Investigating Chinese and English scale point labels for verbalized rating scales in survey research

Research report by

Bernd Rohrmann, Wing-tung Au & Paul Taylor

(Univ. of Melbourne/Australia) (Chinese University of HongKong)

July 2008

Abstract

Rating scales are the most-used response tool in surveys, and the scale levels are commonly described with words (verbal scale point level, "VSPL"). In this study, Chinese VSPLs for questionnaires were identified which employ five-point scales that are psychometrically equivalent to English VSPLs. In several bi-lingual experiments, altogether 61 Chinese and 44 English items were tested which cover three rating modalities: intensity, frequency and agreement with statements. For each VSPL three aspects were measured: position between minimum and maximum, familiarity and appeal. The correspondence between pertinent Chinese and English words was also assessed. Based on these data the best-suited VSPLs are recommended. The findings have significant practical implications for the translation of scales between the two languages.

*** Preface ***

This research belongs to my project "Verbal qualifiers for rating scales: A cross-cultural study [VQS]". It is based on my long-time work about verbal tools used for developing rating scales. The primary study was conducted in Germany and published in 1978.

When at Melbourne University, the original study was repeated and extended, now conducted in English language (1996-2000). A pertinent report, "Verbal qualifiers for rating scales: Sociolinguistic considerations and psychometric data" was placed on this website; a publication is in prep. The report outlines the philosophy and methodology of Project VQS.

The next expansion aimed at an additional language, Chinese, and cross-language comparisons. I realized this together with Assoc-Prof Au and Assoc-Prof Taylor at the Chinese University of Hong Kong, beginning in 2002. The text presented here describes the procedures and outcomes of the experiments conducted in Hong Kong in reasonable detail. A publication focusing on an Asian audience will be created under the direction of Prof Au and in 2009 submitted to the Asian Journal of Social Psychology.

Those interested in more information about the outcomes of Project VQS are welcome to contact me.

Contact Address:

Professor B. ROHRMANN
Roman Research Road, 94 Fenwick St, Carlton-Nth, Vic 3054, AUSTRALIA

E-Mail mail@rohrmannresearch.net

1 The issue: Words in rating scales

1.1 Research field

Studies about human perception, knowledge, thinking, attitudes and behavior have always included to ask people about such issues, and thereby obtaining scientific insights. Of course words are the crucial and indispensable tool - at first within qualitative procedures, then increasingly in quantitative approaches. Academic and applied research across a wide range of social science fields - such as psychology, sociology, education, communication studies - still rely heavily on questionnaires for collecting data (Myers & Winters 2002, Rohrmann 2003). Respondents are asked to provide either open-ended or close-ended answers, and the most frequent use of close-ended response mode is the category rating scale (e.g., “never / seldom / sometimes / often”). Although a handful of previous studies have demonstrated that asking respondents to provide magnitude estimates (e.g., presenting to respondents a light of certain intensity (as an anchor) and assign it an arbitrary scale value of 10, and then ask respondents to rate the brightness of other lights in reference to this anchor) provide both better quality of data and better theoretical support of measurement than category rating scales, the latter remains a dominating scaling method because it is less demanding and time-consuming for both researchers and respondents (e.g., Levine 1994, Lodge & Tursky 1979, Purdy & Pavlovic 1992, Schaeffer & Bradburn 1989, Wegener 1983, Wills & Moore 1994).

The popularity of category rating scales has been demonstrated by a wide range of questionnaires across fields and research method textbooks (e.g., Aiken 1997, Babbie 1989, Dillman 2007, Foddy 1992, Kerlinger & Lee 2000, Krosnick 1999, Krosnick & Fabrigar 1998, Miller 1991, Oppenheim 1992, Sapsford 2007, Schuman 1996, Vaus 2002).

1.2 Category rating scales

Category rating scales usually offer from four to eleven response alternatives, that is, equidistant ordinal scale points (Rohrmann, 2003). Numbers, words or visual/graphic symbols are used to signify the categories. Verbal labeling of rating scales is the most frequently used format (e.g., “totally agree / agree / neutral / disagree / totally disagree”) for qualifying numeric rating scales. These words function as a *verbal scale point label* (in the following abbreviated as "VSPL").

Verbal labeling may be expressed in a variety of formats, e.g., single words or short expressions (e.g., “never / seldom / sometimes / often / always”, “not / slightly / fairly / quite / very”, “bad / poor / fair / good / excellent”, “strongly-disagree / disagree / undecided / agree / strongly-agree”). Sometimes words are used only in the scale endpoints (e.g., from “not-at-all” to “extremely” or from “never” to “always” for a 0 to 10 scale). The most frequently used format is the combination of words that describe a certain attribute or behavior (e.g., “agree” or “excellent”) and those that indicate the diverse level of a dimension (e.g., “never / seldom / sometimes / often / always”).

Events or issues can be scaled in at least three modalities: their intensity or strength, their

frequency, and their probability. Further commonly used rating scale modes are: quality levels, agreement with statements, right/wrong grades.

1.3 Issues concerning verbal labeling

Verbal labeling of numeric response scales provides many advantages. For example, verbal labels are easy to explain and familiar to respondents, and they help capture normative judgment (Rohrmann 2003). However, they should be used with caution so as not to influence the perceptions and answers of the respondents in unintended ways. Karelitz & Budescu (2004) stated that the nature of the qualifiers make verbal scaling susceptible to methodological flaws because of the numerous verbal lexicons and the inconsistent meanings held across individuals.

People have a broad vocabulary bank of verbal qualifiers to choose from when describing different probabilities and certainties, but they develop preferences for some words and tend to use those but not others. Thus individuals with different preference profiles may interpret the intensities of the verbal labels differently. For example, in the study of Budescu, Weinberg & Wallsten (1988), a total of 111 different probability phrases were generated by the 20 participants when they were asked to verbally described a probability display. Similarly, in a review of rating scales employed by researchers in a set of 25 published articles, Karelitz & Budescu (2004) reported that more than 100 distinguished probability phrases were employed. As people may be used to a specific set of probability words or phrases, there may be variations in interpretation of the words.

From a within-subject perspective, the meanings of the verbal labels held by individuals are relatively consistent and reliable over time. From a between-subject perspective, however, different people would have different perception of meanings of the same verbal labels (Xu & Li 2007, Karelitz & Budescu 2004, Budescu & Wallsten 1985, Clarke, Ruffin & Beamen 1992, Johnson 1973, Mullet & Rivet 1991, Reagan, Mosteller, & Youtz 1989). These researchers required participants to convert perceived intensities of verbal labels into numerical probabilities, or to rank and compare labels (Renooija & Witteman, 1999). E.g., Rohrmann (1998) found that, in denoting the intensity of an annoyance scale, respondents differ substantially in assigning suitable verbal labels. Their ratings of the intensities of the verbal labels like “rather”, “quite a bit”, “fairly” and “hardly” deviate across individuals, as reflected by large standard deviations. Together, these studies suggest the absence of universally-perceived intensities of verbal labels for scale points across respondents.

Although the development of standardized verbal scale point labels (VSPLs) is crucial in designing questionnaires, it has received little concern among researchers both for the root words or phrases within the stem of each item (e.g., “How difficult do you find....?”) and the adverbs that indicate the extent or frequency perceived by each respondent (e.g., “Very difficult, somewhat difficult...”) (Budescu & Wallsten 1994, Diefenbach, Weinstein & O’Reilly 1993, Rohrmann 1978, Theil 2002, Windschitl & Wells 1996, Wright, Gaskell &

O'Muircheartaigh 1994). Furthermore, cultural differences in the focus of rating scale anchors (e.g., Chinese tend to focus on the middle scale points whereas Westerners pay more attention to the extreme anchors) may also be considered in deciding the VSPL, or otherwise, the advantages of using the VSPL diminishes (e.g., Auer, Hampel, Moller & Reisberg 2000, Chen, Lee & Steveson 1995, Reid 1995, Van de Vijver 2001, Van de Vijver & Leung 1997, Weinfurt & Moghaddam 2001).

1.4 Translation of scales

A related concern in developing VSPLs is the translation of scales in cross-cultural studies. Studies often relied on scales to study constructs or variables on the target population (Chen, Lee, Stevenson 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, Hampel, Moeller, Reisberg (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., Moxey & Sanford 1993, Presser & Blair 1994, Schwarz, Knauper, Hippler, Noelle-Neumann & Clark 1991). This further renders the cross-cultural comparability of rating scales more difficult, as it may be difficult to develop correspondence phrases or expressions between languages such as Chinese and English, which are the two languages focused upon in this study.

Past research have 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 54.99%, 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.

1.6 Application of psychometric scale features

As mentioned earlier, possible differences in how VSPLs are perceived across individuals, both within and between cultures, has 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, Guski, & Schuemer 1998, Guski, Felscher-Suhr & Schuemer 1998, see also Rohrmann 1998). It is essential to design scale anchors very carefully if equidistant and

unambiguous instruments are to be achieved, and this effort requires psychometric data on scale labels. Pioneering work has been started on scaling English VSPLs (Rohrmann 2003). This was based on previous experiments in which German scale expressions were quantified (first in 1966, then repeated in 1976; cf. Rohrmann 1978); the present study builds on that work. Four or five modalities were investigated, including intensity, frequency, probability, quality levels, and agreement with statements.

1.6 Research aims

The objectives of this research were to study Chinese verbal scale point labels, and to extend our knowledge about English ones. This includes:

- Scaling commonly used Chinese and English VSPLs regarding their position between minimum and maximum and their familiarity,
- exploring which words/expressions are preferred as VSPLs,
- identifying psychometrically equivalent VSPLs between the English and Chinese languages,
- covering main rating scale modalities, i.e.: intensity, frequency and agreement with statements,
- recommending best-suitable VSPLs for rating scale design.

The results shall provide researchers administering questionnaires in both Chinese and English a more methodologically sound basis for developing VSPLs that are equivalent across languages.

2 Project methodology

2.1 Experimental design - overview

A set of experiments was conducted. 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.

Three of the five rating scale modalities investigated in Australia (Rohrmann 2003) were selected: intensity, frequency and agreement with statements, because these are the most-used ones.

Chinese and English VSPLs were either presented separately or jointly, depending on the scaling task.

As respondents both a sample of students and a sample from the general public were chosen. All participants were bilingual.

Chinese VSPLs were developed by first generating a long-list of Chinese terms that might indicate varying levels of intensity, frequency and agreement with statement.

This list was later shortened, and a scaling study was conducted on the short-list. Finally, a cross-language, word-matching study was conducted in order to identify matching English terms. These procedures are described in greater detail below.

2.2 Generation of Chinese qualifiers

The first step in identifying the list of verbal scale point labels was an exhaustive search of Chinese qualifiers that could possibly be used in the three types of attribute scales of interest (intensity, frequency and agreement with statement. Lists of qualifiers were assembled from each of four sources: (1) books on Chinese function words, (2) prior studies using rating scales in Chinese (from Chinese psychological journals and student theses available at the Chinese University of Hong Kong), (3) Chinese translation of English qualifiers identified from previous research, and (4) brainstorming by two Chinese authors (Au and Ho). These sources are described in *Appendix A*. This initial process resulted in 238 agreement, 218 intensity, and 160 frequency qualifiers.

2.3 Preliminary study to short-list Chinese qualifiers

In order to reduce the lists of Chinese qualifiers to manageable sizes of approximately 20 words for each modality for the scaling study, we first conducted a questionnaire-based study with the aim of identifying qualifiers for each of the three attribute domains that had a reasonably high level of familiarity, that covered the full range of the rating scale, and that were linguistically compatible with use in typical rating scales. Methodology of the short-listing study and the results are presented in *Appendix B*. In sum, the preliminary short-listing study resulted in 24 agreement, 19 frequency, and 18 intensity Chinese qualifiers to be used in the scaling study.

2.4 Scaling study of Chinese and English qualifiers

The primary aim of scaling both Chinese and English qualifiers was achieved through a scaling study with several experimental sessions. The preliminary study described earlier yielded the Chinese qualifiers used in this study, while lists of English qualifiers (19 intensity, 12 frequency and 13 agreement with statements) were extracted from Rohrmann (2003).

2.5 Participants

The scaling study was administered through a computer program to a sample of students and a sample from the general public.

Student sample

109 students from The Chinese University of Hong Kong participated in a one-hour experiment. All participants were bilinguals who can read both Chinese and English. Twenty-one students participated in the experiment as a partial requirement of an introductory psychology course, while the others received HK$50 (US$6.41) each for their participation in the experiment.

General public sample

A convenience snowball sample of 191 participants recruited through the authors' personal network participated in the experiment. All participants had worked for at least one

year and attained a minimum level of secondary school education. Their ages ranged from 22 to 60 with a mean of 27 years old. About 87% received tertiary education, and the female/male proportion was even (51:49%). All general public participants received HK$ 50 (US$ 6.40) for completing the experiment for one set of qualifiers.

2.6 Design & procedure

Student sample

Approximately 20 persons participated in each experimental session for the data collection of the student sample. The experiment was conducted in a large computer room in which each participant sat in front of a computer. Upon arrival at the laboratory, participants signed an informed consent form that described the study as a word rating experiment. The tasks associated with each of the three sets of qualifiers were the same except for the list of words presented. We first collected data from 60 participants for the intensity qualifiers in one-hour experimental sessions. We found that all participants were able to complete the entire experiment within 30 minutes. In order to facilitate further data collection, we administered both the agreement and frequency qualifier tasks in the same one-hour experimental session to the remaining 60 participants, having no reason to expect interference in the scaling qualifiers of different modalities. In summary, 60 students participated in the intensity experiment, and another 60 students participated in both the agreement and frequency studies.

General public sample

Participants downloaded the computer program to do the experiment at home at their own pace. Participants were told that the study would take approximately 30 minutes to complete. No time limit was enforced. Feedback from participants indicated that most had finished the study within 30 minutes. After participants emailed their output file back to the researcher, payment and a debriefing form were mailed to them.

2.7 Computer program for running the experiments

The computer program was set up for conducting four tasks in which participants evaluated both the Chinese and English qualifiers: (1) a familiarity rating task; (2) a “number-for-words” task; (3) a “word-for-numbers” task; and (4) a cross-language matching task. These four tasks are described below.

2.8 Familiarity task

The purpose of the Familiarity Task was to evaluate participants’ familiarity with each qualifier. Participants rated familiarity of each verbal qualifier on an 11-point scale ranging from 0 = “Extremely unfamiliar” to 10 “Extremely familiar,” similar to the preliminary short-listing study described earlier. The Chinese and English qualifiers were presented together in a randomized order on the computer screen.

Participants indicated the familiarity rating by using the mouse to click on one of the 11 (0 to 10) radio buttons. The top left panel of *Figure 1* shows this familiarity rating task for scaling both Chinese and English frequency qualifiers.

2.9 Number-for-Words task

The second task presented to participants was a category scaling task, which we refer to here as a “number-for-words” task. The purpose of this task was to assess the scale value of the qualifiers using Thurstone’s method of equal appearing intervals. Each list of qualifiers for a particular modality (e.g., intensity) was shown at the bottom of the screen in a randomized order in a four-rows by ten-columns matrix. Each qualifier was contained in a small rectangle, appearing as a card that could be dragged with the mouse, similar to the functioning of a computer card game. A horizontal scale ranging from 0 to 10 of equal appearing intervals was presented in the vertical center of the screen.

For the intensity modality, the two endpoints were “Extremely low intensity” and “Extremely high intensity”. For the frequency modality, the two endpoints were “Extremely low frequency” and “Extremely high frequency”. For the "agreement with statements" modality, the 0 and 10 endpoints were marked underneath by “Extremely low level of agreement” and “Extremely high level of agreement”.

Participants were instructed to drag each of the qualifiers, using the mouse, to one of the eleven categories (slots) according to the perceived scale values of the words. Participants were free to categorize (i.e., drag and drop) qualifiers in any sequence, and no limit was set on how many qualifiers could be assigned to any single scale value. Participants were able to change their choices while categorizing other words, for words placed in one slot were able to be dragged to another slot. Participants did not need to fill up all slots but each word had to be put into one of the slots. When more than one word was dragged to the same one slot, these words would be aligned on top of each other without overlapping. The vertical positions of the words (in the same slot) had no implication on the scale value of the words. The task would be completed after all words were assigned into categories and the participant clicked on a “Confirm” button. The top right panel of *Figure 1* shows this number-for-words rating task for scaling English frequency qualifiers.

Experimental conditions: In order to counter-balance the order of presentation of the Chinese versus English qualifiers for this and the remaining tasks, and to check for language order effects, we implemented four between-subject experimental conditions. For each modality domain, between 14 and 16 participants were assigned to each of the four conditions. In Conditions 1 and 2 the participants performed the number-for-words task for the Chinese and English words separately: Those in Condition 1 completed the Chinese qualifiers first, and those in Condition 2 completed the English qualifiers first. In both Conditions 3 and 4, participants completed the number-for-words task with both the Chinese and English words presented together as one single categorization task.

2.10 Word-for-Numbers task

The third task, which we refer to as the "word-for-numbers" task, required each participant to identify the single, *best* qualifier for each of the five levels on a five-point rating scale. This task was similar to the previous number-for-words task with the exception that: (a) the horizontal scale presented in the middle of the screen had only five slots (rather than 11), and (b) only one qualifier could be assigned to each of the five slots. Participants were instructed to construct the most appropriate five-point scale while maintaining equal intervals between the five points. The task was completed after one word was assigned to each of the five categories and the participant clicked on a "Confirm" button.

All participants performed this task on the Chinese and English qualifiers separately. In Conditions 1 and 3, participants categorized the Chinese qualifiers first, while the English words were completed first in Conditions 2 and 4. The bottom left panel of *Figure 1* shows this word-for-numbers task for scaling English frequency qualifiers.

2.11 Cross-language matching task

The fourth task was a cross-language matching task. The purpose of this task was to identify, for each Chinese and English qualifier, the best (English or Chinese) translation among the list of qualifiers in the other language. Similar to the previous tasks, the Chinese (or English) qualifiers were presented as cards at the bottom of the screen. In the middle of the screen, the qualifier of the other language (the target word) was shown, one by one. The participants were asked to identify one matching word underneath that had the most similar meaning as the target word by dragging the matching word to an empty box beside the target word shown. The participants could change their decisions by dragging another matching word to the box to replace the prior one before they clicked the "Confirm" button for each target word. The participants matched a target word one at a time until the whole list of qualifiers in that language was completed. Afterwards, the participants did the same task again with the language of the target word and the matching words swapped. In experimental Conditions 1 and 3, Chinese qualifiers were presented as targets first; and English qualifiers were presented first in Conditions 2 and 4. The bottom right panel of *Figure 1* shows this cross-language matching task for matching English to Chinese frequency qualifiers.

In sum, participants completed four tasks for either the intensity qualifiers or for both frequency and agreement qualifiers. Altogether, there were 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. As summarized in *Table 1*, the familiarity task was the same across different conditions, having both Chinese and English qualifiers for the familiarity task at the same time. In the second (number-for-words) task, participants in Condition 1 were presented with the Chinese qualifiers first; in Condition 2, then English qualifiers were presented first; and in Conditions 3 and 4 the Chinese and English qualifiers were presented together. In both the

word-for-numbers and cross-language matching tasks, Chinese qualifiers were presented first in Conditions 1 and 3; while English qualifiers were presented first in Conditions 2 and 4.

3 Hong Kong results

3.1 Pre-analyses

Prior to considering the substantive analyses, we compared results for (1) the student and general public samples, and (2) order-of-language conditions of the experiments. *Appendix C* presents the results of these initial analyses. In sum, we found only negligible differences between the student and the general public samples and between the order-of-language conditions. Thus we aggregated data across the 2 samples and the 4 conditions in the remaining analyses.

3.2 Familiarity

Familiarity rating means and standard deviations are presented in *Tables 2*, *3* and *4* for intensity, frequency and.

Regarding intensity VSPLs, the range of the 18 Chinese qualifiers was 4.9 to 8.9 with an average of 7.1, whereas that of the 19 English words was 3.8 and 8.8 with an average of 6.0. A low familiarity intensity word was “somewhat”. For frequency words, the range of the 19 Chinese qualifiers was 5.3 to 8.2 with an average of 7.1, whereas that of the 12 English words was 3.1 and 8.1 with an average of 6.1. Less well familiar were “moderately” and “fairly often”. Regarding agreement-with-statement items, familiarity ratings were generally higher for Chinese than English words among these Chinese participants. The range of the familiarity ratings of the 24 Chinese items was 5.3 to 8.5 with an average of 6.6, whereas that of the 13 English labels was 3.9 and 6.9 with an average of 5.4. The most unfamiliar VSPL (below 4 on a 0 to 10 point scale) was “mainly disagree”.

Across all three modalities, there were no Chinese qualifiers that had familiarity ratings below 4. In general, we considered that most of these Chinese and English qualifiers were commonly used by this sample of bilingual Chinese respondents in their daily written usage.

Compared with a similar study conducted with English-as-a-first language participants in Australia (Rohrmann, 2003), familiarity ratings of participants in the present study were generally lower. We compared the familiarity ratings of English words in our studies with those in Rohrmann (2003) and found that the correlations across all the three modalities were higher than 0.6 suggesting a reasonable congruence.

3.3 Scaling of intensity qualifiers

Results of the number-for-words and word-for-numbers tasks applied to the intensity qualifiers are presented in *Table 2*, and results for the corresponding cross-language matching task are presented in *Table 5*.

Number-for-Words task

As indicated in *Table 2*, the range of the Chinese intensity qualifiers was 0.6 and 9.7, and that of the English words were 0.7 to 9.8. We compared the scale values of the English words with those reported by Rohrmann (2003) to determine the consistency of scaling across samples from different cultures and first languages (i.e., an Australian sample with English as first language in Rohrmann (2003) and a Hong Kong sample with English as a second language in the present study). The correlation of the scale values of the 19 English qualifiers between the present study and Rohrmann's Australian data is .96 ($p < .001$).

Independent sample *t*-tests found that, compared with Rohrmann's, some of our scale values were clearly higher, e.g. for "not al all" (M's = 1.1 vs. 0.0), "partly" (M's = 4.6 vs. 3.5), "quite" (M's = 6.1 vs. 3.5), "mainly" (M's = 8.0 vs. 6.8), and "very" (M's = 8.6 vs. 7.9), whereas some of our scale values were evidently lower, e.g., for "considerably" (M's = 6.4 vs. 7.6). Scale ratings of 15 of the qualifiers differed between the two studies by less than 0.5 in absolute values. 14 differed by less than 1.0 in absolute values, and 18 out of 19 differed by less then 1.5 in absolute values. The largest difference was 2.6.

Of particular 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". Assuming that western respondents in Rohrmann's study could more accurately evaluate the scale values of the qualifiers, it is likely that some of the Chinese respondents in this study were not proficient enough in English to appraise these five intensity words precisely.

Word-for-Numbers task

As indicated in *Table 2*, the most popular Chinese intensity qualifiers for the five scale levels were (1) YiDianYeBu 一點也不 (75%, 0.6) and Bu 不 (22%, 0.8), (2) ShaoXu 少許 (36%, 3.3) and Bu 不 (25%, 0.8),(3) YiBan 一般 (72%, 5.0), (4) Hen 很 (30%, 7.9) and Po 頗 (23% 6.4), and (5) JiZhi 極之 (43%, 9.6) and WanQuan 完全 (28%, 9.7). For the English intensity qualifiers, we identified one qualifier which was the same as Rohrmann's. Both studies found "not at all" as the most popular qualifier for level 1 (50% & .1 in ours vs. 70% & 0.0 in Rohrmann's), and "not" (38%, 0.7) was our second choice. For level 2, Rohrmann recommended "slightly" (27%, 2.5), which was our second choice (23%, 3.3). Our most popular level 2 qualifier for level 2 was "a little" (30%, 3.1). For level 3 we identified "average" (52%, 4.9) whereas Rohrmann suggested "moderately" (37% & 5.0). For level 4, our choice was "very" (42% & 8.6) whereas Rohrmann's was "considerably" (21% & 7.6). For level 5 both studies found "extremely" to be the most popular word (60% & 9.8 in ours vs. 47% & 9.6 in Rohrmann's). Our second choice was "completely" (33%, 9.8).

3.4 Scaling of frequency qualifiers

The results of the number-for-words and word-for-numbers tasks performed on the frequency qualifiers are presented in *Table 3*, and the cross-language matching task results are presented in *Table 6*.

Number-for-Words task

For the frequency qualifiers, the scale range of Chinese words was 0.0 to 9.6 and that of English words was 0.1 and 9.6. The correlation of the 12 English frequency qualifiers in the present study and Rohrmann's Australian sample is .98 ($p < .001$). Independent sample *t*-tests found that, compared with Rohrmann's Australian sample, our scale values were significantly higher for "occasionally" (M's = 4.4 vs. 3.2), "sometimes" (M's = 5.0 vs. 3.6), "frequently" (M's = 8.3 vs. 7.4), and "mostly" (M's = 8.7 vs. 8.0) and lower for "fairly often" (M's = 5.6 vs. 6.1. In absolute values, eight out of nine differed by less than 0.5.

Word-for-Numbers task

The most popular Chinese frequency qualifiers for the five levels were (1) WanQuanMoYou 完全沒有 (91%, 0.0), (2) HenShao 很少 (33%, 2.0) and ShenShao 甚少 (29%, 1.8), (3) YiBan 一般 (37%, 5.3), JianZhong 間中 (36%, 5.0) and (4) HenDuo 很多 (21%, 8.7), and (5) FeiChangDou 非常多 (51%, 9.6) and JingChang 經常 (38%, 8.9). Our list of English frequency qualifiers matched four out of five results of Rohrmann (2003). Level 1 was "never" (93% & 0.1 vs. 92% & 0.0) in both studies. For level 2 we identified "seldom" (44%, 1.9), whereas Rohrmann proposed "rarely" (49% & 1.3), which was our second choice (26%, 1.5). Level 3 was "sometimes" (50% & 5.0 vs. 50% & 3.6) for both studies. We and Rohrmann also identified "often" as the most popular level-4 qualifier (33% & 6.8 vs. 32% & 6.6), and our second choice was "frequently" (21%, 8.3). For level 5, both studies identified "always" (65% & 9.6 vs. 90% & 10.0).

3.5 Scaling of agreement qualifiers

Results from both the number-for-words and word-for-numbers tasks performed on the agreement-with-statements qualifiers are presented in *Table 4*.

Number-for-words task

Among the agreement qualifiers, the number-for-words scale values of Chinese words ranged from 0.1 to 9.9 and those of English words ranged from 0.2 to 9.8. Again we compared the scale values of the English words with those reported by Rohrmann (2003), i.e., participants for whom English was their first language. Among the 10 agreement words that were scaled in both the present study and Rohrmann's study, the correlation of the scale values is 1.00.

Comparing the number-for-words scale values of each agreement qualifier between the two studies using independent sample *t*-tests, the following resulted: On one hand, we had significantly higher scale values for one item, "mainly agree" (M's = 8.2 vs 7.2). On the other hand, our scale values were significantly lower for "mainly disagree" (M's = 1.9 vs. 2.4). We found no consistent pattern in the way our scale values differed from those reported by Rohrmann. Scale ratings of nine of the qualifiers differed between the two studies by less than 0.5 in absolute values, and all differed by less than 1.0 in absolute values. We concluded that

scale values reported in this study were negligibly different from Rohrmann's (2003) study of English-as-a-first-language participants.

Word-for-numbers task

As can be seen in Table 2, the most frequent choices of agreement qualifiers for the five levels were (1) JueBuTongYi 絕不同意 (selected by 35% of the participants, scale value = 0.1) and WanQuanBuTongYi 完全不同意 (34%, 0.1) for level 1, (2) BuTongYi 不同意 (47%, 2.2) and BuTaiTongYi 不太同意 (46%, 3.0) for level 2, (3) ZhongLi 中立 (88%, 5.0) for level 3, (4) TongYi 同意 (46%, 7.4) and PoTongYi 頗同意 (27%, 6.9) for level 4, and (5) JueDuiTongYi 絕對同意 (37%, 9.9) and WanQuanTongYi 完全同意 (28%, 9.8) for level 5. Turning now to English qualifiers for the five scale levels, we identified only one qualifier as the same as Rohrmann's (2003) finding in an Australian sample. For levels 1 and 5, the most frequently-chosen agreement qualifiers were "fully disagree" (endorsement rate = 53% & scale value = 0.2) and "fully agree" (55%, 9.8), whereas Rohrmann (2003) identified "strongly disagree" (66%, 0.4) and "strongly agree" (68%, 9.6), which were the second popular qualifiers in our study (44%, 0.5; and 44%, 9.5, respectively). For levels 2 and 4, "partly disagree" and "partly agree" (61%, 3.2; and 63%, 6.7) are favoured, which were different from Rohrmann's "somewhat disagree" and "somewhat agree" (38% & 3.2; and 38% & 3.2). *Note:* The essential VSPLs for levels 2 and 4 are either "mainly" or "partly" (dis)agree). "Partly", the preference within the current sample was unfortunately not tested in the Australian study (2003), and therefore pertinent comparisons are not feasible. Both studies identified "neutral" (73%, 5.0 in ours; and 36% & 4.9 in Rohrmann's) as the most popular word for level 3.

3.6 Cross-language matching task

In this task, we asked participants to match Chinese and English qualifiers within each of the three modalities of VSPL's. The results for the tested intensity qualifiers, frequency qualifiers and agreement-with-statements qualifiers are presented in *Table 5*, *6* and *7*. In the following, only *Table 7* will be discussed.

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 top-left cell indicates the mapping between JueBuTongYi (絕不同意) and "Fully Disagree". The first figure in the bracket was the percentage of participants (50% in this case) mapping the English word "Fully disagree" when given the word JueBuTongYi (絕不同意). The second figure in the bracket shows the percentage of participants (35%) choosing the Chinese word JueBuTongYi (絕不同意) when presented with the word "Fully Disagree". The number above the bracket (in bold) is an average of these two figures showing the associations of the Chinese and English qualifiers. In the table, only those figures larger than 20 suggesting that more than 20% of participants associated those two cross-language words together are shown. The rightmost two columns show the English word that was most frequently mapped to the corresponding

Chinese word, and the mapping percentages. E.g., “Fully Disagree” was most frequently chosen (50% of the respondents did) as the English translation for JueBuTongYi (絕不同意). The bottom three rows show the Chinese words that were most frequently mapped to the corresponding English words listed on the top rows. E.g., WanQuanBuTongYi (完全不同意) was chosen by 63% of the respondents as the most appropriate Chinese word among our list for “Fully Disagree”. We will refer to these results when we recommend pairs of Chinese and English VSPLs in the discussion section.

4 Assessing the findings

At first, methodological considerations for selecting verbal scale point levels will be outlined, and then the results from this study assessed.

4.1 Criteria for the utility of verbal scale point labels"

When rating scales measuring psychological variables are to be designed based on psychometric data, essential considerations for choosing a word/expression for a scale point level are (cf Rohrmann 2003):

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

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

4.2 Recommendations for verbal qualifiers in rating scales

Each VSPL has been examined within four different experimental tasks, and the approach to attain recommendations is to consider first, the number-for-words ratings; second, the word-for-numbers numbers; third, the familiarity ratings; and fourth, the cross-language matchings.

In the number-for-words task, we looked for the mean scale values and the 95% confidence interval. In the word-for-numbers task, we checked whether the qualifier exceeded the chance level percentage of responses. In the familiarity task, we inspected whether items were rated as how well-known. Finally, we inspected the corresponding translation of the VSPLs in the cross-language matching task. The recommended VSPLs for the intensity, frequency and agreement-with-statements modality are shown in *Table 8*. The major considerations were:

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 that are closest in mean ratings to these five 'perfect' levels were identified.

Next, we 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 VSPL's 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.

Example regarding the agreement modality: The level "4" Chinese VSPL should be Shi (是) according to the rule of the closest to perfect scale value 7.5. However, because its 95% CI (7.5 +/- 2 * 1.5) covers both adjacent points (level "3" = 5 and level "5" = 10), it did not pass the pertinent criterion. Therefore, it was not chosen. "TongYi" (同意) is the next VSPL with a mean rating (M = 7.4, sd = 0.9) that is nearest to the level "4" perfect point of 7.5. Its 95% CI (5.6, 9.2) does not cover either adjacent level and thus was chosen as a potential level "4" VSPL at this stage.

Criterion 1 is based on 0--2.5--5--7.5--10 as anchors. 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 …", "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.

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-labeling a position in the middle between the endpoints. VSPLs like "medium" or possibly "moderately provide this.

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 *Table 8* which contains suggestions based on both, rationale A and rationale B.

Word-For-Numbers findings

In this task, participants are asked to select one VSPL for each of the levels on the

five-point rating scale, i.e., altogether five items. Restriction on choosing only one VSPL for a certain level helps to identify the most suitable position of a VSPL on the 5-point scale as agreed by most participants. However, the placement of a VSPL on a certain level may also depend on how other VSPLs are placed in other levels. Because of the non-independent nature of this scaling procedure, we considered the results of word-for-numbers as less important, i.e., as a second priority compared to the “number-for-words” ratings. (There's also a risk that respondents state what they believe to be "right", rather than thinking in linguistic terms).

The numbers on the five rightmost columns in *Table 2*, *3* and *4* are the percentages showing the proportion of participants that allocated a particular VSPL to a certain level of the five-point scale. High percentage represents strong endorsement.

An option to assess this is to use the cumulative binomial distribution, in order to identify suitable VSPL in the word-for- numbers task. For example, looking at the English frequency VSPLs: The probability of assigning randomly a particular VSPL (e.g., “often”) out of the 12 VSPLs to a particular level (e.g., level "4") is 1/12. We denote this probability as *p*. The probability that *x* people (e.g., 17) out of a total of *n* participants (e.g., 104) assigned “often” to level 4 is a binomial probability of $\frac{104!}{17!(104-17)!}\left(\frac{1}{12}\right)^{17}\left(1-\frac{1}{12}\right)^{104-17}$.

Thus, the cumulative probability of *x* or more people assigning “often” to level 4 is therefore $\sum_{i=17}^{104}\frac{n!}{i!(n-i)!}p^{i}(1-p)^{n-i}$ = .006. We can then see that any English frequency VSPL that is being randomly assigned 17 or more times (16.3% or above among 104 participants) to a particular level will occur with a probability that is less than .01. Based on similar calculations, we computed that in order to control the random error rate to be below .01, a VSPL for any level should be chosen more than 9.6% and 14.8% among the Chinese and English agreement VSPLs, 11.5% and 16.3% among the Chinese and English frequency VSPLs, and 12.5% and 11.5% among the Chinese and English intensity VSPLs, respectively. These cutoff percentages served as "*criterion 2*" in identifying appropriate VSPL for the five levels.

For example, for level 4 of intensity, although the English VSPL “mainly” would be chosen as it has the closest-to-perfect scale value and meets the 95% confidence interval criterion (first criterion), it was not selected because it was assigned to level 4 by only 8.7% of the participants and it did not meet the 11.5% chance level cutoff (second criterion).

Familiarity

After identifying sound VSPLs by the rules of number-for-words and word-for-numbers, we checked the familiarity ratings of potential VSPLs. Most of the VSPLs chosen so far score at least 5, with 10 being “extremely familiar.”

Pairing of VSPL

In addition to the three appraisals above, we also considered pairings of VSPLs. 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 also *appear* to respondents as equidistant.

Recommended VSPLs that failed a criterion

Because of employing multiple stringent criteria, for some levels of the frequency modality and the intensity modality optimal VSPLs were not available. Therefore it was decided to soften the use of the 95% CI criterion. VSPLs that fail to meet this criterion yet fulfill the second criterion (preferredness) are marked by an asterisk in *Table 8*.

In the frequency modality, relaxing criterion 1 allowed to accept “seldom” (M = 1.9, 95% CI = [-0.44, 4.22] for number-for-words ratings) for level "2", “sometimes” (M = 5.0, 95% CI = [1.81, 8.13]) for level "3", and “often” (M = 6.8, 95% CI = [4.12-9.54]) for level "4". The wide CIs indicate that Chinese participants had diverse interpretations of these VSPLs.

Concerning the intensity modality, loosening up criterion 1 makes “not” (M = 0.7, [-1.16, 2.63]) suitable for level "1" and “a little” (M = 3.1, [-0.29, 6.47]) for level "2" - both regarding an English scale. Regarding a Chinese scale, Bu (不) (M = 0.8, 95% CI = [-1.07, 2.63]) becomes feasible for level "1" and ShaoXu (少許) (M = 3.3, [-0.15, 6.80]) for level "2". These remarks refer to rationale A; the issue does not apply to rationale B.

Cross-language VSPLs

For the translation of VSPLs between Chinese and English, firstly Chinese and English VSPLs were identified independently, using the criteria described above. Then, we tried to match corresponding Chinese and English VSPLs for the respective levels and checked the matching results in *Table 5*, *6* and *7*. Some of the pairs of the cross-language VSPLs matched with the patterns of results in these tables, e.g., 98% matching “neutral” and ZongLi (中立) for the agreement modality and 99% matching “never” and WanQuanMeiYou (完全沒有) for the frequency modality.

However, some of the VSPLs fulfilling criteria 1 and 2 were not often endorsed by participants in their judgments in the cross-language matching task. The pertinent Chinese and English VSPLs are not literal translations of each other; only, they have similar scale level values. For example, in the agreement modality, “partly agree” and TongYi (同意) are choices for level "4", but their meanings are not literally the same.

When designing verbalized rating scales, it is more important for the cross-language scale values of the VSPLs to match with each other than the literal meanings - because the purpose

of a VSPL is to tag on the perceptions representing the thoughts of respondents. For example: The best Chinese and English frequency VSPLs for level "5" are FeiChangDuo (非常多) and “Always”. Even though only 21% of the respondents considered them to be literally equivalent, they can be used interchangeably in translations of a scale, because they are almost on par in scale features (both VSPLs’ mean scale value are 9.6, and they also fulfill criteria 1 and 2 discussed above).

Adding numerical values to a scale layout

Depending on the research issue and type of sample, numerical presentations of scale values in addition to VSPLs help respondents to understand the equidistant interval scale property. Windschitl & Wells (1996) suggested that respondents tended to use deliberate and rule-based reasoning when the measures are in numeric terms, but associative and intuitive reasoning in verbal terms. Similarly, verbal VSPL and numerical value representations may also yield different types of responses from the respondents. For example, in a study about exercise behavior, the frequency modality could be used to measure how often people have done exercises in the past 12 months. If the concern of the research is to investigate respondents’ body health, the numerical value representations may give better usable quantitative data for the reference of the researchers. However, if the research question is mainly about respondents’ efficacy in reducing weight by exercise, respondents’ own qualitative perception of their exercise regularity may be more representative.

During the previous experiments with VSPLs in Germany and Australia (cf. Rohrmann 1978, Rohrmann 2003), additionally tests and interviews about the usability and utility of verbalized rating scales were conducted. These provided evidence that rating scales to be used in experiments or surveys should be designed as a combination of numbers and words, plus possibly graphic/pictorial facets. Regardless whether a rating scale is Chinese or English, the questions is not really "words *or* numbers", the target should be a combination, given that VSPLs are "words *for* numbers".

5 Conclusion and outlook

5.1 Essential findings

The results from this research confirm that 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.

When the data for 61 Chinese and 44 English verbal scale point descriptors were connected, the 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 these 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.

Comparing the results of the English qualifiers in the present study, i.e., the HongKong sample, with those in the prior study, i.e., the English-as-first language sample in Australia (Rohrmann, 2003), either good or moderate similarity was found. Regarding the frequency modality, four of five suggested words chosen for each of the category were the same across the two studies. This suggests a high congruence between the participants from different cultural backgrounds. For the agreement qualifiers, only one qualifier matched exactly in both studies, though Rohrmann's results for levels 1 and 5 appeared as close seconds in our study. (The comparability is restricted though because some essential VSPLs were not yet tested in the prior study). The list of qualifiers on the five levels could be considered largely similar. However, the results for the intensity modality words were quite different in the two studies, except for the similarities in words in levels 1 and 2. The main reason for this was the different rationale developed and utilized in Rohrmann (2003), as outlined in the discussion chapter above. When employing a rationale in which psycholinguistic criteria get more emphasis, four of the five suggested VSPLs are the same.

The psychometric findings of the project are not restricted to 5-point scales - 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 equivalent across languages.

5.2 Further agenda

The function of VSPLs in a rating scale is influenced by both their coherence with the other words in the instrument, and by the presentation, which could be words-only or showing words together with numbers and visual indicators of equidistance. In a field experiment it should be tested firstly whether differently presented scales yield the same substantive results, and secondly which scale type and layout suits best the likings of respondents.

The type of bilingual competence deserves more detailed exploration. For example, people who grew up with one language and only later learned a second language may differ significantly in their two pertinent capabilities. Furthermore, it could be that those growing up bilingually have neither for Chinese nor for English verbalizations a high familiarity.

Cultural differences are also of interest within a country. Regarding Chinese, the sample in this study were Hong Kong Chinese who speak Cantonese, one of seven major dialects in China, and somewhat different from Mandarin. Regarding English, it differs moderately to what degree the 'orthodox' English or ' Australian' English or 'Americanized' English is spoken.

Psychometric research of this issue is yet rare.

Furthermore, both English and Chinese are spoken in a large number of other countries, for example: USA, Canada, South Africa, New Zealand; and Singapore, Taiwan, Malaysia. All these countries have to some degree 'their own' English or Chinese language, and cross-national equivalence can not be taken for granted.

Languages change over time, and therefore it needs to be explored how persistent the appraisals of VSPLs are. A replication of this project a decade later could provide beneficial insights into the steadiness of words and expressions for social and psychological ratings.

A core reason for continued cross-cultural research on scale point labels for verbalized rating scales is, to identify impacts on questionnaire validity - - the means to do this are certainly available.

References

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

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.

Budescu, D. V., & Wallsten, T. S. (1985). Consistency in interpretation of probabilistic phrases. *Organizational Behavior and Human Decision Processes, 36*, 391–405.

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.

Budescu, D. V., Weinberg, S., & Wallsten, T. S. (1988). Decisions based on numerically and verbally expressed uncertainties. *Journal of Experimental Psychology: Human Perception and Performance, 14*, 281–294.

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.

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.

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: John Wiley & Sons, Inc.

Felscher-Suhr, U., Guski, R., & Schuemer, R. (1998). Some results of an international scaling study and their implications 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.

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

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

Johnson, E. M. (1973). *Numerical encoding of qualitative expressions of uncertainty* (Technical Rep. No. 250). U.S. Army Research Institute for the Behavioral & Social Sciences. Arlington, VA: Author.

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.*

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.

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

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.*

Levine, T. R. (1994). Do individuals make interval/ratio level responses to magnitude scaled items? *Journal of Social Behavior and Personality, 9*, 377-38.

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

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

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

Mullet, E., & Rivet, I. (1991). Comprehension of verbal probability expressions in children and adolescents. *Language & Communication, 11,* 217–225.

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.*

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

Presser, S., & Blair, J. (1994). Survey Pretesting: Do different methods produce different results? *Sociological Methodology, 24*, 73-104.

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., 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.

Renooija, S., & Witteman, C. (1999). Talking probabilities: communicating probabilistic information with words and numbers. *International Journal of Approximate Reasoning, 22*, 169-194.

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. (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. (2003). Verbal qualifiers for rating scales: Sociolinguistic considerations and psychometric data. Available on website *www.RohrmannResearch.net*. Retrieved May 17, 2008. {Text will be updated later in 2008.}

Rohrmann, B. (2008 in prep). Rating scale design using verbal qualifiers: Communication facets and psycholinguistic data. Manuscript.

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

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., Knauper, B., Hippler, H.-J., Noelle-Neumann, E., & Clark, L. (1991). Rating scales: numeric values may change the meaning of scale labels. *Public Opinion Quarterly, 55*, 570-582.

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.

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 Univ 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.

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

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, 8*, 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.

Acknowledgement by Bernd Rohrmann

This project was conducted while I was Visiting Professor at the Chinese University of Hong Kong - - yet it would not have been realized without the enormous academic and organizational efforts of my host, Associate Professor Wing-tung Au. Furthermore, Associate Professor Paul Taylor from New Zealand, at that time working at the CUHK, provided most valuable methodological and practical advice. I sincerely thank them very much!

The tables, figures and appendices for this text are provided on the following pages!

Table 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

Table 2

Scaling results for intensity verbal scale point labels

			Familiarity		Nr-for-Words		Word-for-Numbers				
	Verbal Scale Point Label VSPL		M	sd	M	sd	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.

Table 3

Scaling results for frequency verbal scale point labels

			Familiarity		Nr-for -Words		Word-for -Numbers				
	Verbal Scale Point Label VSPL		M	sd	M	sd	1	2	3	4	5
C01	完全沒有	WanQuanMoYou	5.3	2.8	0.0	0.3	91				
C03	很久沒有	HenJiuMoYou	5.4	2.7	1.2	0.8					
C04	甚少	ShenShao	6.5	2.4	1.8	1.2		29			
C02	很少	HenShao	7.3	2.2	2.0	1.0		33			
C07	少許	ShaoXu	6.0	2.3	3.0	1.4					
C05	曾	Zeng	7.1	2.4	3.4	1.9					
C06	有點	YouDian	7.3	2.1	4.1	1.4					
C08	可能	KeNeng	8.1	2.1	4.3	1.8					
C12	有時候	YouShiHou	7.7	1.8	4.6	1.4					
C09	有些	YouXie	7.5	1.9	4.7	1.3					
C10	間中	JianZhong	7.3	2.1	5.0	1.2			36		
C11	一般	YiBan	7.2	2.0	5.3	1.0			37		
C15	常	Chang	7.0	2.2	7.0	1.2				18	
C13	通常	TongChang	7.7	1.9	7.6	1.3					
C14	多	Duo	7.9	1.8	7.6	0.9				15	
C18	時常	ShiChang	6.8	2.1	8.2	1.4					
C17	很多	HenDuo	8.0	1.5	8.7	0.7				21	
C16	經常	JingChang	8.2	2.0	8.9	1.1				14	38
C19	非常多	FeiChangDuo	6.3	2.2	9.6	0.7					51
E05	Never		6.6	3.1	0.1	0.5	93				
E09	Rarely		4.8	2.3	1.5	1.3		26			
E10	Seldom		6.0	2.4	1.9	1.2		44			
E06	Occasionally		5.7	2.4	4.4	1.7		16	17		
E11	Sometimes		7.9	1.8	5.0	1.6			50		
E02	Fairly often		3.8	2.5	5.6	1.5					
E08	Moderately often		3.1	2.4	6.0	1.4					
E07	Often		7.4	2.2	6.8	1.4			16	33	
E12	Very often		6.8	2.2	8.2	1.2				19	
E03	Frequently		6.7	2.2	8.3	1.1				21	
E04	Mostly		6.1	2.1	8.7	1.3					
E01	Always		8.1	2.0	9.6	0.8					65

Note. Familiarity ratings range from 0 = “Extremely unfamiliar” to 10 = “Extremely familiar.” Number-for-words ratings range from 0 = “Extremely low frequency” to 10 = “Extremely high frequency”. 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 11.5% for Chinese VSPLs and 16.3% for English VSPLs are not shown.

Table 4

Scaling results for verbal scale point labels for agreement with statements

	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
C03	絕不同意	JueBuTongYi	5.4	3.2	0.1	0.3	35				
C04	完全不同意	WanQuanBuTongYi	5.5	3.2	0.1	0.4	34				
C02	極不贊成	JiBuZanCheng	5.4	3.0	0.3	0.5	16				
C01	非常不同意	FeiChangBuTongYi	5.4	3.5	0.7	1.0	13				
C06	不是	BuShi	7.6	2.8	1.9	1.3					
C05	不同意	BuTongYi	7.7	2.4	2.2	1.1		47			
C07	不太同意	BuTaiTongYi	6.9	2.1	3.0	0.9		46			
C09	難於決定	NanWuJueDing	5.7	2.7	4.8	0.6					
C24	中立	ZhongLi	6.7	2.4	5.0	0.5			88		
C08	少許同意	ShaoXuTongYi	5.7	2.4	5.8	1.0					
C10	有點是	YouDianShi	5.3	2.8	6.0	0.7					
C11	有一點兒同意	YouYiDianErTongYi	5.4	2.7	6.0	0.8					
C13	頗同意	PoTongYi	6.9	2.2	6.9	0.9				27	
C15	同意	TongYi	8.5	1.5	7.4	0.9				46	
C14	是	Shi	8.5	2.0	7.5	1.5					
C12	贊成	ZanCheng	8.3	1.7	7.7	1.0					
C17	很同意	HenTongYi	7.1	2.2	8.2	0.8					
C23	肯定	KenDing	6.7	2.6	8.6	1.2					
C16	十分同意	ShiFenTongYi	7.7	2.3	8.7	1.3					
C18	非常同意	FeiChangTongYi	7.0	2.7	9.3	0.7					
C20	非常贊成	FeiChangZanCheng	6.7	2.7	9.4	0.6					
C21	極為同意	JiWeiTongYi	5.8	2.8	9.4	1.3					17
C19	完全同意	WanQuanTongYi	6.8	2.6	9.8	0.6					28
C22	絕對同意	JueDuiTongYi	6.1	2.9	9.9	0.5					37
E02	Fully disagree		4.0	2.8	0.2	0.5	53				
E12	Strongly disagree		6.1	3.0	0.5	0.6	44				
E04	Mainly disagree		3.9	2.5	1.9	1.3					
E08	Partly disagree		6.0	2.4	3.2	0.9		61			
E10	Somewhat disagree		4.7	2.5	3.6	0.9		20			
E05	Neither agree nor disagree		5.9	3.0	4.9	0.6			19		
E13	Undecided		5.2	2.9	4.9	0.5					
E06	Neutral		6.8	2.4	5.0	0.4			73		
E09	Somewhat agree		4.5	2.5	6.4	0.6				17	
E07	Partly agree		6.7	2.0	6.7	0.8				63	
E03	Mainly agree		4.7	2.5	8.2	0.8					
E11	Strongly agree		6.9	2.4	9.5	0.5					44
E01	Fully agree		4.5	2.8	9.8	0.5					55

Note:. Familiarity ratings range from 0 = “Extremely unfamiliar” to 10 = “Extremely familiar.” Number-for-words ratings range from 0 = “Extremely low level of disagreement” to 10 = “Extremely high level of agreement”. 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 9.6% for Chinese VSPLs and 14.8% for English VSPLs are not shown.

Table 5

Cross-language matching results for agreement verbal scale point labels

		E02	E12	E04	E08	E10	E05	E13	E06	E09	E07	E03	E11	E01		
		Fully disagree	Strongly disagree	Mainly disagree	Partly disagree	Somewhat disagree	Neither a. nor d.	Un-decided	Neutral	Somewhat agree	Partly agree	Mainly agree	Strongly agree	Fully agree	*Most chosen English VSPL*	%
C03	JueBuTong Yi	**32** (50, 14)	**32** (44, 20)												Fully disagree	50
C04	WanQuan BuTongYi	**68** (73, 63)													Fully disagree	73
C02	JiBuZan Cheng		**49** (73, 25)												Strongly disagree	73
C01	FeiChang BuTongYi		**60** (71, 48)												Strongly disagree	71
C06	BuShi			**22** (43, 2)											Mainly disagree	43
C05	BuTong Yi			**46** (53, 38)											Mainly disagree	53
C07	BuTaiTong Yi				**63** (38, 87)	**67** (48, 85)									Somewhat disagree	48
C09	NanWu JueDing						**29** (23, 34)	**83** (75, 90)							Undecided	75
C24	ZhongLi						**34** (2, 67)		**98** (97, 100)						Neutral	97
C08	ShaoXu TongYi									**34** (48, 20)	**36** (48, 23)				Partly agree	48
C10	YouDian Shi									**30** (50, 10)	**31** (48, 14)				Somewhat agree	50
C11	YouYiDian ErTongYi									**48** (50, 46)	**44** (48, 40)				Somewhat agree	50
C13	PoTong Yi									**25** (36, 14)	**24** (30, 18)	**25** (27, 23)			Somewhat agree	36

C15	TongYi													Mainly agree	56
												39 (56, 22)			
C14	Shi													Mainly agree	49
												24 (49, 0)			
C12	ZanCheng													Mainly agree	50
												27 (50, 4)			
C17	HenTong Yi													Mainly agree	50
												34 (50, 17)			
C23	KenDing													Fully agree	38
C16	ShiFen TongYi												**27** (52, 3)	Strongly agree	52
C18	FeiChang TongYi												**53** (74, 33)	Strongly agree	74
C21	JiWeiTong Yi												**50** (68, 32)	Strongly agree	68
C20	FeiChang ZanCheng												**46** (75, 17)	Strongly agree	75
C19	WanQuan TongYi													**68** (76, 60) Fully agree	76
C22	JueDui TongYi												**20** (37, 4)	**36** (57, 15) Fully agree	57

Most chosen Chinese VSPL	WanQuan BuTongYi	FeiChang BuTongYi	BuTongYi	BuTaiTong Yi	BuTaiTong Yi	ZhongLi	NanWuJue Ding	ZhongLi	YouYiDian ErTongYi	PoTongYi	FeiChang TongYi	WanQuan TongYi	WanQuan BuTongYi
Percentage	63	48	38	87	85	67	90	100	46	40	23	33	60

Note:. The first figure in the bracket is the percentage of an English VSPL chosen to match with a Chinese VSPL. The second figure in the bracket is the percentage of a Chinese VSPL chosen to match with an English VSPL. The figure on top of these (in bold) is the average of the two numbers in the bracket. Entries with a percentage smaller than 20% are not shown. The names of Chinese items written in a Chinese font can be seen in table 4 (above).

Table 6

Cross-language matching results for frequency verbal scale point labels

		E05	E09	E10	E06	E11	E02	E08	E07	E12	E03	E04	E01		
		Never	Rarely	Seldom	Occasionally	Sometimes	Fairly often	Moderately often	Often	Very often	Frequently	Mostly	Always	*Most chosen English VSPL*	%
C01	WanQuan MoYou	**99** (100, 97)												Never	100
C03	HenJiu MoYou		**29** (54, 4)	**21** (38, 4)										Rarely	54
C04	ShenShao		**55** (56, 54)	**40** (40, 39)										Rarely	56
C02	HenShao		**40** (46, 34)	**43** (45, 41)										Rarely	46
C07	ShaoXu			**21** (32, 10)										Seldom	32
C05	Zeng				**23** (40, 5)									Occasionally	40
C06	YouDian					**22** (38, 6)								Sometimes	38
C08	KeNeng				**23** (41, 5)									Occasionally	41
C12	YouShiHou					**74** (87, 62)								Sometimes	87
C09	YouXie					**31** (55, 7)								Sometimes	55
C10	JianZhong				**41** (34, 49)	**30** (42, 18)								Sometimes	42
C11	YiBan						**25** (34, 15)							Fairly often	34
C15	Chang								**37** (53, 21)					Often	53

C13	TongChang								22 (27, 16)					Often	27
C14	Duo													Often	33
C18	ShiChang								**24** (18, 29)	**21** (21, 20)			**27** (36, 18)	Always	36
C17	HenDuo											**28** (25, 31)		Very often	28
C16	JingChang									**23** (15, 31)	**31** (18, 43)		**53** (51, 55)	Always	51
C19	FeiChangDuo									**24** (30, 17)			**21** (26, 15)	Very often	30
Most Chinese	*chosen VSPL*	WanQuan MoYou	ShenShao	HenShao	JianZhong	YouShiHou	JianZhong	JianZhong	ShiChang	JingChang	JingChang	HenDuo	JingChang		
	Percentage	97	54	41	49	62	18	25	29	31	43	31	55		

Note:. The first figure in the bracket is the percentage of an English VSPL chosen to match with a Chinese VSPL. The second figure in the bracket is the percentage of a Chinese VSPL chosen to match with an English VSPL. The figure on top of these (in bold) is the average of the two numbers in the bracket. Entries with a percentage smaller than 20% are not shown. The names of Chinese items written in a Chinese font can be seen in table 4 (above).

Table 7

Cross-language matching results for agreement verbal scale point labels

		E02 Fully disagree	E12 Strongly disagree	E04 Mainly disagree	E08 Partly disagree	E10 Somewhat disagree	E05 Neither a. nor d.	E13 Undecided	E06 Neutral	E09 Somewhat agree	E07 Partly agree	E03 Mainly agree	E11 Strongly agree	E01 Fullya gree	*Most chosen English VSPL*	%
C03	JueBuTong Yi	**32** (50, 14)	**32** (44, 20)												Fully disagree	50
C04	WanQuan BuTongYi	**68** (73, 63)													Fully disagree	73
C02	JiBuZan Cheng		**49** (73, 25)												Strongly disagree	73
C01	FeiChangBu TongYi		**60** (71, 48)												Strongly disagree	71
C06	BuShi			**22** (43, 2)											Mainly disagree	43
C05	BuTongYi			**46** (53, 38)											Mainly disagree	53
C07	BuTaiTong Yi				**63** (38, 87)	**67** (48, 85)									Somewhat disagree	48
C09	NanWuJue Ding						**29** (23, 34)	**83** (75, 90)							Undecided	75
C24	ZhongLi						**34** (2, 67)		**98** (97, 100)						Neutral	97
C08	ShaoXu TongYi									**34** (48, 20)	**36** (48, 23)				Partly agree	48
C10	YouDianShi									**30** (50, 10)	**31** (48, 14)				Somewhat agree	50
C11	YouYiDian ErTongYi									**48** (50, 46)	**44** (48, 40)				Somewhat agree	50
C13	PoTongYi									**25** (36, 14)	**24** (30, 18)	**25** (27, 23)			Somewhat agree	36

C15	TongYi											**39** (56, 22)		Mainly agree	56
C14	Shi											**24** (49, 0)		Mainly agree	49
C12	ZanCheng											**27** (50, 4)		Mainly agree	50
C17	HenTongYi											**34** (50, 17)		Mainly agree	50
C23	KenDing													Fully agree	38
C16	ShiFenTongYi												**27** (52, 3)	Strongly agree	52
C18	FeiChangTongYi												**53** (74, 33)	Strongly agree	74
C21	JiWeiTongYi												**50** (68, 32)	Strongly agree	68
C20	FeiChangZanCheng												**46** (75, 17)	Strongly agree	75
C19	WanQuanTongYi													**68** (76, 60) Fully agree	76
C22	JueDuiTongYi												**20** (37, 4)	**36** (57, 15) Fully agree	57
Most-Chin.	*chosen VSPL*	WanQuan BuTongYi	FeiChang BuTongYi	BuTongYi	BuTai TongYi	BuTai TongYi	ZhongLi	NanWuJue Ding	ZhongLi	YouYiDian ErTongYi	PoTongYi	FeiChang TongYi	WanQuan TongYi	WanQuan BuTongYi		
	Percentage	63	48	38	87	85	67	90	100	46	40	23	33	60		

Note:. The first figure in the bracket is the percentage of an English VSPL chosen to match with a Chinese VSPL. The second figure in the bracket is the percentage of a Chinese VSPL chosen to match with an English VSPL. The figure on top of these (in bold) is the average of the two numbers in the bracket. Entries with a percentage smaller than 20% are not shown. The names of Chinese items written in a Chinese font can be seen in table 4 (above).

Table 8

VSPLs recommended for creating a 5-point scale

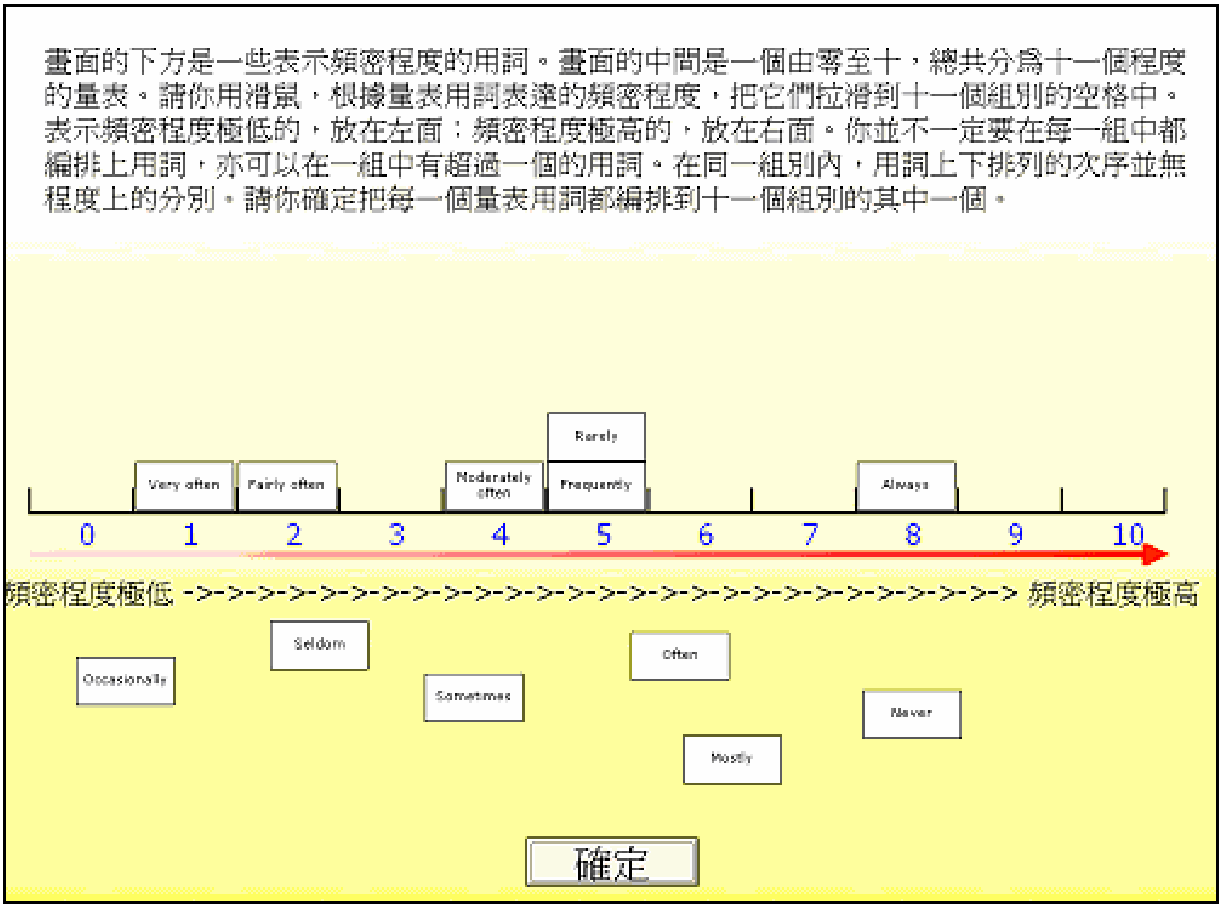
	Intensity modality				**Frequency** modality		**Agreement with statements** modality	
	Rationale A		*Rationale B*		*Rationale A*		*Rationale A*	
Scale level	Chinese	English	Chinese	English	Chinese	English	Chinese	English
1st	Bu	Not*	Bu *or* YiDianYe Bu	Not *or* Not at all	Wan Quen Mei You	Never	Jue Bu 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	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 1" refers to criterion 1 and criterion 2 outlined in the text; VSPLs which do not fully meet these are marked with "*". "Rationale 2" is based on different anchors for scale levels (cf. text).

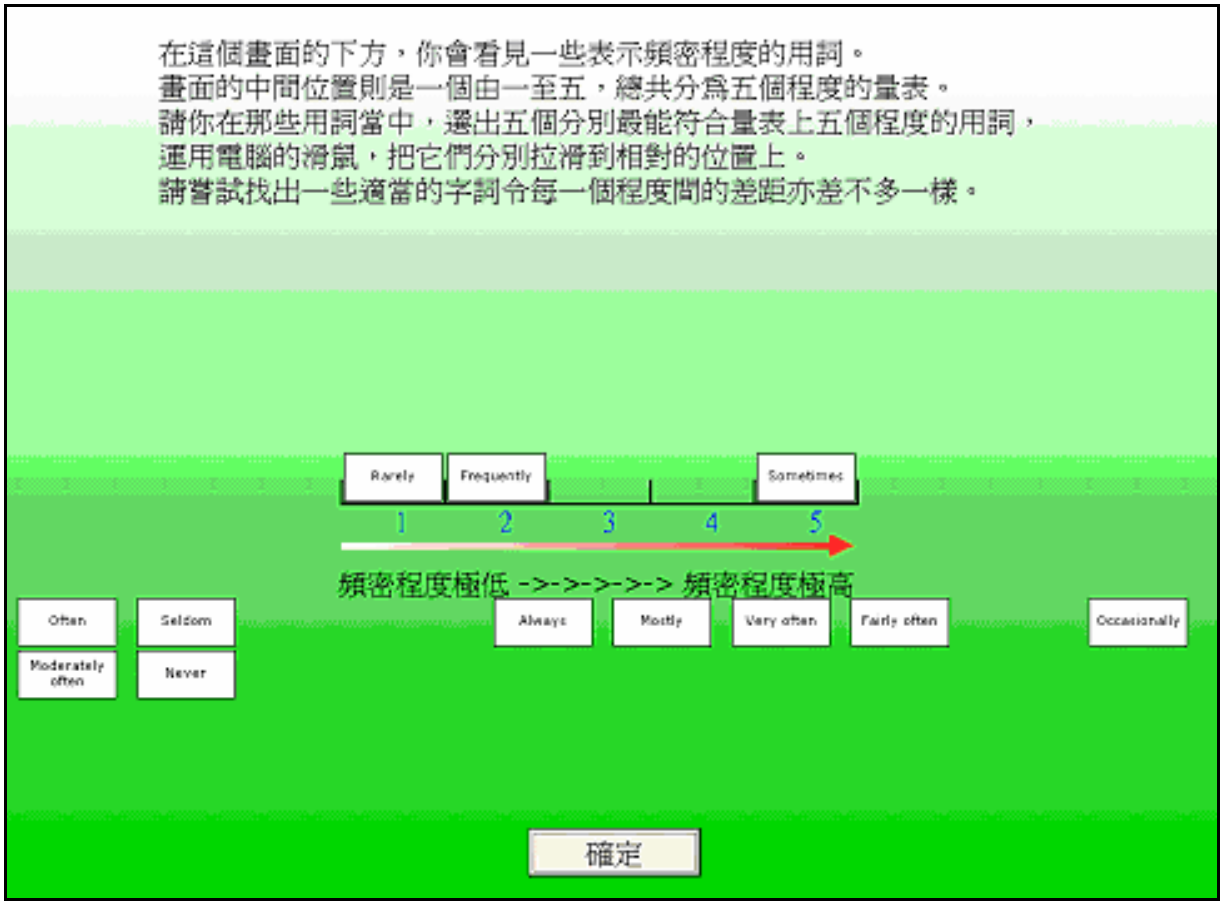
Figure 1. Four screen shots of the computer-based experiment

1A Familiarity task

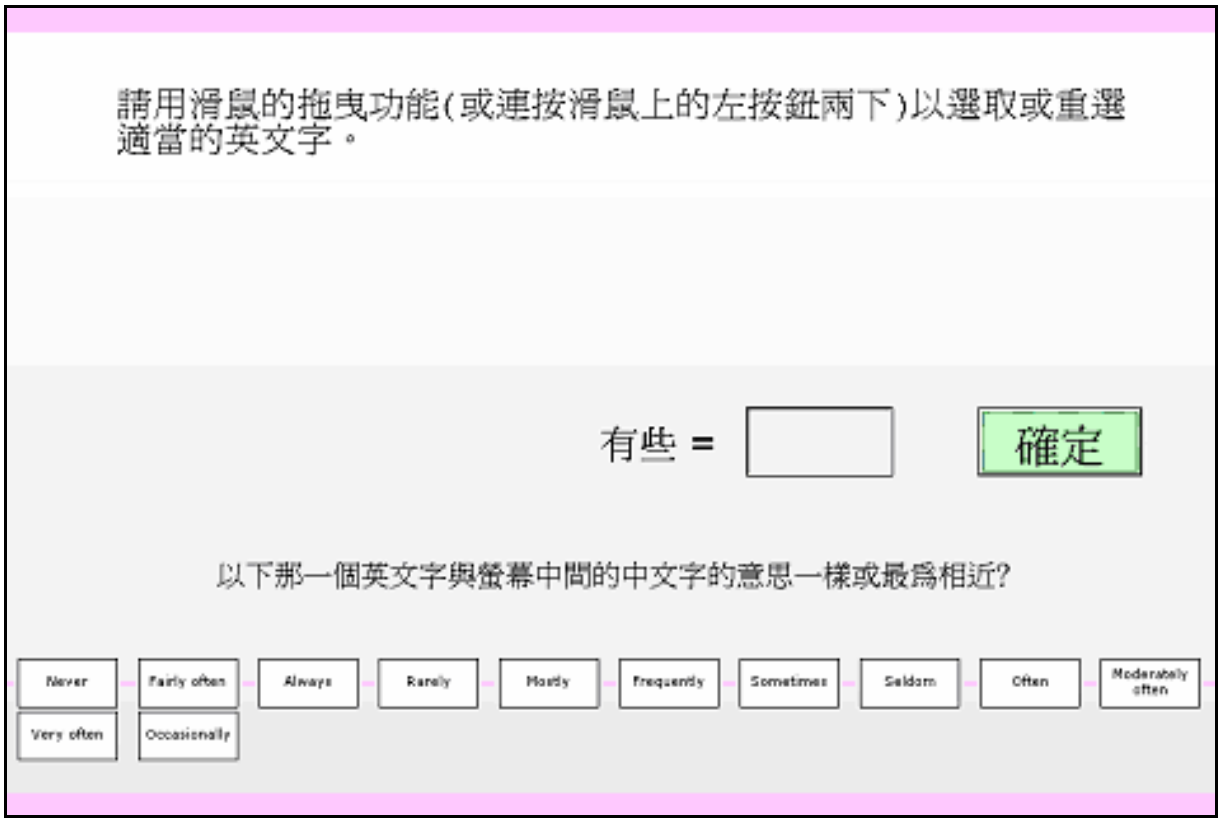
1B Number-for-words task



1C Word-for-numbers task.



1D Cross-language matching task



Appendix A **Wing-tung Au**

Generation of Chinese qualifiers

Books on Chinese function words. Qualifiers are considered "function words" in the Chinese language. From *A Student Handbook for Chinese Function Words* and "現代漢語" (*XianDaiHanYu*) << reference>> we identified 69 and 174 expressions, respectively, that could be used as qualifiers for the three types of attributes.

A top-tiered Chinese psychology journal and student theses. Publications in *The Journal of Acta Psychological Sinica* are generally considered the top psychological journals in Mainland China, Hong Kong, and Taiwan. A search of the 1999 - 2002 issues identified 51 unique qualifiers.

We also consulted Chinese student research theses that have used Chinese rating scales. An electronic database of 589 student theses submitted to the Chinese University of Hong Kong Psychology Department during the 17 years between 1986 and 2002 produced 172 unique qualifiers.

Translation of English qualifiers studied in previous study. A previous study in which English qualifiers were identified (Rohrmann, 2003) yielded 129 qualifiers that were translated to Chinese by the two Chinese authors.

Brainstorming. Finally, the two Chinese authors independently brainstormed qualifiers for each of the three domains resulted in a list of 56 unique qualifiers.

Final amendments. A total of 1,033 qualifiers were identified through the five sources. The two Chinese authors then categorized them into three attribute domains. After deleting duplicating entries, a total of 189 words were identified in the agreement category, 160 in frequency, and 140 in intensity.

Qualifiers for agreement (e.g., "*strongly*" agree) typically reflect intensity, and so we were able to further augment the list of intensity words by retrieving the qualifiers of the agreement words. Among the agreement words, we first removed the "同意" ("agree") portion of the identifiers leaving the intensity words, e.g., trimming "completely disagree" as "completely". Those remaining phrases which were not proper intensity words (e.g., "有點是這樣" "a little bit like this") as judged by the two Chinese authors were discarded. Seventy-eight additional intensity words were identified from this process, and were amended to the original list of intensity words to give a total of 218 intensity words in the revised intensity word list.

Similarly, we amended the list of agreement words by combining intensity words (of the revised list) with "agree", e.g., "extremely" + "agree" → "extremely agree". After removing duplicating entries we compiled a revised list of 238 agreement words.

This initial process resulted in 238 agreement, 218 intensity, and 160 frequency qualifiers.

Appendix B **Wing-tung Au**

Method and results of the short-listing study

Method

Thirty university students participated in this 90-minute pre-study, each receiving HK$75 (~US$9.62). The questionnaires were administered in a group setting in a lecture hall. Three packets of questionnaires, one for each attribute domain, were prepared. The structures of all three questionnaires were similar and the participants completed these questionnaires in a random order.

The first part of the questionnaire was the familiarity rating task. On a scale of 0 to 10 ("extremely unfamiliar" to "extremely familiar"), participants rated how familiar they were with each of the qualifiers. The qualifiers were presented in a randomized order.

The second part of the questionnaire was a scaling task to provide a preliminary assessment of each qualifier's strength. On an 11-point (0 to 10) scale, respondents rated the strength of each of the qualifiers. Scale endpoints were as follows for each of the three attribute domains:

1) agreement – 0 "extremely low level of agreement" to 10 "extremely high level of agreement"
2) frequency – 0 "extremely low level of frequency" to 10 "extremely high level of frequency"
3) intensity – 0 "extremely low level of intensity" to 10 "extremely high level of intensity"

An example of a scaling task of 11 probability words was first presented to respondents for practice, followed by the focal scaling tasks for the agreement, intensity, or frequency words. The list of qualifiers within each of the three attribute domains was randomized.

Results

An initial assessment of the data suggested that respondents seemed to misinterpret the instructions to the scaling task for the intensity qualifiers. After inspection of the pattern of item means, we divided respondents into three groups. The first group of respondents consisted of 13 persons who understood our instructions correctly. The second group was six participants who misunderstood our instructions. They assigned small scale values to "negative" qualifiers like "extremely not", large scale values to "positive" qualifiers like "extremely", and medium scale values to "neutral" qualifiers like "uncertain". The third group of respondents was inconsistent in interpreting the scales. We included only the first group of respondents who understood the task correctly in the analyses. We did not find any patterns of misinterpretation for the agreement and frequency qualifiers because the polarity and extremity of these two types of qualifiers go in the same direction.

The aim of this preliminary short-listing study was to select qualifiers that (a) are familiar to respondents, (b) have a high consensus on the scale values; (c) cover the full range of the rating scale, and (d) (in the case of the intensity qualifiers) grammatically fit typical rating scale statements. We screened qualifiers on the first two of these criteria by selecting those with mean familiarity rating above 6.5, and a standard deviation of the scale values below 2.0.

Agreement qualifiers. Forty-nine of the 238 agreement qualifiers met these first two criteria (mean familiarity > 6.5; SD < 2.0). After rounding-off the scale values to the nearest integer, the number of qualifiers distributed across the 11 (from 0 to 10) scale levels was 3, 5, 5, 1, 4, 9, 0, 5, 6, 10, and 1, respectively. We noted that for some scale levels there (e.g., "6" [the 7th level]) there was no corresponding qualifiers. This is not a concern because the ultimate aim is to construct a five-point scale and we do not need to have qualifiers on each of the 11 scale value. It will be a problem if there were no qualifiers near the five scale values of 0, 2.5, 5, 7.5, or 10. These 49 qualifiers were trimmed down by the two Chinese experimenters independently to select VSPLs that can cover the full span of the scale while maximizing mean familiarity and linguistic compatibility. The trimmed list contained 24 agreement qualifiers--19 were selected by both Chinese authors, and 5 others were selected by only one author. □

Frequency qualifiers. Applying the familiarity (mean > 6.5) and scale value variance (SD < 2.0) criteria reduced the list of 94 frequency qualifiers to 33. The number of qualifiers distributed across the 11 levels was 1, 0, 3, 0, 3, 3, 6, 3, 13, 1, and 0. The two Chinese authors than selected 20 qualifiers (16 suggested by both authors, four suggested by only one author).

Intensity qualifiers. In addition to selecting intensity qualifiers on the basis of familiarity and consensus, we also considered their grammatical fit in typical rating statements. Specifically, we included only words that their mean linguistic compatibility ratings were above 5.0. Thirty-nine of the 218 intensity qualifiers match all three criteria. The number of qualifiers distributed to the 11 levels was 0, 1, 0, 2, 6, 3, 3, 4, 4, 4, and 0. Eighteen of these 39 intensity qualifiers were chosen by the Chinese authors for the subsequent scaling study (17 suggested by both authors and one suggested by one author).

In sum, the preliminary short-listing study resulted in 24 agreement, 19 frequency, and 18 intensity Chinese qualifiers to be used in the scaling study.

Appendix C **Wing-tung Au**

Initial analyses comparing samples and experiment conditions

Student and general public samples. Mean familiarity ratings between the student and general public samples for each of the qualifiers were compared in terms of the relative order of the qualifiers in terms of familiarity between the two samples, as well as differences between the means of the two samples. Correlations in familiarity ratings between the student and general public samples for the Chinese and English agreement words were .92 and .91, respectively, those for Chinese and English frequency words were .96 and .98, respectively, and those for Chinese and English intensity words were .98 and .96, respectively. These results indicate that participants' familiarity with the qualifiers could be considered similar in terms of relative order between the student and general public samples.

Next we focus on the differences between mean familiarity ratings of the student and general public samples. Setting the Type-I error at .05 for each t-test comparison, 4 of 37 agreement qualifiers, 3 of 31 frequency qualifiers, and 9 of 37 intensity qualifiers differed significantly in familiarity ratings between the student and general public samples. Across all of these 16 qualifiers for which significant differences were found, students rated them as more familiar than the general public did. This finding may reflect the fact that the student sample, which consisted of all university students, had higher language proficiency than the general public sample did. In general, the effect sizes of public vs. student samples on familiarity ratings were small. The 105 effect sizes (d) ranged from 0.00 to 0.85, with a mean of 0.21 and a median of 0.17.

We also compared the student and general public samples on the number-for-words task, in which participants assigned qualifiers to one of 11 categories on a scale of 0 to 10. Again, we compared the mean scale values of the student and general public samples. Correlations in mean scale values between the student and general public samples for the Chinese and English agreement words were 1.00 and 1.00, respectively, those for Chinese and English frequency words were .99 and 1.00, respectively, and those for Chinese and English intensity words were .99 and .99, respectively. Setting the Type-I error at .05 for each t-test comparison, 1 of 37 agreement qualifiers, 8 of 31 frequency qualifiers, and 5 of 37 intensity qualifiers differed significantly in scale values between the student and general public samples. Using 5 as the midpoint on this 0 to 10 point scale, six of these significant differences between samples appeared at the high end whereas eight appeared at the low end of the scale. Among all 14 significantly different scale values, the absolute differences were between 0.4 and 1.3 with a median of 0.7. The general public sample gave a higher scale value than the student sample did in all but one occasion.

Separately for each qualifier in the three modalities, chi-square analyses were used to examine whether the public and student samples resulted in comparable placement of qualifiers on the five scale values in the word-for-numbers task. Using .05 as the Type-I error cutoff, we found statistically significant differences in one of the agreement qualifiers and four of the intensity qualifiers. Four of these chi-square analysis tables had 50% to 100% of the cells with expected cell counts less than five that rendered the statistical results questionable. The only analysis that drew our attention was an English intensity qualifier ("Not"). Similar numbers of the public (n=20) and student (n=19) participants assigned this qualifier to the first scale point. For the second scale point, however, there were more public participants (n=13) than student participants (n=3). Due to the small number of statistically significant differences, we concluded that the public and student samples were similar on the word-for-number task.

We did the same chi-square analyses to examine equivalence between the student and the public sample in the cross-language matching task. Using .05 as the Type-I error cutoff, we found statistically significant differences in eight of the agreement qualifiers and three of the frequency qualifiers. Ten of these eleven qualifiers had 40% to 70% of the cells with expected cell counts less than five and we disregarded these results. We noted that, however, for the frequency qualifier FeiChangDuo (非常多), students (n=16) were more likely than the public (n=4) to map it with "mostly" whereas the public (n=21) were more likely than students (n=10) to map it with "very often". Due to the few statistically significant differences, we concluded that results of the cross-language matching task were equivalent between the students and the public sample.

Although the general public seemed to assign some qualifiers to higher scale values than the students, it occurred on only 13% (14 out of 105) of the qualifiers, and the absolute differences were small. Given the near perfect correlation in rank ordering of the scale values between the two samples, and the negligible differences in familiarity ratings, word-for-numbers assignments and cross-language matchings, we concluded that data from the student and general public sample could be combined and treated as a homogenous data set for subsequent analyses.

Order of language presentation. Participants performed the number-for-words task in one of three ways. About one-quarter of the participants scaled the Chinese qualifiers first (Condition 1); English first for another one-quarter of the participants (Condition 2), and both the Chinese and English qualifiers together (Conditions 3 & 4) for the remaining half. We tested whether scaling results differed among these procedures by analyzing the mean scale values of each of the qualifiers' by a one-way ANOVA with three levels (both Chinese &

English words vs. Chinese then English vs. English then Chinese). The *F* (2, 112) values for the 37 agreement qualifiers ranged from .00 to 10.3; the *F* (2, 101) values for the 31 frequency words ranged from .03 to 3.52; and the *F* (2, 101) values for the 37 intensity words ranged from .03 to 6.67. Four agreement, one frequency, and five intensity qualifiers reached the statistically significance level of .05. The standardized mean differences in scale ratings between the Chinese-first and English-first conditions ranged from 0.00 to 1.01 with a mean of 0.29 and a median of 0.26. Standardized mean differences between the Chinese-first and mixed conditions ranged from 0.00 to 1.13 with a mean of 0.22 and a median of 0.19. Finally, those between the English-first and mixed conditions ranged from 0.00 to 0.83 with a mean of 0.19 and a median of 0.16.

For the word-for-numbers task, the orders of presenting the set of Chinese and English qualifiers were counter-balanced. Roughly half of the participants scaled Chinese qualifiers first (Conditions 1 & 3) and the other half scaled English qualifiers first (Conditions 2 & 4). Chi-square analyses were conducted for each of the qualifiers to examine whether assignment of qualifiers to the five scale points were affected by the order of presenting the two languages. No statistically significant chi-square differences (Type-I error rate at .05) were found except for one Chinese intensity qualifier. We regarded the effect of language presentation order on the word-for-numbers task negligible.

The same analyses were performed for the cross-language matching tasks. No statistically significant chi-square differences (Type-I error rate at .05) were found for all 105 qualifiers. We concluded the effect of language presentation order on the cross-language matching task negligible.

In sum, the order of language effects was generally quite small, and so we averaged across the order conditions in the subsequent, substantive analyses.

{ END OF REPORT }