometric mean; Nm= number modality, standardized raw scores; X_{nl}= scores based on merged number/lines responses.

"Prefered label": respondents had to suggest one verbal label for each of the levels "1" to "5".

"--": No data collected.

Table 3-S. Main Results for "Agreement" Qualifiers for Statements

Scaling task	CATEGORIAL				MAGNITUDE			PREFERED LABEL					FAMILIA-	
	(0...10 scale)				<Nm>		$<X_{nl}>$	for levels (%)					RITY	
Context:	all		noise		all			all					all	
	M	*sd*	*M*	*sd*	*M*	*sd*	*GM*	*1*	*2*	*3*	*4*	*5*	*M*	*sd*
Verbal label														
agree	8.2	0.9	8.2	0.8	--	--	--				29	13	9.0	1.4
disagree	1.6	1.0	1.6	1.0	--	--	--	15	30				8.9	1.5
don’t agree	1.9	1.2	1.9	1.3	5.0	8.0	12						--	--
fairly true for me	6.6	0.9	6.5	1.0	--	--	--						5.8	2.6
fully agree	9.8	0.5	9.8	0.5	29.9	28.8	97					33	--	--
fully disagree	0.2	0.4	0.2	0.5	1.5	1.6	02	28					--	--
half-half	5.0	0.4	5.0	0.5	14.9	14.5	47						--	--
in-between	4.9	0.5	4.9	0.5	--	--	--						6.0	2.4
mainly agree	7.4	0.7	7.5	0.7	22.3	21.7	72				31		--	--
mainly disagree	2.4	0.9	2.3	1.0	7.4	6.6	20		29				--	--
mostly true f. me	7.7	1.0	7.8	1.1	--	--	--						5.7	2.7
n. agree n. disagr.	4.9	0.4	4.9	0.5	14.7	14.6	46			15			7.2	2.5
neutral	4.9	0.4	4.9	0.6	15.5	15.2	49			36			6.9	2.5
not true for me	1.2	1.0	1.1	1.0	--	--	--						6.3	2.8
right	8.6	1.1	8.3	1.4	--	--	--						8.1	2.3
somewhat agree	6.4	0.9	6.6	0.9	19.0	18.3	61				34		6.1	2.3
somewhat disagree	3.2	0.9	3.0	1.0	10.6	10.8	29		38				6.0	2.3
some. true for me	6.0	1.2	6.1	1.2	--	--	--						5.5	2.5
strongly agree	9.6	0.6	9.6	0.5	27.9	26.5	92						8.6	1.7
strongly disagree	0.4	0.6	0.3	0.5	2.2	2.6	04	66				68	8.7	1.5
true for me	8.4	1.2	8.4	1.1	--	--	--						6.7	2.6
undecided	4.8	0.6	4.9	0.6	--	--	--			22			7.7	2.4

Notes:

“Magnitude” data: GM= geometric mean; Nm= number modality, standardized raw scores; X_{nl}= scores based on merged number/lines responses.

“Prefered label”: respondents had to suggest one verbal label for each of the levels “1” to “5”.

"--": No data collected.

Figure 1. Mean Ratings and SD for Intensity VSPLs (Task "NW")

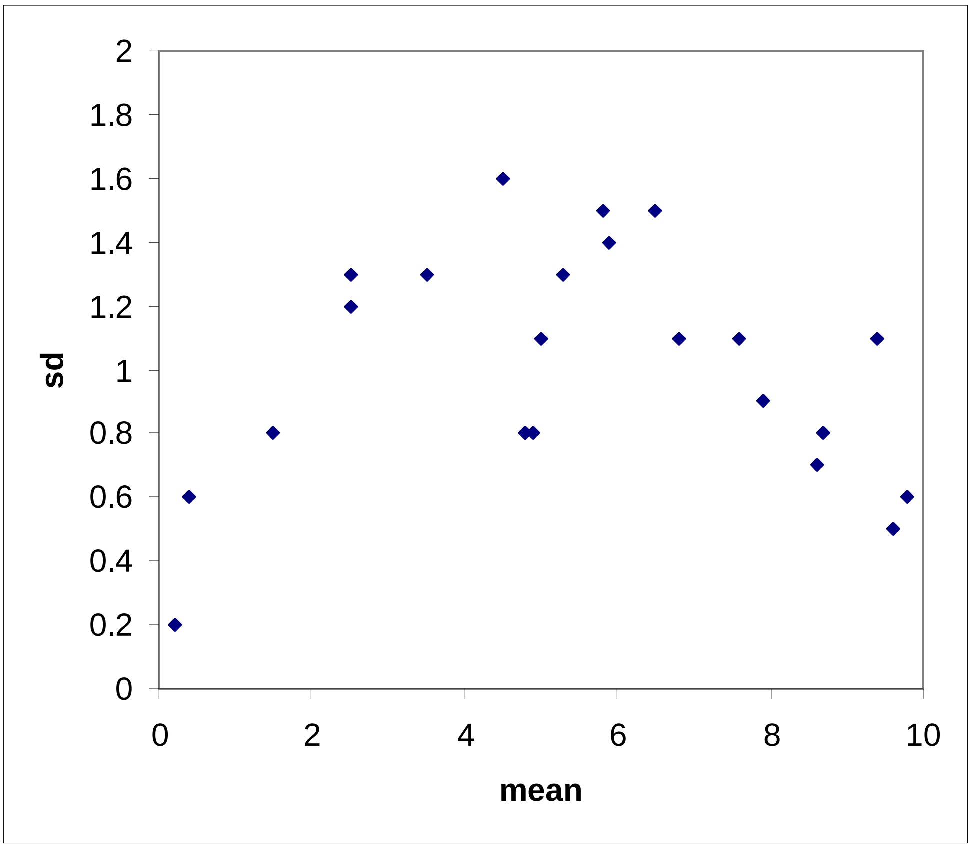

Figure 2. Relationship between Category and Magnitude Scaling Results, for Intensity VSPLs

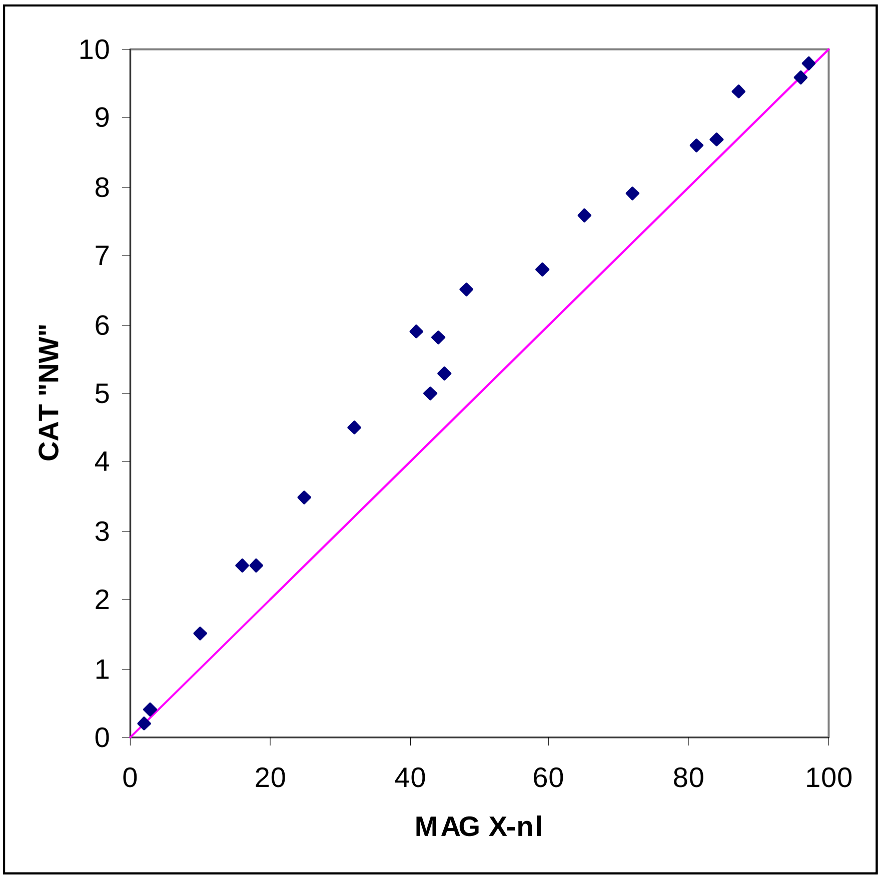

Figure 3. Examples of Multimodal Rating Scales

```
 --         -          :          +         ++
  |---------|----------|----------|----------|
bad   inadequate    average     good    excellent
```

```
  !        !!          !!!       !!!!       !!!!!
  |---------|----------|----------|----------|
never    seldom    sometimes    often      always
```

{end of doc}