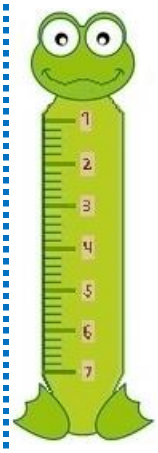
ROHRMANN, Bernd

Designing verbalized rating scales: Sociolinguistic concepts and psychometric findings from three cross-cultural projects

Reports, Roman Research Road, Melbourne 2015

CONTENT



CCC CONFUSION CODING CHECKLIST

my level of understanding is einsteinian
surely grasping whats goin' on
just a little uncertain - not to worry yet
maybe i'm confused? or not? who knows?
some fog is closing in on my thinking
thinx are about "klar wie klossbruehe"
everything is contradictory, don't even know why

This report summarizes the design, essential results and utilization of Project "VQS" ~ "Verbal qualifiers for rating scales: A cross-cultural psychometric study", which was running 1965-1967 and 1976-1978 in Germany and 1996 to 2010 in Australia and 2003 to 2004 in Hong Kong/China, based on an international perspective. Suppositions and practicality were the core interest. At the end, some metaphorical thoughts about the findings' utility and an outlook are presented.

Contact address:

Prof. Dr. B. Rohrmann *Website:* www.rohrmannresearch.net
Venture *"RomanResearchRoad"*, Melbourne/Australia *E-Mail:* mail@rohrmannresearch.net

Abstract

Designing verbalized rating scales: Sociolinguistic concepts and psychometric findings from three cross-cultural projects

Methodological issue: Surveys based on questionnaires are the dominant data collection method in psychology, sociology and other social sciences, and most use rating scales as response mode. Within category scaling, verbal labelling of rating scales has become the primary approach to enhancing usability. The labels are used as "qualifiers", either for the two 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 a restricted number of publications provide pertinent information.

Research approach: The principal idea underlying the presented research is, to create rating scales using verbal labels which reflect the cognitions of respondents and for which socio-linguistic and psychometric data are available. Therefore a series of studies was conducted to clarify the measurement features of relevant words or expressions and to develop methodologically sound response scales which are useful for both basic and applied research. This started in Germany, was later repeated and extended in a cross-national approach in Australia, and then in HongKong/ China (i.e., projects 1, 2, 3).

Data collection: A large number of 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. Furthermore, linguistic relations between different languages (here: English and Chinese) were quantified. This provided multiple information about the linkage between words and quantitative scale levels. The samples for the sub-studies in the three projects (N=122, N=229, N=300) were recruited from both university students and the general population, to widen the validity.

Outcomes and conclusions: The results provide a comprehensive body of quantitative information about common scale labels and enable the systematic design of response formats with using distinctive words or expressions. The recommended format is multi-modal to enhance both psychometric quality and user-friendliness. To widen the validity scope, on-going research is suggested, namely, to cover further languages (e.g., Arabic or Slavic ones), to inspect cultural distinctions within a country/language, and to explore stability over time of verbal scale point qualifiers..

Preface

This report is mostly based on previous publications or brochures. For "VQS-1": Rohrmann 1978 (in German language), for "VQS-2": Rohrmann 1998, 2007; for "VQS-3": Rohrmann, Au & Taylor 2008, 2011.

The presentation of methodology and outcomes of chapter 3 = VQS-1 in Germany, chapter 4 = VQS-2 in Australia and chapter 5 = VQS-3 in HongKong/China is somewhat overlapping. The reason is that each project description should be readable on its own.

<1> RATING SCALES IN SOCIAL SCIENCE RESEARCH: PRINCIPAL ISSUES

1.1 *Scales for judgments ~ "rating scales" in the social sciences*

For more than hundred years, various types of questionnaires are by far the most-used method of data collection in psychology, sociology and other social sciences, and about all of them use rating scales as primary response mode when exploring judgements, attitudes and behaviours. Countless articles have followed the seminal work of authors such as Freyd (1923), Thurstone (1928), and 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.

1.2 *Presenting and labelling scale points*

Commonly 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, Dawes & Smith 1985, Foddy 1992, Haertel 1993, Jensen et al. 2011, Krosnick & Fabrigar 1997, McIver & Carmines 1993, Myers & Winters 2002, Preston & Colman 2000, 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 widespread 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 this is: combining an adjective with adverbs). How scale points are denoted is very likely to affect response behavior (cf. e.g. Barilli et al. 2010, Christian & Dillman 2004, Dixon et al. 1984, French-Lazovik et al. 1984, Freyd 1923, Hartley et al. 1984, Hippler et al. 1991, 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 used to be 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, Barilli et al. 2010, 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 1967, 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 et al. 1982, Wegener 1983). Category rating and magnitude estimation differ fundamentally, as they are based on different cognitive operations, that is, thinking in differences or thinking 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, McColl & Fucci 2006, 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, 2011, Bryman 2012, Czaja & Blair 2005, Dillman 2007, Foddy 1992, Gerring 2011, Kerlinger & Lee 2000, Krosnick 1999, Krosnick & Fabrigar 1998, Lanier et al. 2014, Miller 1991, Montello & Sutton 2013, Oppenheim 1992, Robson 2011, 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 - comparing values and behaviors is very much a cultural phenomenon (Hofstede 2001). Only for one topic, the intensity of noise annoyance, has this complex matter been researched systematically (cf. Felscher-Suhr et al. 1998, Fields et al. 2001, Guski et al. 1998; Yano et al. took care of Japanese and further Asian languages; 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.

<2> PROJECT VQS: OUTLINE FOR THREE INVESTIGATIONS

2.1 *Original considerations*

The issue got vivid in 1965, while preparing a very large study in Germany on the impact of aircraft noise on residents, focussed on noise annoyance. It became obvious that the data collection should be based on interviews with people living around airports, rather than university students. Many of those people are not familiar with numerical scales, nor with finely graded viewpoints either. Their spontaneous responses were words, not numbers, e.g., "very", "sometimes", "a little", or just "yes" versus "no". Rating scales using such words did exist for a long time, yet their scale quality was uncertain - may be ordinal scales, or even less. Convincing rating scales hardly existed at that time.

Thus it was decided to run a methodological study to create suitable instruments. The principal idea underlying this research was: 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 explorations and experiments was planned to clarify the measurement features of relevant verbal scale point labels and to develop methodologically sound response scales - that is, instruments which are useful for both basic and applied research ventures.

Research questions to be addressed included:

(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?

In the following years it was discussed whether the set of studied words/expressions for rating scales was sufficient or needed extension, and whether the results can be taken as stable or not. Both issues were taken up in a complete repetition 10 years later. The crucial research question was

(4) Has the subjective interpretation of frequency and intensity expressions shifted over time?

These two investigations about "verbally qualified scales" became Project VQS-1; it is described in Chapter 3.

An overview list of the 6 dealt-with research questions is provided in *Box 2-1.*

2.2 *Continuation and extension*

The German language is spoken in several countries, yet the relevance of the findings is nevertheless restricted because the dominating research language is clearly English. This led to the decision to conduct a further project regarding verbally qualified scales - which was conducted from 1996 onwards in Australia. The essential research questions (1), (2) and (3)

were maintained. However, the selection of to-be-studied words and expressions had to start from scratch.

Box 2-1

Core tasks of Project VQS

(1)	Which are the best verbal labels for rating scales with 5 to 9 points in terms of equidistance, linguistic distinctiveness, comprehensibility?	VQS-1, -2, -3
(2)	Is the modifying function of a VSPL influenced by the content and context of the scaling task at hand?	VQS-1, -2, -3
(3)	To what extent is the perception of VSPLs homologous for people of different educational background?	VQS-1, -2, -3
(4)	Has the subjective interpretation of frequency and intensity expressions shifted over time?	VQS-1
(5)	Do category scaling and magnitude estimation provide coherent information about VSPLs?	VQS-2
(6)	Is it possible to create ratings scales in different languages which are mutually equivalent in terms of their VSPLs?	VQS-3

Furthermore, it appeared worthwhile to extend the psychometric procedure for measuring the intensity level of words. The pertinent research question was

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

Thereby words of interests could be cross-validated.

This investigations about verbally qualified scales in English became Project VQS-2; it is described in chapter 4.

A long debated question is how to compare rating scale results gained in countries with different languages. This is already problematic for pure numeric scales, because peoples habits in using extreme levels of a scale differ across cultures (cf. e.g., Chen et al. 1995, Harzing 2005). Obviously it's even more difficult for verbal scales.

Therefore a visiting professorship in HongKong/China in 2002/2003 was utilized to investigate verbally qualified scales in Chinese language.

Again research questions (1), (2) and (3) were pursued. Furthermore, a vital issue was added - research question (6):

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

To deal with this topic, data collections regarding English and Chinese items were necessary, plus linking procedures.

This investigations, the most complex one and first bilingual one about verbally qualified scales, became Project VQS-3; it is described in chapter 5.

Remark:

The description of the three projects VQS-1, VQS-2 and VQS-3 in the following chapter 3, chapter 4 and chapter 5 overlap to some degree, because they are based on shared concepts and methods.

This was accepted for this report, in order to make each project description readable on its own.

The overarching appraisals and conclusions are presented in the final chapter 6.

<3> CRAFTING VERBALLY QUALIFIED SCALES - VQS-1 - IN GERMANY

3.1 *The original problem space*

The whole issue - namely, which rating scales to use in a very large survey regarding the impacts of aircraft noise on the population - started when introductory interviews revealed that a lot of respondents were not familiar with demanding scales, that the at the university common zero-to-ten scales were too complex for them, that purely numerical scales were too abstract, and that quite a few people did not like formal scales at all, because they would prefer to reply in their own words to the interviewers' questions. Furthermore, in their mind many issues were obviously less widely graded beyond "yes" versus "no" - it seems that Miller's famous statement "The magical number seven, plus or minus two, some limits on our capacity for processing information" (1956) was somewhat optimistic, if you apply it to the scope of rating scales?

The sketched situation was around 1965, in Hamburg in Germany. The author was the researcher responsible for creating the necessary questionnaires for this large field study about noise effects, which included a lot of scaling formats. He had a double background, on one hand a high-level university education in experimental psychology, on the other hand empirical investigation methods in linguistics.

Yes, purely verbal response modes are liked by folks, yet as an experimental psychologist one will realize that a verbal frequency scale such as "never--seldom--often--frequently-always" is not equidistant (plus, there is no clear 'middle' category, and the meaning of "often" and "frequently" is pretty much the same). However, equidistance is a metric prerequisite if the data are to be analyzed by common statistical tools, which mostly request that the data have interval scale quality, not just ordinal scale level.

Numerical rating scales such as 1--2--3--4--5, or a 1-to-10 scale, are principally equidistant, yet as an empirical linguist one will wonder what, for example, "4" actually means to a respondent - does he/she think of "often", or "frequently", or even "mostly" or "very often", when rating the frequency of something as "4"? And how about social factors in using language?

At that time of Social Science Research, the tools and options of rating scales, as well the related troubles, were recurrently discussed, internationally (see chapter 2) and in Germany (cf. e.g., Clauss 1968, Friedrichs 1973, Hoermann 1967, Hofstaetter & Wendt 1966, Koenig 1965, Kristof 1966, Sixtl 1967) - yet convincing modes for practical survey research were missing.

Thus, connecting psychometric and socio-linguistic concepts, the decision was made to measure the link of words to scales, to investigate their familiarity as well, and then to create verbalized numeric rating scales based on empirical criteria - that is, instruments suitable for surveys in the general population, and for stern statistical analyses as well. This was a novel enterprize.

The venture was undertaken as a methodological sub-study to the German national aircraft noise project (DFG 1974; see also Rohrmann 1974), fully supported by the principal of the Social Psychology section of the DFG Project, Prof Irle. A report about part A of VQS-1 was provided by Rohrmann 1967 and Irle & Rohrmann 1968. For a full publication of Project VQS-1 (in German language) see Rohrmann 1978.

3.2 *Purpose of Project VQS-1*

The primary aim was to gain rating scales which were easy-to-use in surveys yet nevertheless satisfying psychometric and socio-linguistic standards. Accordingly, the research questions to be addressed were #1 and #2 of the research program outlined in chapter 2:

(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?

This meant as start, to identify as much relevant words or expressions as feasible, and to consider relevant types of appraisals dealt within basic or applied research actions.

3.3 *Research design*

Preface: Except of Box 3-1, all boxes contain tables or figures from the original German publication (Rohrmann 1978); some of these were scanned, and the quality may be inferior.

Overview

Conducting VQS-1 in 1966 was organized into four phases:

- Documentation of verbal scale point labels (VSPL) used so far in investigations,
- Experiments: Category scale rating of 77 VSPLs (words/expressions),
- Application of findings to scale construction for questionnaires,
- Qualitative interviews about their value.

This program is presented in the left part of *Box 3-1.*

Selection of words/expressions

The first step was an intense search for words or expressions which had been used, or could be used, in verbal rating scales. It was attempted to find all existing rating scales, and dictionaries were checked out as well (the almost 'official' one in Germany is the Duden (1959, 1964, 1972).

Verbal qualifiers are used to grade the degree to which a particular attribute is given. There are four fundamental judgement dimensions:

- *Intensitaet ~ Intensity [I]*, e.g., not, a little, rather, very, extremely;
- *Hauefigkeit ~ Frequency [F]*, e.g., never, sometimes, often, always;
- *Wahrscheinlichkeit ~ Probability [P]*, e.g., unlikely, hardly, possibly, for sure.
- *Bewertung von Aussagen ~ Agreement with statements [S]*, e.g., don't accept, agree, true for me.

Box 3-1

Overview data collection in VSQ-1

Study "1" - *Original investigation*	**Study "2" - *Retest of first investigation***
Exploration of suitable words/expressions	Modification of item list
Sample G = general public (N=2x30)	Sample G = general public (N=29) Sample S = students (N=33
[:] Experiments NW and WN	[:] both: Experiments NW and WN [:] both: Apraisal of scaling modes
Survey with interviewers (N=10) [:] Evaluation of new rating scales	Survey with interviewers (N=20) [:] Evaluation of new rating scales
Tested items: i=77	*Tested items: i=65*

Intensity and Frequency are the most general and most used dimensions, Probability has induced most experimentation. For each of these finally 18 items were chosen, plus 23 items for Agreement, altogether n=77. They are listed within *Box 3-2.* (The allocated letters are random nr's needed when designing experiments).

Scaling tasks

The core task was to quantify the meaning of all investigated Verbal Scale Point Labels (VSPLs); cf. *Box 3-1* above. To achieve this, the participants were asked to position each of the 77 items on a 9-level "equal appearing interval scale" (Thurstone 1928). "-4" was presented as lowest and "+4" as highest level of the respective dimension/attribute, e.g., Intensity. Using the resulting scalings, each word/ expression could get a descriptive number. This experiment was labelled *"Numbers for Words" (NW).* A similar procedure was first time explored by Thorndike in 1910; cf. Guilford 1954, p. 204).

In kind of a 'cross-road', a novel approach was instigated as a second scaling task. Participants were provided with a large model of a 5-point scale, labelled [-2|-1| 0 |+1|+2]. For each of the four modes, they should then suggest those five of the 18 or 23 available VSPLs which in their view (that is, their language) are the best markers for the 5 numerical scale levels. This second experiment was labelled *"Words for Numbers" (WN).* The resulting frequency distribution would indicate the best-suitable words/expressions for verbalizing a numerical rating scale.

Both tasks were handled via a fully standardized questionnaire. The 77 items were presented as cards. The sequence of the 4x2=8 tasks was randomized. The whole process was conducted by trained interviewers.

3.4 *Data collection and selected results*

Sampling

Given that the scaling tasks were quite demanding and time-consuming, it was decided to

Box 3-2

List of studied words & expressions in VSQ-1 - Study 1 & Study 2

Dimension H Häufigkeit:

oft (a), häufig (b), immer (d), gelegentlich (e), manchmal (g), kaum (j), selten (k), einigemal (l), nie (n), sehr oft (q), sehr selten (r);

(nur 1966:) vielmals (c), mehrfach (f), ein paar Mal (h), wenig (i), oftmals (m), niemals (o), immerzu (p);
(nur 1976:) ziemlich selten (w), ab und zu (x), ziemlich oft (y), meistens (z).

Dimension I Intensität:

ziemlich (a), sehr (b), außerordentlich (e), völlig (h), etwas (i), mittelmäßig (j), einigermaßen (k), nicht (l), wenig (m), kaum (n), gar nicht (o), annähernd (r);

(nur 1966:) ungemein (c), besonders (d), überaus (f), ganz (g), ein wenig (p), schwerlich (q);
(nur 1976:) halbwegs (w), überwiegend (y), teilweise (z).

Dimension W Wahrscheinlichkeit:

gewiß (a), zweifellos (b), vielleicht (d), wahrscheinlich (e), kaum (g), keinesfalls (i), sicher nicht (j), sehr wahrscheinlich (l), mit Sicherheit (m), möglicherweise (q), eventuell (r);

(nur 1966:) sicher (c), unter Umständen (f), schwerlich (h), wohl nicht (k), unter keinen Umständen (n), vermutlich (o), unwahrscheinlich (p);
(nur 1976:) wahrscheinlich nicht (w), ganz sicher (x), ziemlich wahrscheinlich (y), wenig wahrscheinlich (z).

Dimension B Bewertung von Aussagen:

sehr richtig (b), ziemlich richtig (c), etwas richtig (e), annähernd richtig (f), etwas falsch (n), annähernd falsch (o), sehr falsch (s), ziemlich falsch (t);

(nur 1966:) völlig richtig (a), starke Zustimmung (d), schwache Zustimmung (g), mehr richtig als falsch (h), unentschieden (i), ungewiß (j), neutral (k), teils richtig teils falsch (l), weder Zustimmung noch Ablehnung (m), schwache Ablehnung (p), mehr falsch als richtig (q), völlig falsch (r), starke Ablehnung (u), richtig (v), falsch (w);
(nur 1976:) teils/teils (x), etwas dagegen (y), etwas dafür (z), stimmt nicht (A), stimmt wenig (D), stimmt mittelmäßig (G), stimmt ziemlich (H), stimmt sehr (K), trifft wenig zu (M), trifft ziemlich zu (L), trifft gar nicht zu (P), trifft völlig zu (U).

Note: H= Frequency, I= Intensity, W= Probability, B= Agreement with Statements

have two samples, each N=30. Each sample got 2x2=4 tasks, to be carried out for all pertinent items.

The sample, "G", was set up as a quota sample of the general public, with an age range 21-60 years.

Mean scale positions: category scaling

Regarding the results for the category scaling task *"Numbers for words*" (NW), the data for one facet, Intensity, are presented in *Box 3-3,* upper left part. Listed are mean (M), modus (Md) and median (Mdn), plus standard deviation (s) re means, for all 18 items. The items are ordered according to their mean.

These NW results show: The VSPLs at the end of the range from 1 to 9 have strict means and low dispersion. For those who don't speak German: Rough translations are: "o" ~ not at all, "l" ~ not, "b" ~ very, "e" ~ very much or extraordinarily. The only other item with clear features is "j" ~ medium or middle level. In *Box 3-4*, the means are shown for five essential Frequency items (translated: never, seldom, occasionally, often, always), which shows a similar structure. If thinking about how to design a verbalized 5-point rating scale, the results are less indicative in the middle than the outer scale levels.

Box 3-3

Ratings of VSPLs in Study 1 and in Study 2 - Facet Intensity

Results Study "1" - Tasks NW and WN - Sample B1

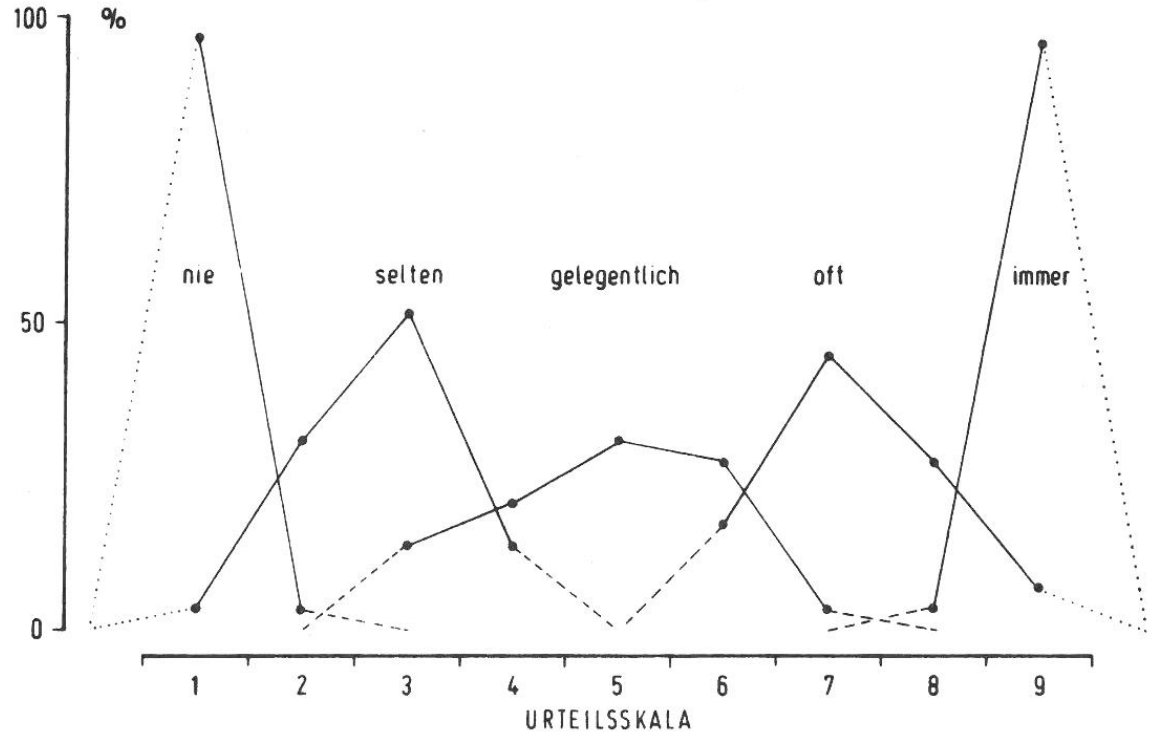
Informanten: „B1"	Skalierung WZ				Gruppierung ZW				
	Mittlere Skalenwerte				genannt für Stufe				
Begriffe	M	Md	Mdn	s	−2	−1	0	+1	+2
o gar nicht	1.0	1	1.0	0.18	!!!				
l nicht	1.4	1	1.0	0.63	!				
m wenig	2.8	2	2.5	1.22		!			
n kaum	3.1	2	2.8	1.48		!			
q schwerlich	3.4	3	3.2	1.79					
p ein wenig	3.7	2	3.4	1.57					
i etwas	4.7	6	5.1	1.71					
j mittelmäßig	5.1	5	5.1	0.85			!!		
r annähernd	5.2	6	5.4	1.42					
k einigermaßen	5.4	6	5.6	1.33					
a ziemlich	6.7	8	7.5	1.80				!	
d besonders	7.8	8	7.8	0.89					
g ganz	8.0	9	8.3	1.07					
f überaus	8.2	9	8.5	1.10					
h völlig	8.2	9	9.0	1.01					
c ungemein	8.4	9	9.0	0.86					
b sehr	8.5	9	9.0	0.78					!!
e außerordentlich	8.7	9	9.0	0.92					!!!

Results Study "2" - Tasks NW and WN - Samples B2 & S2

Informanten:	„B2"		„S2"		„B2" + „S2"				
	Skalierung WZ				Gruppierung ZW				
Begriffe	M	s	M	s	−2	−1	0	+1	+2
o gar nicht	1.0	0.00	1.0	0.00	!!!				
l nicht	1.5	0.51	1.5	0.57					
m wenig	2.6	0.68	2.7	0.57		!			
n kaum	2.8	0.64	2.5	0.57		!!			
i etwas	3.7	1.22	3.7	0.96					
k einigermaßen	4.9	1.22	4.7	1.13					
w halbwegs	5.0	0.73	5.0	0.61					
z teilweise*	5.0	1.03	4.4	0.87					
j mittelmäßig	5.2	1.07	5.1	0.66			!!		
r annähernd	5.4	1.40	5.3	1.33					
a ziemlich	6.0	1.14	6.4	0.90				!	
y überwiegend	6.9	0.95	7.2	0.68				!!	
b sehr	8.2	0.76	8.1	0.74					
h völlig	8.5	0.57	8.7	0.57					!!!
e außerordentlich	8.6	0.74	8.6	0.60					!

Box 3-4

Means and distribution for core frequency items - task NW



This is confirmed by the outcomes of experiment *"Words for Numbers" (*WN), shown in the upper right part of *Box 3-3*. For "-2" and "+2", the majority clearly confirmed one or two special words. The same is true for scale point "0". However, neither for "-1" nor for "+1" strong preferences occurred. Three words were nevertheless significant suggestions. (Note: in *Box-3-3,* !!! indicates at least ½, !! at least ⅓, ! at least ¼ of suggestions are for the shown item).

The data for the other three investigated judgement dimensions, i.e., Frequency, Probability and Agreement with statements (not shown in this report) provided mainly similar outcomes.

3.5 *Created rating scales*

In order to create verbalized rating scales which are functional in both lab and field research, the directive was: Logical, coherent, easy-to-understand devices. Furthermore, it was to decide which role, beside verbal means, numerical and graphical means should have. Consequently, three decision were needed, number of scale levels, their designation, and the layout of the instrument.

Number of levels: Two aims are in conflict, on one hand, to measure very detailed, on the other hand, to reflect how fine respondents can differentiate. In academic research, at that time 5 to 10 levels were used (remember Miller's "7 plus or minus 2" view, based on Psychology and Information Theory). In demoscopic surveys, "yes/no" or "low/medium/high"

were common. Given the context of field research, a 5-point scale was selected as best solution. It could be explained to unexperienced respondents as two levels of yes and two levels of no, around the neutral mid-category. The original numbering as [-2|-1| 0 |+1|+2] was later changed into [1 | 2 | 3 | 4 | 5] - this was clearer perceived as equidistant, and easier to handle in interviews.

The essential task was selecting VSPLs. First options were words/expressions for 1/3/5/7/9 or 2/3.5/5/6.5/8. Pretests made clear that the 1/3/5/7/9 option leads to a rating scale in which the endpoints are too strict and hardly used, and that the risk of the 2/3.5/5/6.5/8 option is endpoints being too meagre. Therefore the pattern 1.5/3.25/5/6.75/8.5 was used for selecting verbalizations.

The final choice had to follow these principles:

(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 particular scale,
(4) sufficient familiarity of the expression,
(5) reasonable likelihood of utilization when used in substantive research,
(6) practicability in oral interviews.

Of course not all rules could be realized perfectly; altogether points (1) and (3) turned out to be essential.

In *Box 3-5* three of the created rating scales are presented - these are the German originals.

Both the Frequency and the Intensity VSPLs are important beyond a tool by itself, because they are adverbs and can be connected with adjectives or verbs for second-order scales. Some examples: "I am never happy, seldom happy, sometimes happy, often happy, always happy (Happiness scale)", or "He struggled not/a little/moderately/quite a bit/very much (Competence scale)". Importantly, Cliff's Law (Cliff 1959, Kristof 1966) states that the graduation impact of qualifiers is quite stable across different contexts.

Regarding the scale layout: The examples in *Box 3-5* show that several means were employed to enhance the perception of an equidistant scale: The five levels are presented as an equally-sized structure, which compensates that the words differ in length, and the symbols [--| - | **:** | + |++] are added. Furthermore, different colours are used for the four rating scales, which facilitates the communication with respondents.

Lastly, an exploration with interviewers (N=10) was conducted, to get an evaluation of the new rating scales in terms of their practicality, and this was utilized for final adjustments.

All scales were then immediately used in the surveys of the German national aircraft noise project (final report: DFG 1974), within its Phase I and then Phase II, and soon employed by other investigators as well.

The involved researchers, from several Social Science disciplines, provided mostly very positive feedback, regarding both methodology and usability aspects.

Box 3-5

New verbalized rating scales developed in Project VQS-1

Intensität ~ Intensity

Häufigkeit ~ Frequency

Bewertung von Aussagen ~ Apraisal of statements

3.6 *Replication study*

Issues

The outcomes of Project VQS was surprisingly successful - yet over the years a discourse about validity issues also started. Responding to this, main questions were identified:

- ♦ Are the identified features of the chosen words/expressions stable over time?
- ♦ Would the judgements of university students be similar to the population sample or not?
- ♦ What about scale labels which are meanwhile familiar yet were not included in the original study?

Retest design

The principal design of Study 1, regarding the testing of words/expressions as VSPLs for rating scales, was maintained (see right part of *Box 3-1*), however, the agenda was extended as follows:

Items:

After an expert debate, 23 new items were added, and items rated as less important were removed; this resulted in 15+15+15+20=65 items (*Box 3-2* for details). The four judgement dimensions, Frequency, Intensity, Probability, Agreement with statements, were maintained.

Sample:

To extend the validity, two samples were investigated: Sample "G", N=29, set up as a quota sample of the general public, with an age range 21-60 years (equivalent to Study 1). Sample "S", N=33, University students in Social Science.

Data collection:

The procedure was similar to Study 1, based on a fully standardized questionnaire, with enhanced instructions. The 65 items were again presented as cards. T

Main retest results

There are 2x4x2=16 data sets, for 2 samples, 4 dimensions, 2 tasks (NW, WN). In the second part of *Box 3-3* (above), the results for the facet "Intensity" are listed.

The means (M) for the "Numbers for Words" task provide a clear structure, and most standard deviations (s) are mostly below 1.0. The judgements of the two groups are quite similar and correlate 0.97; only for one item, "z" (partly) the difference is significant.

Regarding "Words for Numbers", the results of the two groups were similar and thus merged. The respondents agreed considerably in suggested verbal labels - for the five scale levels one or two words/expressions dominated the responses (see the right lower part in *Box 3-3*).

3.7 *Comparison of results from Study 1 and Study 2 of VQS-1*

For one facet, Intensity, the results can be compared in *Box 3-3*; however, be aware that only 12 Intensity items were investigated in both studies, and that only sample "P" can be used. The scores are very similar (mostly differing less than 0.5), and the rank correlation for the 12 maintained items is 0.99. The only instable item is "a" (rather or quite), with means of 6.0 and 6.7. Furthermore, about 2/3 of the "Words for Numbers" judgements are the same. Even though the context of the 12 maintained items (within 18 or 15 tested items) of course varied, this had little influence on the scores.

Altogether the retest - conducted 10 years after the primary investigation - showed that the appraisals of the examined Verbal Scale Point Labels are mostly stable, especially regarding words/expressions which had been selected for verbalized rating scales in Study 1 (shown in *Box 3-5*).

In addition to the experiments, again a survey with interviewers was carried out (N=20, 10 each linked to samples "P" and "S"), in order to collect qualitative feedback about verbalized scales, and to explore the familiarity of pertinent instruments. Altogether their views were in favour of the scales resulting from Study 1.

This group was also interrogated about their views regarding 5-point scales and experimental 11-point rating scales; these are shown in *Box 3-6*. In group "P" keenness for

5-point scales dominated, while in group "S" most preferred one of the 11-point scales, mainly the fully verbalized version "X".

Box 3-6

Two examples for verbalized 11-point scales

Scale X: Appraisal ***(bi-polar 11-point)***			***Scale Y: Graduation*** ***(unipolar 11-point)***		
außerordentlich gut	+5	|	.	überhaupt nicht	0
sehr gut	+4	|	I		1
gut	+3	|	II		2
ziemlich gut	+2	|	III		3
mehr gut als schlecht	+1	|	IIII		4
mittelmäßig	0	|	IIIII		5
mehr schlecht als gut	-1	|	IIIIII		6
ziemlich schlecht	-2	|	IIIIIII		7
schlecht	-3	|	IIIIIIII		8
sehr schlecht	-4	|	IIIIIIIII		9
außerordentlich schlecht	-5	|	IIIIIIIIII	außerordentlich	10

3.8 *Decision about verbalized 5-point scales*

The final question in Study 2 was whether the gathered reliability and validity information would justify to maintain the verbalized 5-point rating scales from Study 1, or whether they needed to be changed.

The six principles outlined above (see section 3.5) were again applied to the VSPLs. The new data from task NW justified the existing verbalizations. A further supporting point was that the suggested words for a five-point scale (task WN) came out quite similarly. Finally, none of the new words/expressions in Study 2 were coherently better than the not-fully-convincing VSPLs in Study 1 (example: level 2 and level 4 in the Intensity scale).

Consequently, after a careful team discussion it was decided to maintain the four created instruments as they are, and continue to use them in both experimental and applied research.

3.9 *Evaluation of the VQS approach*

Realizing the two studies of Project VQS was a considerable effort, and methodogically quite demanding as well. Altogether the two - almost contradictive - aims, namely, to create a decent psychometric instrument, and to design a user-friendly tool, could be reasonably achieved.

The outcomes re practicality of the generated verbalized rating scales are mostly positive; they have been employed for many years in a multitude of ventures and are still in use. This is certainly relevant for research outside the 'lab', especially population surveys.

The issue how equidistant the produced instruments are is more of a problem. If words/expressions shall be used for verbalizing a rating scale, then their psychometric features are not the only factor - they must be satisfactory in socio-linguistic terms as well. These two criteria clash to some degree, and require a compromising choice. The scale for Intensity may be used as an example, it is [nicht | wenig | mittelmäßig | ziemlich | sehr]. The pertinent data from the NW task are [1.4 | 2.8 | 5.1 | 6.7 | 8.5] in Study 1, [1.5 | 2.7 | 5.1 | 6.4 | 8.2] in Study 2. The ideal scores would have been [1.5 | 3.25 | 5.0 | 6.75 | 8.5]. This means that the selected words were near to the target yet not perfect. They were chosen because they are very common for naming five levels, plus, the considerate design of the scale layout, including numbers (as shown in *Box 3-5*) was meant to increase the perception of equidistance.

Interestingly, the collected data, especially regarding task NW, can be utilized to inspect the words/expressions used in existing verbal rating scales - doing this will identify a set of reasonable instruments, yet also quite a few instruments which are, in psycho-linguistic terms, simply wrong.

The reported two studies had some shortcomings. First of all, the sample sizes were very small. Furthermore, potential regional differences (German was/is the language in 4 countries) could not be investigated. Finally, the chosen methodology did not include a multi-dimensional scaling.

When publishing VQS-1, in 1978, a couple of suggestions for further research were outlined. This included: To investigate verbalizing 7-point scales; to compare uni-polar and bi-polar scales, to explore further combined scales (such as intensity and personality attributes), and to test systematically the impact of scale features such as numbers, words, symbols, and combinations thereof (Hennig 1975 claimed that these elements stabilize scale validity).

Finally, researching "verbally qualified scales" ~ VQS can, and should, be done in other languages as well, especially those which are important in Psychology and Sociology research, such as English, French, Spanish, Russian, Chinese, Japanese language and so on, all part of the social science research community.

<4> COMPOSING ENGLISH VERBALIZED RATING SCALES - VQS-2 - IN AUSTRALIA

4.1 *Objectives of Project VQS-2*

The success and considerable use of verbalized rating scales in German language, which were crafted with psycholinguistic and psychometric concepts, led to the plan to achieve the same for scales in English language. This initiative was enhanced by the fact that social science research is pretty much dominated by Anglo-American countries.

Consequently, the three main research objectives for Project VQS-1 (detailed above in Chapter 3) were underlying VQS-2 as well:

(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?

Accordingly, both theory and methodology (as discussed in Chapter 2) had to be maintained as firm as possible - however, before developing the definite research plan, it was vital to check out whether particular characteristics of the English language would request a modified approach.

Also, the existing studies from authors in GB and USA (Budescu & Wallsten 1994, 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, Theil 2002, Windschitl & Wells 1996, Wright et al. 1994) had to be carefully studied. Most of them dealt with probability scaling. The goal was to include words/expressions into VQS-2 which had repeatedly been examined.

Besides, the findings were expected to help with a long-discussed matter, whether plain verbal rating scales should have 5 or 6 or 7 scaling levels (e.g., Cox 1980, Preston & Colman 2000, Sauro 2010, Spector 1993) - results could be used for various scale lengths.

4.2 *Research design*

Principal approach

Following the rationale of Project VQS-1 in Germany, the principal concept was, 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 - now English ones - 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 acquire the necessary information, again 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 (Dunn-Rankin 1983, Wegener 1983). Furthermore, context effects were to be controlled by using several linguistic 'frames' for the qualifiers under study.
The project was organized into four 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,
- Application of findings to scale construction for questionnaires.

The outcomes will be presented in a condensed concise mode; for a detailed report see Rohrmann 2007.

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, focussed on English-speaking countries.

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

- *Intensity [I]*, 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,
- *Agreement with statements [S]*, e.g., don't accept, agree, true for me.

These had been investigated in Project VQS-1. They can be used in manifold combinations with substantive attributes, usually expressed as either verb phrases (e.g., I am happy, I use trams, my agreement is) or adjectives (e.g. satisfactory, annoyed).

One further type of judgments is frequently used in social science research and therefore deserves attention:

- *Quality [Q],* e.g., bad, acceptable, satisfactory, excellent;

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 German

items of VQS-1. An Australian study by Clarke et al. 1992 was also considered. *Box 4-1* provides a list of all items.

Box 4-1

List of all items used in Project VQS-2

<F> FREQUENCY	<P> PROBABILITY	<S> (DIS-) AGREEMENT WITH STATEMENTS
always	about 50-50	agree
fairly often	a very. good chance	disagree
frequently	certainly	* don't agree
mostly	certainly not	fairly true for me
never	for sure	* fully agree
occasionally	likely	* fully disagree
often	no chance at all	* half-half
moderately often	perhaps	in-between
rarely	possibly	* mainly agree
seldom	probably	* mainly disagree
sometimes	probably not	mostly true for me
very often	quite likely	neither agree/disag
	unlikely	neutral
<I> INTENSITY	under some circumstances	not true for me
a little	under most circumstances.	right
average	with certainty	somewhat agree
completely		somewhat disagree
considerably	<Q> QUALITY	s/what true for me
extremely	adequate	strongly agree
fairly	* average	strongly disagree
*fully	bad	true for me
hardly	dissatisfied	undecided
highly	excellent	
* in-between	fair	
* mainly	good	
medium	inadequate	
moderately	medium	
	mostly dissatisfied	
not	mostly satisfied	
not at all	not too bad	
partly	outstanding	
quite	poor	
* quite a bit	satisfactory	
rather	satisfied	
slightly	so so	
somewhat	unsatisfactory	***Note:***
very	very good	*Items labelled with* *****
very much	very satisfied	*were not used*
	very dissatisfied	*in all sub-studies*

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:*

In Project VQS-1, some information about the familiarity of VLPs had been gained in a qualitative exploration. Now this feature was measured: On a 0-to-10 scale, for each VSPL it was judged how common and familiar it is in everyday language.

Furthermore, the studied 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. An overview is provided in *Box 4-2*. The participants were recruited from psychology students and the general population; for each sub-group, the target N was 40.

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 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.

Box 4-2

Data collection VQS-2: 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.

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

4.3 *Data collection and selected results*

Preface: The Project VQS-2 resulted in 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. *Box 4-2*). 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 *Box 4-3*, which consists of 5 parts. 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.

Box 4-3-I

Main results for "Intensity" qualifyers

Scaling task	CATEGORY (0...10 scale)				MAGNITUDE			PREFERED LABEL for levels (%)					FAMILIA-RITY	
Context:	all		noise		<NM>	<X_{nl}>		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.

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 negative as "bad" (rated 1.0).

Box 4-3-F

Main results for "Frequency" qualifyers

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 1 verbal label for each of the levels "1" to "5".

Box 4-3-P

Main results for "Probability" qualifyers

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 1 verbal label for each of the levels "1" to "5".

Box 4-3-Q

Main results for "Quality" qualifyers

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 1 verbal label for each of the levels "1" to "5".
"--": No data collected.

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 *Box 4-4* (see further below) illustrates the relationship between M and sd for the intensity labels.

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)? Unfortunately this is difficult to assess, as the

Box 4-3-A

Main results for "Agreement" qualifyers for statements

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														
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

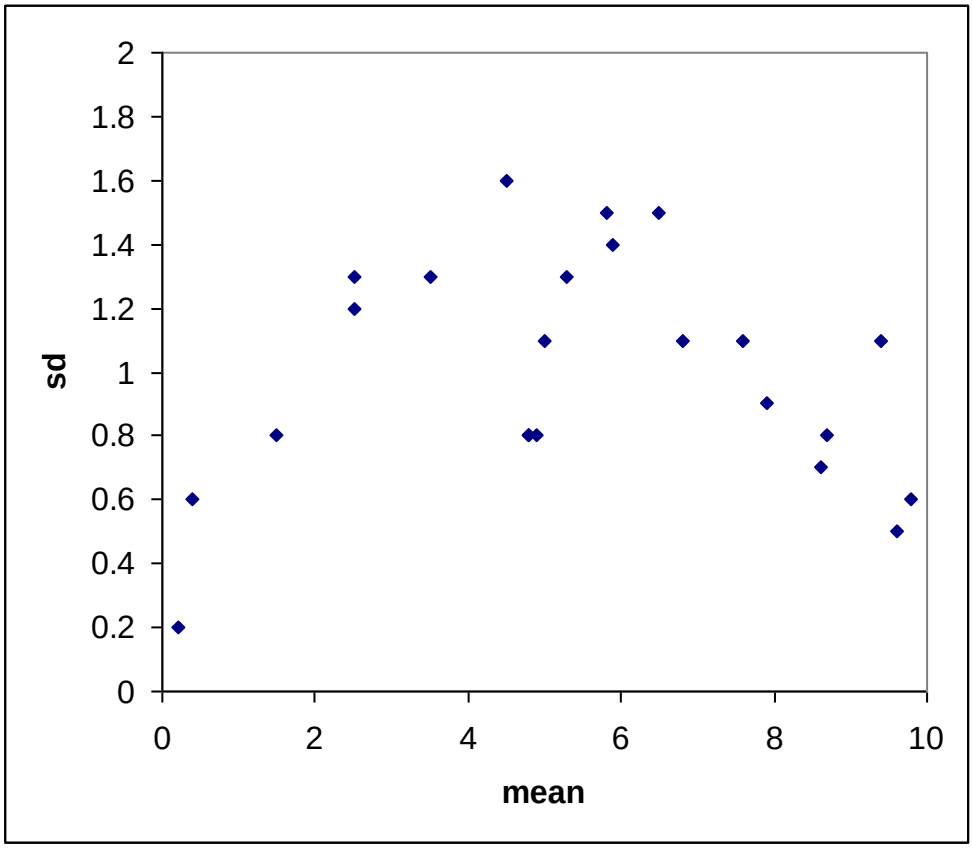
Notes:
"Magnitude" data: GM= geometric mean; Nm= number modality, standardized raw scores; X_{nl}= scores based on merged number/lines responses.
"Prefered label": respondents had to suggest 1 verbal label for each of the levels "1" to "5".
"--": No data collected.

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. *Box 4-3*). 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).

Box 4-4

Mean ratings and sd for Intensity VSPLs (task "NW")



The application of the scaling results to rating scale construction will be discussed in the final section of chapter 4.

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 three tables of *Box 4-3* 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".

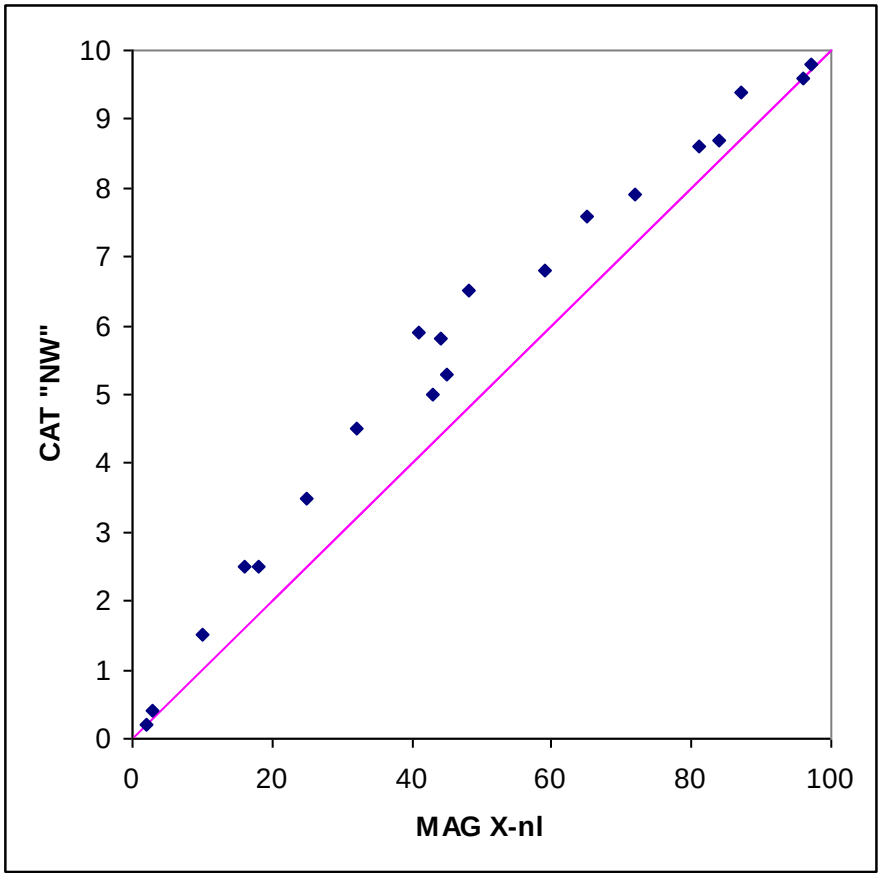
However, it seems questionable to take these data literally (*sensu,* "very good" is 10

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. *Box 4-5* 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).

Box 4-5

Relationship between Category and Magnitude Scaling Results, for Intensity VSPLs



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 the five tables in *Box 3*. All words/expressions are rated as at least moderately familiar (i.e., 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 the tables of *Box 4-3*. They 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 *Box 4-3*, 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 pre-tests 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 effects of age, education, and language preferences in sub-cultures. Given the small 'general population' sample (this part of the project could not yet be completed), only exploratory analyses were 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 data seem to indicate that the understanding of the VSPLs scaled in this project is consistent and not specific for a student population.

4.4 *Created rating scales*

As the results presented in *Box 4-3* show, for both 5-point and 7-point rating 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 rating 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. Pre-tests can help to decide whether it is better to avoid the top-end VSPL.

In addition to the labelling 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 visual/graphical means, in order to enhance both psychometric quality and user-friendliness.

In *Box 4-6*, two of the finally chosen scales are shown. Like in the German scales developed in VQS-1, a multi-modal scale design is realized, using --//-//0//+//++ as numerical notes, and equidistant frames as visual/graphical means.

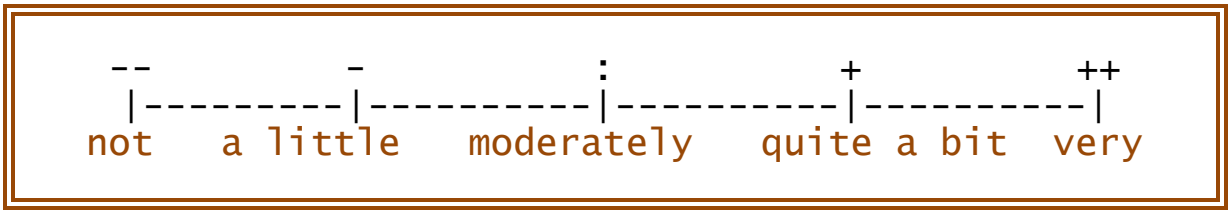
The relevance of the available design attributes was explored in 'meta-interviews' with interviewers who were trained to present different scale modes to their interviewees; the same had been done in the original German study VQS-1. The findings were considered for the final suggestions for verbalized multi-mode rating scales. In any case, modified or new instruments do need to be tested carefully with relevant target groups before installing them for permanent use.

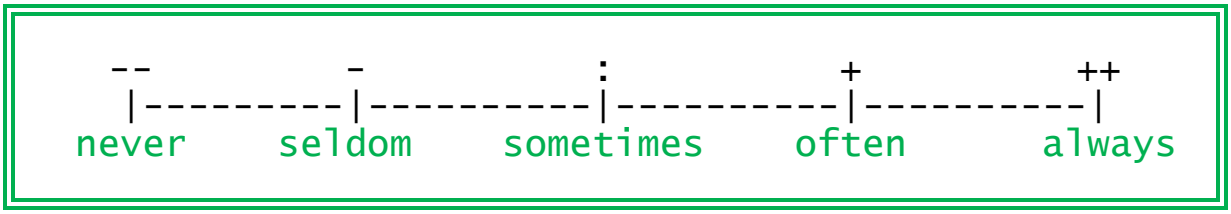
4.5 *Appraisal and conclusions*

Altogether the outcomes of Project VQS-2 were encouraging - they conformed, as Project VQS-1 had done in Germany, that English-language rating scales can be enriched by verbal

Box 4-6

Final English-language scales - examples Intensity and Frequency





scale point markers, and that careful psychometric research can provide instruments which are reasonably equi-distant.

Consequently this leads to the question how the psycholinguistic links between rating scale labels in different languages are - as stated in the introduction as research aim

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

Thus at the conclusion of Project VQS-2 its continuation was discussed. One straight-forward option is a study which compares verbalized rating scales in English and German language, focussing on the self-created ones. A wider-reaching approach would be to compare English or German instruments with rating scales used in another language.

Obviously these kinds of experiments are designed for bi-lingual respondents - which means that the sampling of research participants is much more demanding. Nevertheless it should be worthwhile because such cross-cultural investigations hardly exist.

<5> CREATING CHINESE SCALES & TESTING CHINESE-ENGLISH LINKAGE: VQS-3

5.1 *Disparities among Chinese and 'Western' ratings*

Cross-cultural issues regarding verbalized rating scales were first realized when scales had to be created within or translated into another language.

Concerning the VQS design, during prestudies for Project VQS-2 it was noticed that there seem to be principal differences between people from a Chinese background and those in 'Western' countries when rating behaviors or objects. For example, the focus of rating scale anchors is a factor - Chinese tend to focus on the middle scale points whereas Westerners pay more attention to the extreme anchors when making judgements. This needs to be considered in deciding about VSPLs, otherwise the advantages of using a VSPL-set may diminish (see e.g. Auer et al. 2000, Chen et al. 1995, Van de Vijver 2001).

The role of VSPLs in the translation of scales provides important insights. Studies often relied on scales to study constructs or variables on the target population (Chen et al. 1995), and in cross-cultural studies, scales are typically translated from a host language to another with an emphasis on readability and convenience of the respondents. According to Auer et al. (2000), cultural differences, linguistic problems and the psychometric properties of the scale have to be addressed in considering the translation of scales. Past research has noted that the choice of words in VSPLs has played a large part in deficiencies on measurement (e.g., Schwarz et al. 1993). This further renders the cross-cultural comparability of rating scales more difficult, as it may be intricate to develop correspondence phrases or expressions between languages such as Chinese and English, which are the two languages focused upon in this study, VQS-3.

Prior research found differences in Chinese and English probability root words in constructing VSPLs. Lau & Ranyard (1999) suggested one of the differences is that the Chinese root words' numerical probability meanings may be vaguer compared to English. For example, the term "Keneng" (可能) has a range of perceived numerical probability from 20 to 95%, whereas the corresponding word in English, "possibly", has a range from 25 to 75%. Similarly, Xu & Li (2007) have demonstrated the non-correspondence and ambiguous nature of root words. They found that Keneng (可能), with a mean numerical probability is 55%, is usually translated to correspond with "probable", "possible" and "perhaps" – words with substantially different mean numerical probabilities (74, 38 and 39%, respectively). Therefore, the translation of the VSPL "probably" in a questionnaire from English to "Keneng" (可能) in Chinese is likely to yield a different pattern of response, owing to differences, not in the construct of concern, but rather to differences in scale interpretation.

As mentioned earlier, possible differences in how VSPLs are perceived across individuals, both within and between cultures, have received little attention. To our knowledge, only the development of a noise annoyance scale has studied this problem systematically (Fields et al. 2001, Felscher-Suhr et al. 1998, Guski et al. 1998; Yano et al.

1998 took care of Japanese and further Asian languages; see also Rohrmann 1998;). This utilized the approach created in VQS-1 (cf. Rohrmann 1978). It is essential to design scale anchors very carefully if equidistant and unambiguous instruments are to be achieved, and this effort requires psychometric data for scale labels. For the English language, VQS-2 (cf. Rohrmann 2007) provided pioneering work, which was to be exploited in VQS-3 as well. Informing about risk levels is also affected by the used scaling words (Barilli et al. 2010, Dieffenbach et al. 1993, Rohrmann 2000, Theil 2002); this is likely to be different between cultures and countries.

5.2 *Project VQS-3: New aims*

Operating in a university where many lecturers as well as students came from Asian countries revealed firstly that these cultures were not much considered in designing suitable response scales for Social Science research, and secondly that the knowledge about comparable words/expressions in English was quite restricted.

Thus VQS-3 was shaped, utilizing a research stay at the Chinese University of Hong Kong. Two vital research objectives of Project VQS-1 and VQS-2 were maintained:

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

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

These tasks had now to be pursued in a cross-cultural approach, and the essential ambition was the new research question (6):

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

To deal with this topic, data collections regarding Chinese and English items were necessary, plus linking procedures.

For conducting this third VQS investigation, the most complex and first bilingual one about verbally qualified scales, it was certainly intended to maintain both theory and methodology as stringent as possible - however, the new context, China, obviously requested an adapted conceptualization.

5.3 *Designed sub-studies*

A set of experiments was set up. The principal approach was to first collect series of VSPLs and then to quantify their core characteristics. The item ratings were administered through a computer program. Chinese and English VSPLs were either presented separately or jointly, depending on the scaling task. VQS-3 was carried out in five phases:

- Documentation of verbal scale point labels (VSPL) used in Chinese studies,
- Pre-study to select most relevant VSPLs,
- Scaling of Chinese and English qualifiers,
- Cross-language matching tasks,

- Using results for creating verbalized 5-point scales.

Three of the five rating scale modalities investigated in VQS-2 were covered: Intensity, Frequency and Agreement with statements, because these are the most-used ones.

5.4 *Experiential and experimental data collection*

Collecting VSPLs

At first an exhaustive search of Chinese qualifiers for the rating scales of interest was conducted. This included Chinese research journals, dictionaries, handbooks (e.g. Howard 2002) and brainstorming by two Chinese authors. It resulted in 238 Agreement, 218 Intensity, and 160 Frequency qualifiers.

In order to reduce these very large lists, in a prestudy those VSPLs were identified which have a high level of familiarity, cover the full range of appraisals, and are linguistically compatible with use in typical rating scales. The outcome was 24 Agreement, 19 Frequency, and 18 Intensity Chinese qualifiers.

Finally, a list of 44 English-language qualifiers (19 Intensity, 12 Frequency and 13 Agreement with statements) was extracted from Project VQS-2.

Scaling experiments for Chinese and English qualifiers

The experimental setting, presented in *Box 5-1*, gives an overview of the four tasks of scaling Chinese and English words, and shows that counterbalancing conditions were set up for each experiment.

Box 5-1

Experimental setting: counterbalancing conditions of scaling Chinese and English words

	Counterbalancing conditions			
	1	2	3	4
Task 1 Familiarity	Always Chinese & English words together			
Task 2 Number-for-Words	Chinese then English words	English then Chinese words	Both Chinese & English words together	
Task 3 Word-for-Numbers	Chinese then English words	English then Chinese words	Chinese then English words	English then Chinese words
Task 4 Cross-language Matching	Match English to Chinese words then match Chinese to English words	Match Chinese to English words then match English to Chinese words	Match English to Chinese words then match Chinese to English words	Match Chinese to English words then match English to Chinese words

For all experiments the data collection was mostly handled via visuals on computer screens, not in a "paper+pencil" mode, and responses were made via mouse. This is fully described in Au et al. 2011.

Sampling:

Two samples were employed, (a) a student sample of N=110, and (b) a general public sample of N=190. Either 60 or 50 of the students were allocated to the main experiments. All participants were bilinguals who can read both Chinese and English. The general (non-student) sample was recruited through the Chinese authors' personal network; it got a less complex set of scaling tasks. The age range was 22 to 60 years; the female/male proportion was even (51:49%).

Data collection:

Regarding students, groups of 20 at a time did the scaling experiments at computers in a prepared research labs. Regarding the general sample, all participants had computers and e-mail connection at home, did the tasks as steered by the experimental instructions, and then returned their responses.

5.5 *Selected results*

A modest selection of outcomes will be presented below; for detailed information see the publication by Au et al. 2011 and the report by Rohrmann, Au & Tailor 2008.

Task (1) Familiarity:

For the Chinese VSPLs, the mean ranking on a 0-to-10 point scale was 7.1 for Intensity items, 7.1 for Frequency items, and 6.6 for Agreement-with-statement items. For the English items, the pertinent scores were 6.0, 6.1 and 5.4. These scores came lower because of just a few words/expressions uncommon to some of the Chinese participants. For "Intensity", the results can be seen in Box 5-2 (below).

In general, it appeared that most of the investigated Chinese and English qualifiers were commonly used by this sample of bilingual Chinese respondents in their daily written usage.

Task (2) Number-for-Words:

In the evaluation "Numbers for Words" all words/expressions get rated on a 0-to-10 scale, as in projects VQS-1 and VQS-2 (see chapters 3 or 4 for the methodology). The results for one facet, "Intensity", are listed in *Box 5-2.* (For full results see Rohrmann et al. 2008).

The range of the means for Chinese Intensity qualifiers is 0.6 and 9.7, and that of the English words is 0.7 to 9.8. Standard deviations (sd) range from 0.4 to 2.1. A few of the English VSPLs had a high sd. It is likely that some Chinese respondents in this study were not proficient enough in English to appraise all Intensity words precisely.

Task (3) Words-for-Numbers:

In the "Words for Numbers" task, respondents had to suggest for each of five numeric levels (1-2-3-4-5) the best-suited word/expression as descriptor. For "Intensity", these results are given in the right part of *Box 5-2*.

The preferences are reasonably clear for levels "1", "3" and "5", less so for "2" and "4". This is the case for both Chinese and English words. For English, preferences were similar yet not identical in Projects VQS-2 and VQS-3.

Box 5-2

Scaling results for intensity verbal scale point labels

	Verbal Scale Point Label VSPL		Familiarity M	Familiarity sd	Nr-for-Words M	Nr-for-Words sd	Word-for-Numbers 1	2	3	4	5
C12	一點也不	YiDianYeBu	5.2	2.7	0.6	1.9	75	¤	¤	¤	¤
C10	不	Bu	8.6	2.2	0.8	0.9	22	25	¤	¤	¤
C06	少許	ShaoXu	6.2	2.2	3.3	1.7	¤	36	¤	¤	¤
C01	或許	HuoXu	6.7	2.1	3.6	1.5	¤	¤	¤	¤	¤
C05	也許	YeXu	6.5	2.5	3.7	1.7	¤	¤	¤	¤	¤
C02	有點兒	YouDianEr	6.4	2.4	3.9	1.9	¤	13	¤	¤	¤
C07	稍為	ShaoWei	4.9	2.4	4.5	1.7	¤	¤	¤	¤	¤
C03	有些	YouXie	8.1	1.7	4.8	1.6	¤	¤	¤	¤	¤
C04	一般	YiBan	7.8	1.6	5.0	1.0	¤	¤	72	¤	¤
C09	大概	DaGai	7.1	2.1	5.2	1.7	¤	¤	¤	¤	¤
C08	頗	Po	6.8	2.0	6.4	1.3	¤	¤	¤	23	¤
C11	很	Hen	8.9	1.5	7.9	0.9	¤	¤	¤	30	¤
C16	十分	ShiFen	8.5	1.6	8.7	0.9	¤	¤	¤	13	¤
C14	非常	FeiChang	8.8	1.4	8.9	0.7	¤	¤	¤	¤	¤
C17	肯定	KenDing	6.9	2.3	9.0	1.3	¤	¤	¤	¤	¤
C13	最	Zui	8.1	2.1	9.6	0.8	¤	¤	¤	¤	13
C18	極之	JiZhi	6.1	2.4	9.6	0.5	¤	¤	¤	¤	43
C15	完全	WanQuan	6.7	2.3	9.7	1.0	¤	¤	¤	¤	28
E12	Not		8.7	2.2	0.7	0.9	38	15	¤	¤	¤
E13	Not at all		5.7	2.8	1.1	2.1	54	¤	¤	¤	¤
E08	Hardly		5.6	2.3	2.0	2.5	¤	¤	¤	¤	¤
E01	A little		5.9	2.2	3.1	1.7	¤	30	¤	¤	¤
E17	Slightly		5.8	2.3	3.3	1.8	¤	23	¤	¤	¤
E18	Somewhat		3.8	2.7	4.0	1.7	¤	¤	¤	¤	¤
E14	Partly		5.8	2.3	4.6	1.5	¤	¤	¤	¤	¤
E02	Average		6.3	2.2	4.9	0.9	¤	¤	52	¤	¤
E10	Medium		4.5	2.4	5.0	0.9	¤	¤	14	¤	¤
E11	Moderately		4.7	2.5	5.3	1.1	¤	¤	13	¤	¤
E16	Rather		6.3	2.1	5.3	1.8	¤	¤	¤	¤	¤
E06	Fairly		5.2	2.3	5.5	1.7	¤	¤	¤	¤	¤
E15	Quite		8.0	2.0	6.1	1.7	¤	¤	¤	¤	¤
E04	Considerably		4.5	2.6	6.4	1.9	¤	¤	¤	14	¤
E09	Mainly		6.1	2.0	8.0	1.0	¤	¤	¤	¤	¤
E19	Very		8.8	1.4	8.6	0.7	¤	¤	¤	42	¤
E07	Fully		5.3	2.4	9.5	1.1	¤	¤	¤	¤	¤
E03	Completely		6.0	2.4	9.8	0.5	¤	¤	¤	¤	33
E05	Extremely		6.7	2.3	9.8	0.4	¤	¤	¤	¤	60

Note: .Familiarity ratings range from 0 = “Extremely unfamiliar” to 10 = “Extremely familiar.” Number-for-words ratings range from 0 = “Extremely low intensity” to 10 = “Extremely high intensity”. The

numbers under the Word-for-numbers columns are percentages of respondents choosing that VSPL for a particular five-point scale level, (separately for Chinese and English VSPLs). Percentages smaller than 12.5% for Chinese VSPLs and 11.5% for English VSPLs are not shown.

Task (4) Cross-language Matching:

In this task, the most demanding one, the participants were asked to match Chinese and English qualifiers within each of the three modalities of VSPLs. As described in *Box 5-1* (above), this was carried out in four experimental conditions, with different sequence of presenting the two languages in various tasks so as to counterbalance the possible bias responses due to the sequencing of the language presented. It was conducted for all three tested VSPL types - Intensity, Frequency and Agreement-with-statements qualifiers. For one of these, Intensity, the results are presented here, see *Box 5-3*.

The Chinese qualifiers are listed in the leftmost column and the English qualifiers are listed in the top row. The entries in each cell indicate the percentages of participants mapping the Chinese and English qualifiers with each other. For example, the 2nd left cell indicates the mapping between YouDianEr and “A little”. The first figure in the bracket was the percentage of participants (52% in this case) mapping the English word "A little" when given the word YouDianEr. The second figure in the bracket shows the percentage of participants (27%) choosing the Chinese word YouDianEr when presented with the word "A little". The number above the bracket (39%, in bold) is an average of these two figures showing the associations of the Chinese and English qualifiers. In this table, only scores larger than 20% of participants are shown. The rightmost two columns show the English word that was most frequently linked to the corresponding Chinese word, and the pertinent percentage. The bottom rows show the Chinese words that were most frequently linked to the corresponding English words listed on the top row.

These findings are essential for the cross-cultural agenda of Project VQS-3. The data show a wide variety of clear Chinese-English allocations (above 50%, up to 93%) versus VSPLs for which no convincing counterpart came out (i.e., below 50%, down to 22%). Obviously this means that Chinese-English and English-Chinese translations are not, and cannot be, an easy venture.

Comparison of VQS-2 and VSQ-3 outcomes for English VSPLs

For the English VSPLs, the VQS-3 data collected via bi-lingual samples could by compared with the findings in VSQ-2, gained in with an Anglo-Australian sample. Of main interest are the "Numbers-for-Words" (NW) results, which were measured on 0-to-10 scales.

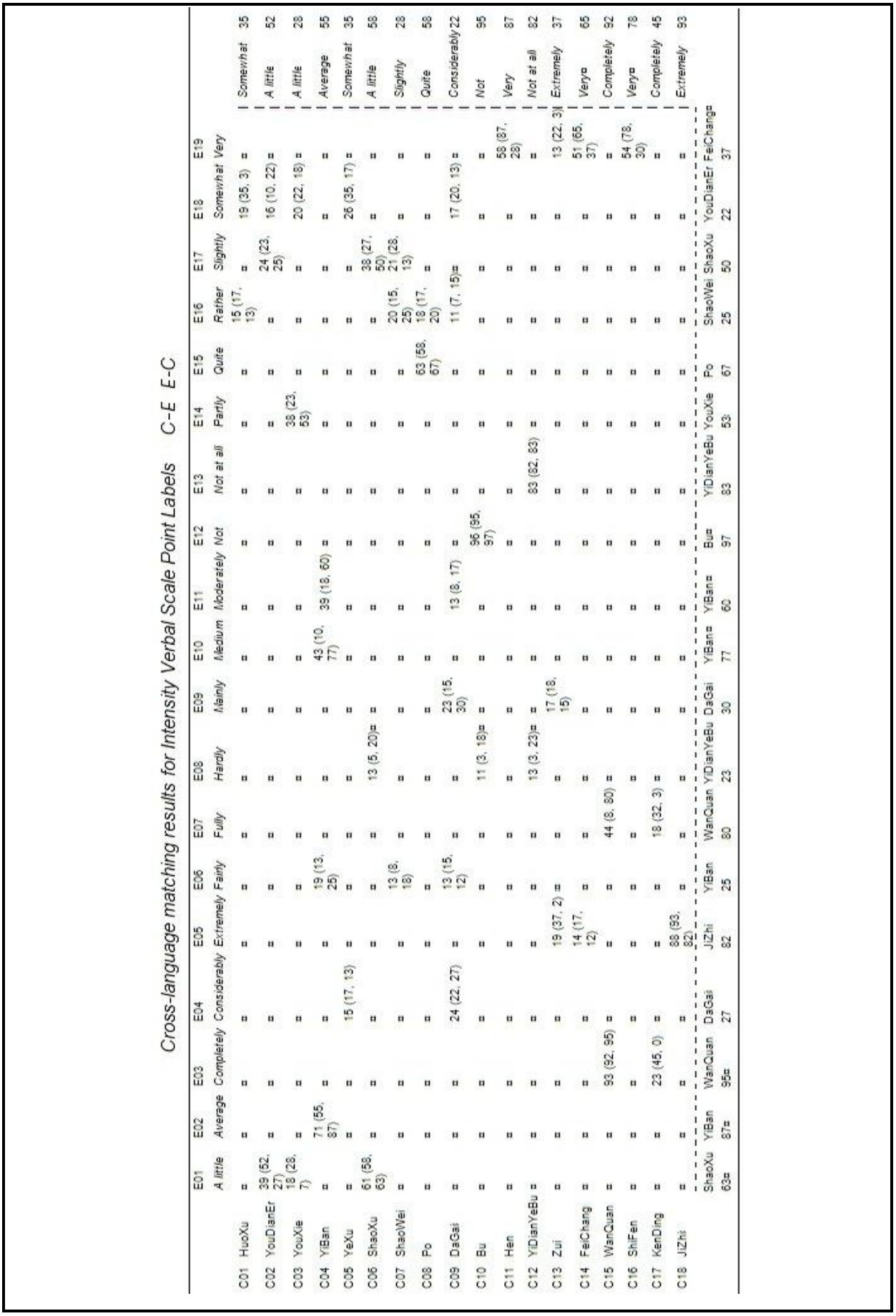
Regarding the qualifiers for the "Intensity" modality, ratings of 15 VSPLs differed between the two studies by less than 0.5, 14 differed by less than 1.0, and 18 out of 19 differed by less then 1.5 in absolute values. The largest difference was 2.6.

Of concern are the qualifiers differing by more than 1.0 point on the 11-point scale. These words include “not at all”, “partly”, “considerably”, “mainly”, and “quite” - some of which are frequently utilized for verbalized response tools.

However, altogether the NW scores from VQS-2 and VQS-3 correlate highly, r=0.96. Thus for quite a few common 5-point rating scales their cross-national consistency is likely to be satisfactory.

Box 5-3

Cross-language matching - Results for Intensity Verbal Scale Point Labels



Cross-language matching results for Intensity Verbal Scale Point Labels C-E E-C

		E01	E02	E03	E04	E05	E06	E07	E08	E09	E10	E11	E12	E13	E14	E15	E16	E17	E18	E19	
		A little	*Average*	*Completely*	*Considerably*	*Extremely*	*Fairly*	*Fully*	*Hardly*	*Mainly*	*Medium*	*Moderately*	*Not*	*Not at all*	*Partly*	*Quite*	*Rather*	*Slightly*	*Somewhat*	*Very*	*Somewhat* 35
C01	HuoXu	¤	¤	¤	¤	¤	¤	¤	¤	¤	¤	¤	¤	¤	¤	¤	15 (17, 13)	¤	19 (35, 3)	¤	
C02	YouDianEr	39 (52, 27)	¤	¤	¤	¤	¤	¤	¤	¤	¤	¤	¤	¤	¤	¤	¤	24 (23, 25)	16 (10, 22)	¤	*A little* 52
C03	YouXie	18 (28, 7)	¤	¤	¤	¤	¤	¤	¤	¤	¤	¤	¤	¤	38 (23, 53)	¤	¤	¤	20 (22, 18)	¤	*A little* 28
C04	YiBan	¤	71 (55, 87)	¤	¤	¤	19 (13, 25)	¤	¤	¤	43 (10, 77)	39 (18, 60)	¤	¤	¤	¤	¤	¤	¤	¤	*Average* 55
C05	YeXu	¤	¤	¤	15 (17, 13)	¤	¤	¤	¤	¤	¤	¤	¤	¤	¤	¤	¤	¤	26 (35, 17)	¤	*Somewhat* 35
C06	ShaoXu	61 (58, 63)	¤	¤	¤	¤	¤	¤	13 (5, 20)¤	¤	¤	¤	¤	¤	¤	¤	¤	38 (27, 50)	¤	¤	*A little* 58
C07	ShaoWei	¤	¤	¤	¤	¤	13 (8, 18)	¤	¤	¤	¤	¤	¤	¤	¤	¤	20 (15, 25)	21 (28, 13)	¤	¤	*Slightly* 28
C08	Po	¤	¤	¤	¤	¤	¤	¤	¤	¤	¤	¤	¤	¤	¤	63 (58, 67)	18 (17, 20)	¤	¤	¤	*Quite* 58
C09	DaGai	¤	¤	¤	24 (22, 27)	¤	13 (15, 12)	¤	¤	23 (15, 30)	¤	13 (8, 17)	¤	¤	¤	¤	11 (7, 15)¤	¤	17 (20, 13)	¤	*Considerably* 22
C10	Bu	¤	¤	¤	¤	¤	¤	¤	11 (3, 18)¤	¤	¤	¤	96 (95, 97)	¤	¤	¤	¤	¤	¤	¤	*Not* 95
C11	Hen	¤	¤	¤	¤	¤	¤	¤	¤	¤	¤	¤	¤	¤	¤	¤	¤	¤	¤	58 (87, 28)	*Very* 87
C12	YiDianYeBu	¤	¤	¤	¤	¤	¤	¤	13 (3, 23)¤	¤	¤	¤	¤	83 (82, 83)	¤	¤	¤	¤	¤	¤	*Not at all* 82
C13	Zui	¤	¤	¤	¤	19 (37, 2)	¤	¤	¤	17 (18, 15)	¤	¤	¤	¤	¤	¤	¤	¤	¤	13 (22, 3)	*Extremely* 37
C14	FeiChang	¤	¤	¤	¤	14 (17, 12)	¤	¤	¤	¤	¤	¤	¤	¤	¤	¤	¤	¤	¤	51 (65, 37)	*Very* 65
C15	WanQuan	¤	¤	93 (92, 95)	¤	¤	¤	44 (8, 80)	¤	¤	¤	¤	¤	¤	¤	¤	¤	¤	¤	¤	*Completely* 92
C16	ShiFen	¤	¤	¤	¤	¤	¤	¤	¤	¤	¤	¤	¤	¤	¤	¤	¤	¤	¤	54 (78, 30)	*Very* 78
C17	KenDing	¤	¤	23 (45, 0)	¤	¤	¤	18 (32, 3)	¤	¤	¤	¤	¤	¤	¤	¤	¤	¤	¤	¤	*Completely* 45
C18	JiZhi	¤	¤	¤	¤	88 (93, 82)	¤	¤	¤	¤	¤	¤	¤	¤	¤	¤	¤	¤	¤	¤	*Extremely* 93
		ShaoXu	YiBan	WanQuan	DaGai	JiZhi	YiBan	WanQuan	YiDianYeBu	DaGai	YiBan¤	YiBan¤	Bu¤	YiDianYeBu	YouXie	Po	ShaoWei	ShaoXu	YouDianEr	FeiChang¤	
		63¤	87¤	95¤	27	82	25	80	23	30	77	60	97	83	53	67	25	50	22	37	

5.6 *Assessment of the VQS-3 outcomes*

Validity deliberations

While the outcomes of Project VQS-3 are certainly significant and beneficial, because it is the first enterprize of this type, nevertheless validity constraints need to be considered. Firstly, the sample size was not very large, given the complexity of the conducted experiments. Secondly, not all participants had a high level of competence in English, their second language. Thirdly, only one of the Chinese languages could be covered - the sample in this study were Hong Kong Chinese who speak Cantonese, one of seven major dialects in China. This is somewhat different from Mandarin, the main language in China.

Nevertheless, altogether it seems that the principal findings are by and large valid, while results regarding particular words should not be generalized without further exploration.

Conclusions about recommended words/expressions for rating scales

The investigated VSPLs have been examined within four different experimental tasks. Following this approach, to achieve conclusions at first the "Number-for-Words" scores are considered; second, the "Word-for-Numbers" allocations; third, the "Familiarity" ratings; and fourth, the outcomes of the "Cross-language Matchings" trial. The recommended VSPLs for the Intensity, Frequency and Agreement-with-statements modality are shown in *Box 5-4*.

By the way, the conceptual rules for utilizing the results of VQS-3 were handled even stricter than in VQS-2, because of the bilingual context and decision-making.

Number-for-Words findings

The Number-for-Words task induces less scale context bias than the Word-for-Number task because no order-ranking is required and participants can assign the same scale value to different VSPL they found suitable. Therefore, the scaling of one VSPL does not affect the scaling of another VSPL.

As first step, the most suitable VSPLs for a 5-point scale were indicated. For this, out of the 11-point ratings, five equidistant points were defined, 0--2.5--5--7.5--10, and then VSPLs were identified which are closest in mean ratings to these five 'perfect' levels .

Next, it was examined whether the 95% confidence interval of the VSPL rating, i.e., mean +/- 2 sd, covered adjacent 'perfect' points. The rationale of this confidence interval criterion was to choose VSPLs with mean scale values that do not overlap with adjacent levels.

For example, a VSPL for level "2" should not overlap with level "1" or level "3". That is, for level "1", the upper bound of 95% CI of a VSPLs mean scale value should be smaller than 2.5; the 95% CI of level "2" VSPL should be within 0.1 to 4.9; the 95% CI of level "3" VSPL should be between 2.6 to 7.4; the 95% CI of level "4" VSPL should be within 5.1 to 9.9; and the lower bound of the 95% CI of level "5" VSPL should be larger than 7.5. Yet unfortunately this feature, used as *criterion 1,* is harder to match for levels "2" and "4" than the other three levels.

Box 5-4

VSPLs recommended for creating a 5-point scale

	Intensity modality				**Frequency** modality		**Agreement with statements** modality	
	Rationale A		*Rationale B*		*Rationale A*		*Rationale A*	
Scale level	Chinese	English	Chinese	English	Chinese	English	Chinese	English
1st	Bu	Not*	Bu *or* YiDianYe Bu	Not *or* Not at all	Wan Quen Mei You	Never	Jue Bu Tong Yi *or* FeiChang Bu Tong Yi	Fully disagree *or* Strongly disagree
2nd	Shao Xu	A little*	Shao Xu	A little	Hen Shao	Seldom*	Bu Tai Tong Yi	Partly disagree
3rd	Yi Ban	Average	*A* Yi Ban	Medium *or* Moderately	Jian Zhong	Sometimes*	Zhong Li	Neutral
4th	Hen	Very	Po	Quite	Duo	Often*	Tong Yi	Partly agree
5th	Wan Quen	Completely*	Fei Chang	Very	Fei Chang Duo	Always	Jue Dui Tong Yi *or* Fei Chang Tong Yi	Fully agree *or* Strongly agree

Notes: Rationale *A* refers to criterion 1 and criterion 2 outlined in the text; VSPLs which do not fully meet these are marked with "*". Rationale *B* is based on different anchors for scale levels (cf. text).

Using 0--2.5--5--7.5--10 as anchors sounds logical - however, for statistical reasons it is a thorny decision when designing scales how extreme an endpoint to choose. There is a risk: Extreme levels at the bottom or top of a scale may not be used very often (e.g., in questions such "how satisfied are you with …", or "how angry are you about …", etc), meaning that the 5-point response scale becomes practically a 4-point or 3-point one. An alternative rationale is to make 1--3--5--7--9 the target values for items calibrated on a 0-to-10 scale, this is *criterion 2*. These two concepts are the facets of rationale A.

Some verbal labels also induce linguistic trouble. A pertinent example is "average", the favoured VSPL for level "3" - - it is not a good label in language terms, because the point is not whether a rating is "average" in relation to other's judgments, it is to function as best-labelling a position in the middle between the endpoints. VSPLs like "medium" or possibly "moderately" provide this better.

The Intensity modality is the bar far most-used type of rating scales, either as standing-alone scale or added to substantive adjectives. Therefore, aiming at functionality, a second approach, "rationale B", was developed for intensity-scaling, giving more weight to the psycholinguistic considerations outlined above. The result can be found in Box 5-4 which, only for Intensity, contains suggestions based on both, rationale A and rationale B. For this

modality, and English VSPLs, rationale B appears to provide the better outcome in pragmatic terms.

Words-for-Number and Familiarity findings

The suggestions presented in Box 5-4 are roughly in line with data from these two experiments; not-fitting words/expressions were excluded.

Regarding English VSPLs, the suggestions for all three modalities are similar to the conclusions resulting from VQS-2.

Cross-language Matchings findings

When designing verbalized rating scales "pairing" of VSPL can work as a smart concept, especially for "Agreement" measures. Therefore, in addition to the appraisals above, suitable pairings were explored. A conceptual aim is to select pairs of VSPL that literally contain directly opposite meanings. For example, in the agreement-with-statements modality, “mainly agree” (M = 8.2 in the number-for-words task) was originally considered as a recommended VSPL for level "4" because it is closer to the perfect scale value (7.5) than “partly agree” (M = 6.7). For level "2" (the symmetry of level "4"), however, because our only recommendation was “partly disagree” (that met both the first and the second criterion), we resorted to preferring its mirror VSPL “partly agree” rather than “mainly agree” for level "4". The rationale for preferring literally opposite VSPLs between levels "1" and "5" and also between levels "2" and "4" is to design scales that *appear* to respondents as equidistant (see *Box 5-4*).

Final remarks

The results from this research confirm that in both languages, Chinese and English, words or expressions are reliably linked to the numerical levels of scales, and that thereby verbalized rating scales are valid. This is true for all three investigated modalities of making judgments - the *Intensity* or the *Frequency* of something, and *Agreement* with statements.

Furthermore, when the data for 61 Chinese and 44 English verbal scale point descriptors were connected, well-matching Chinese and English VSPLs could be identified. This is essential knowledge for the translation of surveys across the two languages.

The preferences of the respondents, i.e., which words/expressions were preferred as VSPLs for 5-point scales, are predominantly quite strong, especially with respect to the borders of a scale.

The psychometric findings of the project are not restricted to 5-point scales though - the data gained in the Number-for-Word experiment for the three scale modalities can be utilized for designing 4-point or 6-point or 7-point scales as well. Such scales are used by researchers who do not want to offer a mid-point, or need three levels on the upper and the lower part of their scale.

To sum up - prior studies have investigated the numerical values of VSPLs, but few studies have systematically applied the findings to constructing rating scales within experiments or surveys.

For researchers administering questionnaires in both Chinese and English, now a methodologically sound basis is available for developing VSPLs that are reasonably equivalent across languages.

Nonetheless, Project VQS-3 should be seen as the beginning rather than the end of *"Creating Chinese rating scales & testing Chinese-English linkages"* - further psychometric, socio-linguistic and ethnographic notions deserve on-going research.

<6> APPRAISAL AND ISSUES FOR FURTHER PSYCHOMETRIC RESEARCH

6.1 *Validity constraints*

Obviously the external validity of the findings from these three VQS projects must be restricted, as smallish non-random samples were employed, and the non-student groups are certainly too small. Furthermore, not all germane variations of context conditions could be realized in the experiments. 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 Verbal Scale Point Labels (VSPL) studied in this project.

Regarding internal validity, some participants 'struggled' to understand the instructions, especially for the magnitude scaling tasks in VQS-2, and the explanation of the familiarity task may have been phrased too indistinct; both is likely to have increased response dispersion, beyond what was expected.

Finally, there are epistemological issues to be considered. From a cognitive psychology or socio-linguistic 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 first project (VQS-1 in Germany, 1966, and repeated a decade later) 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 therefore these cognitions on average didn't change much over 10 years (cf. Rohrmann 1978).

Cross-cultural subject matters turned out to be more worrying. Almost all of the analyzed words/expressions have a 'soft' meaning, as the variance around the midpoint of the ratings indicates. In Project VQS-3 this was the case for both the English and the Chinese items, the best equivalences were not easily identified. Thus it was valuable that the design of Project-3 included cross-validation procedures.

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 which reflect how humans think and talk.

6.2 *Overall implications for designing rating scales*

The outcomes of this research can be utilized as a general rationale for the systematic construction of verbalized scales measuring psychological or sociological 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 particular scale;

(4) sufficient familiarity of the expression;

(5) reasonable likelihood of utilization when used in substantive research.

With respect to cross-cultural research in the social sciences, a further facet becomes relevant:

(6) availability of comparable VSPLs (re scale position) in the other language.

The rating scale at whole needs to be linguistically coherent in the employed words and easy to communicate to research participants.

As the results from VQS-1 and VQS-2 show, for both 5-point and 7-point scales fitting words/expressions can be found. In *Box 6-1*, data for 5-point scales from VQS-2 are presented.

Box 6-1

Best-suitable Verbal Scale Point Labels for designing 5-point rating scales

---

FREQUENCY

<1>	never (0.1)		
<2>	seldom (1.7)	rarely (1.3)	
<3>	sometimes (3.6)	occasionally (3.2)	*gap <3> to <4>*!
<4>	often (6.6)	frequently (7.4)	
<5>	always (10.0)	very often (8.3)	*gap <4> to <5>!*

INTENSITY

<1>	not at all (0.0)	not (0.4)	
<2>	slightly (2.5)	a little (2.5)	
<3>	fairly (5.3)	moderately (5.0) average (4.8)	
<4>	considerably (7.6)	quite a bit (6.5) mainly (6.8)	*not good: rather(5.8)*
<5>	extremely (9.6)	very much (8.7) very (7.9)	

PROBABILITY

<1>	certainly not (0.2)		
<2>	probably not (1.9)	unlikely (1.7)	
<3>	possibly (5.0)	about 50:50 (4.8)	
<4>	probably (6.8)	quite likely (7.4) likely (6.9)	
<5>	certainly (9.6)	for sure (9.8)	

QUALITY

<1>	bad (1.0)	poor (1.5)	
<2>	inadequate (1.9)	unsatisfactory (1.8)	*no good word available*
<3>	fair (5.2)	medium (5.0) average (4.9)	
<4>	good (7.2)		
<5>	excellent (9.7)	very good (8.5)	

AGREEMENT

<1>	fully disagree (0.2)	strongly disagree (0.4)	
<2>	mainly disagree (2.4)	somewhat disagree (3.2)	*not good: disagree (1.6)*
<3>	neutral (4.9)	undecided (4.8)	
<4>	mainly agree (7.4)	somewhat agree (6.4)	*not good: agree (8.1)*
<5>	fully agree (9.7)	strongly agree (9.6)	

Based on these data, the best options for an English 5-point scale seem to be (as was outlined in chapter 4):

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 satisfactory 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 the above suggestions had been cross-checked in a seminar with a dozen psychologists.

To select optimal words/expressions for the two endpoints of a rating scale is always a difficult deecision, especially for 5-point scales. It depends on how extreme the two anchors are meant to be, and what the typical response pattern in a country are. Very 'far-out' labels are unlikely to be chosen be respondents, leading to unused scale levels. In chapter 4 (section 4.4), this issue is contemplated for the English language.

Here again a 'multi-modal' approach for the to be created rating scales is functional, that is, to utilize words *and* numbers *and* symbols *and* visual/graphical means in order to achieve the best-feasible psychometric quality.

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

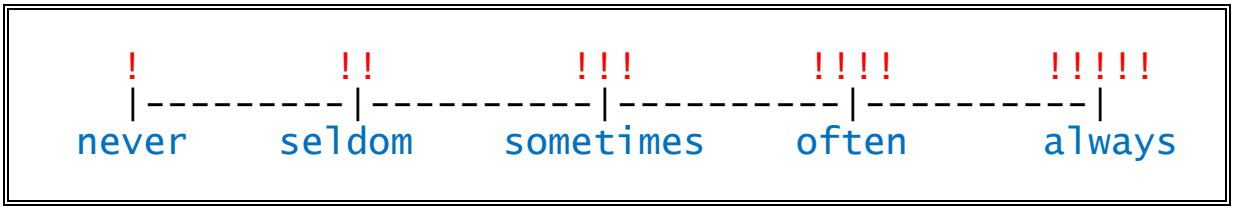
Numerical: 1/2/3/4/5 or -2/-1/0/+1/+2 or --/-/0/+/++;

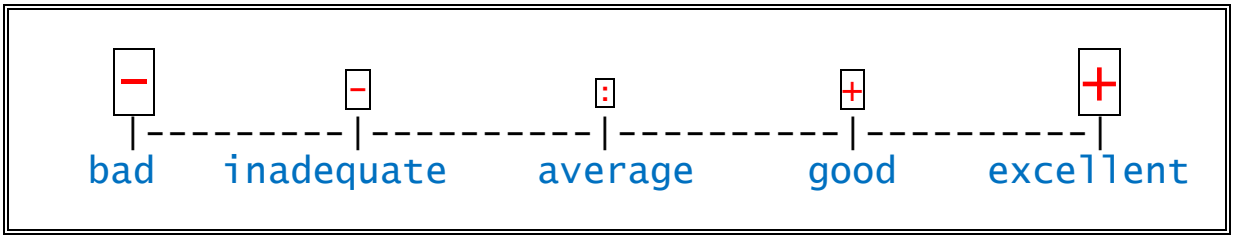
Visual/graphical means: equidistant frames, or scaled lines, or !/!!/!!/!!!/!!!!, or sympols in different size, or faces, etc.

A few examples are shown in *Box 6-2.*

Box 6-2

Examples of multi-modal rating scales for frequency and quality





The relevance of the multi-mode design has been explored in 'meta-interviews' with interviewers who were prepared to show their interviewees different types of scales; this was done in VQS-1 (German language tools) and VQS-2 (English language tools). The results about the usability and utility of verbalized rating scales provided evidence that rating scales to be used in experiments or surveys should be designed as a combination of numbers and words. The chosen layout needs to be adapted to the questionnaire mode, e.g., in printed/mailed or internet 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 may be an easy mode, yet in personal communication situations (see Karelitz & Budescu 2004) again words may be the preferred manner. In web-based surveys participants simply need to tick a box.
Of course any newly constructed rating scale should be pretested for comprehensibility and acceptability with relevant target groups.

6.3 *Directions for on-going research*

To widen the validity scope of the presented projects, further research is indispensable. Recent reviews (e.g., Jensen et al. 2011, Vaerenbergh & Thomas 2013) point at quite a few issues which are not yet sufficiently clarified, as did the critical evaluation of the VQS outcomes. Various lines of on-going enquiries should be deliberated.

One issue is stability over time. This was task (4) "Has the subjective interpretation of frequency and intensity expressions shifted over time?" - - so far only realized in VQS-1, 35 years ago (!), for German VSPLs. A further repetition is strongly recommended. (By the way, this was planned for 2006, but unfortunately no suitable research partner in Germany could be found).

For the extremely common verbalized rating scales in English language (especially those used in USA) almost no retests exist, and consequently this appears as a worthwhile venture as well.

A further issue are direct language comparisons, as realized in VQS-3 for VSPLs in English and Chinese language. This is intended for German and English language (potential Project VQS-4) but not yet instigated. It should be focussed on VSPLs which have been regularly utilized in rating scales

The findings would 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 fear of crime or noise annoyance or residential satisfaction in a community). However, comparisons between countries can only be valid if the quantitative meaning of the utilized response scale and especially the top-end item word - e.g., "very", "sehr", "tres", is sufficiently similar.

A third issue is to cover further languages. It seems that so far there are no explorations of the words/expression used in rating scales for the Arabic or Slavic languages. Even within

widely used languages, such as English or Chinese, cultural differences are given. Hence diverse national types of English could be compared, such as, English, American and Australian English, or Mandarin versus Cantonese Chinese in China..

In case of further research, sampling is an important concern. In order to clarify whether the interpretation of qualifiers is consistent across different levels of age and education, much larger samples are indicated. Within multi-cultural societies it is a topic whether findings for natural speakers are valid for people for whom English or German is the second language. A special complexity is bilingual competence. In Hong Kong many people grow up bilingually - which actually means that for them neither Chinese nor English verbalizations have a high familiarity.

The outlined in-depth 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.

To sum up this report, "designing verbalized rating scales", and to do so based on proficient research, is still a valuable enterprize - social science surveys are thriving and addressing wider communities than ever.

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.

Au, W., Rohrmann, B., Au, W. & Taylor, P., Ho, J.M.,Yeung, S. (2011). Developing equivalent Chinese and English scale point labels for rating scales used in survey research. Asian Journal of Social Psychology, 14, 91-111.

Auer, S., Hampel, H., Moeller, H.-J., & Reisberg, B. (2000). Translations of measurements and scales: Opportunities and diversities. International Psychogeriatrics, 12 (Suppl. 1), 391-394.

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

Babbie, E. (2011). The basics of social research. Londonon: Sage.

Barilli, E., Savadori, L., Pighin, S. (2010). From chance to choice: The use of a verbal analogy in the communication of risk. Health, Risk & Society, 12, 546-559.

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.

Bryman, A. (2012). Social research methods. Oxford, Oxford Univ. Press.

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.

Clauss, G. (1968). Zur Methodik von Schaetzskalen in der empirischen Forschung. Probleme und Ergebnisse der Psychologie, 26, 7-52.

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.

DFG-Forschungsbericht 1974. Fluglaermwirkungen - Eine interdisziplinaere Untersuchung ueber die Auswirkungen des Fluglaerms auf den Menschen (2 Baende). Boppard.

Diefenbach, M. A., Weinstein, N. D. and O'Reilly, J. (1993) Scales for assessing perceptions of health hazard susceptibility. Health Education Research, 8, 181–92.

Dillman, D. (2007). Mail and Internet surveys: The tailored design method (2nd ed., updated). New Jersey: 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.

Duden 1959. Der Grosse Duden Bd. 1, Rechtschreibung. Weinheim.

Duden 1964 & 1972. Der Grosse Duden Bd. 8, Sinn- und sachverwandte Woerter und Wendungen. Mannheim.

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.

Friedrichs, J. (1973). Methoden empirischer Sozialforschung (Kap 6: Skalierungsverfahren). Reinbek: Rowohlt.

Gerring, J. (2011). Social science methodology - A unified framework. Cambridge: Cambridge Univ. Press.

Guilford, J.P. (1954). Psychometric methods. New York: Wiley.

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.

Hennig, W. (1975). Schaetzskalen. In Friedrich, W., & Henning, W, Der sozial-wissenschaftliche Forschungsprozess (pp 345-367). Berlin: Humboldt:

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. (1967). Psychologie der Sprache. Berlin: Springer.

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.

Hofstaetter, P.R., & Wendt, D. (1966). Quantitative Methoden der Psychologie. Muenchen: Barth.

Hofstede, G.H. (2001). Culture's consequences: Comparing values, behaviors, institutions and organizations across nations. Hilldsdale/NJ, Erlbaum.

Howard, J. (2002). A student handbook for Chinese function words. Hong Kong: The Chinese University Press.

Irle, M, & Rohrmann, B. (1968). Gesamtbericht über die Hamburger Voruntersuchung zum DFG-Projekt Fluglärmforschung - Sektion Sozialpsychologie. Report Mannheim Universität.

Jensen Hjermstad, M., Fayers, P.M., Haugen, D., Caraceni, A., Hanks, G.W., Loge, J.H., Fainsinger, R. (2011). Studies comparing numerical rating scales, verbal rating scales, and visual analogue scales for assessment of pain intensity in adults: A systematic literature review. Journal of Pain and Symptom Management, 41, 1073–1093.

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

Karelitz, T. M., & Budescu, D. V. (2004). You say "probable" and I say "likely": Improving interpersonal communication with verbal probability phrases. Journal of Experimental Psychology: Applied, 10, 25-41.

Kerlinger, F. N., & Lee, H. B. (2000). Foundations of behavioural research. 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.

Koenig, R. (Ed.) (1965). Das Interview. Koeln: Springer.

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

Kristof, W. (1966). Das Cliff'sche Gesetz im Deutschen. Psychologische Forschung, 29, 22-31.

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

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

Lanier, M.M., Ford, C.A., Reid, J.C., & Strickland, K.M. (Eds.) (2014). Advanced research methods for the social sciences. Cognella Academic Publishing.

Lau, L. Y. & Ranyard, R. (1999). Chinese and English speakers' linguistic expression of probability and probabilistic thinking. Journal of Cross-Cultural Psychology, 30, 411-421.

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 attitude. Archives of Psychology, 140, 1-55.

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

McColl, D. & Fucci, D. (2006). Measurement of speech disfluency through magnitude estimation and interval scaling. Perception & Motor Skills, 102, 454-460.

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.

Miller, G. A. (1956). The magical number seven, plus or minus two - some limits on our capacity for processing information. Psychological Review, 63, 81-97.

Montello, D.R., Sutton, P. (2013). An introduction to scientific research methods in geography and environmental studies. 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. (2000). Communicating quantities: A review of psycholinguistic evidence of how expressions determine perspectives. Applied Cognitive Psychology, 14, 237-255.

Myers, K., & Winters, N. C. (2002). Ten-year review of rating scales. I: Overview of scale functioning, psychometric properties and selection. Journal of the American Academy of Child & Adolescent Psychiatry, 41, 114-122.

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). Survey Pretesting: Do different methods produce different results? Sociological Methodology, 24, 73-104.

Preston, C.C., & Colman, A.M. (2000). Optimal number of response categories in rating scales: Reliability, validity, discriminating power and respondent prerences. Acta Psychologica, 104, 1-15.

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.

Robson. C. (2011). Real world research: A resource for users of social research methods in applied settings. Chichester, Wiley.

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

Rohrmann, B. (1974). DFG-Forschungsbericht "Fluglaermwirkungen - Eine interdisziplinaere Untersuchung ueber die Auswirkungen des Fluglaerms auf den Menschen", Kurzbericht (Band 3). Boppard.

Rohrmann, B. (1978). Empirische Studien zur Entwicklung von Antwortskalen fuer die sozialwissenschaftliche Forschung [Empirical studies to develop standardized rating scales for the social sciences]. Zeitschrift fuer Sozialpsychologie, 9, 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.

Rohrmann, B. (2000). Risk perception research - Review and documentation. Research Center Juelich: RC Studies, Report #68.

Rohrmann, B. (2007). Verbal qualifiers for rating scales: Sociolinguistic considerations and psychometric data. Project Report, University of Melbourne/Australia.
Available on website RohrmannResearch.net

Rohrmann, B., Au, W. & Taylor, P. (2008). Investigating Chinese and English scale point labels for verbalized rating scales in survey research. Project Report, University of Melbourne/Australia.
Available on website RohrmannResearch.net

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

Sauro, J. (2010). Should you use 5 or 7 point scales? Denver, USA: Measuring Usability Newsletter.

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.

Sixtl, F. (1967). Messmethoden in der Psychologie. Weinheim: Hogrefe.

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.

Vaerenbergh, Y.v., Thomas, T. (2013). Response styles in survey research: A literature review of antecedents, consequences and remedies. International Journal of Public Opinion Research, 25, 195-217.

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. 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.

Xu, J. H. & Li, S. (2007). An exploratory research on the numerical translation of Chinese verbal probabilistic expression. Chinese Journal of Ergonomics, 12, 15-19.

Yano, T., Masden, K. & Kawai, K. (1998). A survey on Japanese and English Descriptors of Annoyance. N. Carter & R. F. S. Job, 7th International Congress on Noise as a Public Health Problem, 519-522, Sydney, Noise Effects '98.

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.