Survey Research Methods (2007)
Vol.1 , No.2 , pp. 97-108
ISSN 1864-3361
http://www.surveymethods.org

Bias and Changes in Perceived Intensity of Verbal Qualifiers Effected by Scale Orientation

Joeri Hofmans, Peter Theuns, Sven Baekelandt, Olivier Mairesse,
Niels Schillewaert* and Walentina Cools
Department of Work and Organisational Psychology, Vrije Universiteit Brussels
* Department of Marketing, Vlerick Leuven Gent Management School Ghent

The objective of this study is to examine whether manipulating the orientation of a rating scale impacts on the perceived intensity of the verbal qualifiers. An experiment is designed to assess the perception of intensities of verbal qualifiers in an agreement scale. A first finding is that participants seem to adopt one of two response strategies. Those who show the 'extreme null-point strategy' report perceived intensities that monotonically increase along with the scale from 'fully disagree' to 'fully agree'. However, other respondents seem to adopt a 'middle null-point strategy', where the highest perceived intensity coincides with both extreme qualifiers of the scale and the lowest intensity is experienced for qualifiers near the middle. An orientation effect is supported for the 'extreme null-point strategy' group, and manifests itself in less agreement about the intensity of the qualifiers when placed on a decremental scale (e.g. fully agree - rather agree - neutral - rather disagree - fully disagree) as opposed to an incremental scale (e.g. fully disagree - rather disagree - neutral - rather agree - fully agree). Next, the existence of a primacy-effect, an orientation effect found in previous research, was tested by means of a web survey-experiment and is rejected in favour of a more differentiated effect.

Keywords: Category rating scales, orientation effects, primacy effect, cross-modality matching.

Introduction

In social sciences, questionnaires can be said to be the dominant method for data collection Myers (2002). In these questionnaires, rating scales constitute the most common response modality (Belson 1966; Poulton 1989; Rohrmann 2002). Many different formats of rating scales exist, but the most frequently used format are itemized rating scales where each response option is accompanied by a specific verbal qualifier (e.g. fully disagree - rather disagree - neutral - rather agree - fully agree) (Breakwell et al. 2000; Cools et al. 2006; Rohrmann 2002).[1] Rating scales have many advantages and are mostly used because of their low cost relative to the size of the covered target group, and because of their simplicity in use (Breakwell et al. 2000). Moreover, respondents prefer rating scales because they experience those as convenient and supportive to express their true opinion (Garland 1990). However, rating scales may be prone to response biases and are suspected to yield data of inferior quality with regards to the attained measurement level.

Some of the biases in rating scales originate from the scaling context. Context in this research applies solely to the rating scale itself and does not for example encompass the question or the order in which the questions are presented. With this in mind, each response option is relative to the presented scale and so the judgment of the intensity or value of a particular response option takes place within a certain scaling context (Mellers and Birnbaum 1982). It is, for example, possible that the judgement of the verbal qualifier 'agree' is different when the rating scale also includes the verbal qualifier 'fully agree' than when 'agree' is the end anchor of the rating scale. According to Strack and Martin (1987), an appraisal, and by inference a persons' judgment within the context of a questionnaire, occurs along 4 stages: understanding, judging, preparing the answer, and treating or adapting the answer, and each of these stages is prone to different biases. Our research focuses on a bias occurring in the third stage, when preparing the answer. Here, the respondent's task is to translate his or her opinion into one of the provided response alternatives and one kind of bias originates from the orientation of the scale, referred to as orientation effects.

Orientation effects are defined as "changes in answers to closed-ended survey questions produced by varying the order in which response options are presented" (Krosnick and Alwin 1987:202). There are two types of orientation effects: recency effects and primacy effects. A recency effect occurs when placement of a response option near the end

Contact information: Vrije Universiteit Brussel, Faculty of Psychology and Educational Science, Department of Work and Organisational Psychology, Brussels, Belgium (joeri.hofmans@vub.ac.be)

[1] Since in this research, the response options are always accompanied by a verbal qualifier, we will simply use the term "response option" when referring to "response option with a verbal qualifier".

of a list increases the probability that the response option will be selected while with primacy effects this probability increases when the response option is near the beginning of a list (Krosnick and Alwin 1987).

Past research on orientation effects in Likert-type rating scales yielded inconsistent results: in certain studies, participants altered their responses when the orientation of the scale changed, whereas in other studies participants' responses remained unaffected (Weng and Cheng 2000). The differences in conclusions might result from the fact that some researchers manipulated scale orientation as a between-subjects variable while others did the manipulation within-subjects (Weng and Cheng 2000). In this research, we opt for a within-subjects design because comparisons of judgments between subjects can be misleading, especially when examining context effects. This is due to the fact that different people choose different contexts when judging different stimuli (Birnbaum 1999). Because the context is obviously the same when using the same people to evaluate different stimuli, the within-subjects design allows us to compare the stimuli without being worried about a confounding interaction between the context and the stimuli.

While most theories suggest that orientation effects impute to the position of a response option by itself, Chan (1991) argued that orientation effects can also be due to a change in perceived intensity of the verbal qualifier, resulting from another position of this qualifier on the scale. This means that the perceived intensity of a verbal qualifier on position x may differ from that of the same verbal qualifier occurring on position y. We surmise that both the lexicographical meaning of the verbal qualifier, or the subjective intensity of the verbal qualifier by itself, and also its position on the scale, can influence the appraisal made by respondents. To our knowledge, changes in value or perceived intensity of verbal qualifiers, effected by displacement of the verbal qualifiers on the scale, have not been investigated systematically.

The measurement of the perceived intensity has, since decades, been object of study in psychophysics. In the 1950s, Stevens (1951) proposed the power law, which relates the intensity of a stimulus to the evoked sensation, based on magnitude estimation (see equation (1)). In magnitude estimation experiments, the task of the participants is to provide estimates of the intensity or magnitude of their sensation evoked by certain stimuli (Falmagne 1985). This can be presented as:

$$\text{stimulus} \rightarrow \text{subjective magnitude}(\Psi) \rightarrow \text{response } (R)$$

Stevens (1951) found that the mean response R can be described by a power function of the physical intensity Φ (Falmagne 1985). Since the response (R) is, according to Stevens (1951), proportional to sensation magnitude (Ψ), one can easily substitute R for Ψ (Gescheider 1988):

$$\Psi = R = k\Phi^{x} \tag{1}$$

The exponent x in equation (1) is specific for each response modality (e.g., numerical estimation, line length, sound, light ...(cf. Gescheider 1988). Since these exponents are empirically well established, one can validate the responses made by the participants by comparing the participant's exponent to the empirically, well established, exponent (Lodge 1981). An example will illustrate this: one could ask a person to judge the thickness of 10 lines (Φ) by allocating a number (R) to each line. Since it is empirically well established that numerical estimation, the allocation of a number, has an exponent of 1 (x) (Stevens 1951), we expect an exponent equal to 1 for our participant when bringing his/her numerical estimates in equation with the stimulus intensities, the thickness of the lines. If the exponent is not statistically different from 1, thus when the judgments are a linear function of the stimulus intensity, we know that the participant is able to use numbers to express his/her impressions of line thickness.

On the other hand, if we validate appreciations of social stimuli like words or the attractiveness of several persons, this method is not adequate since social stimuli lack known metric properties, stated otherwise, we don't know the value of Φ for these stimuli. A related method, based on direct measurement, provides a technique to validate the measurement of social stimuli. In this method, called cross-modality matching, participants are instructed to judge the intensity of the stimuli by means of two response modalities (Lodge 1981). Consequently, the two power laws can be rewritten as a function of the stimulus magnitude Φ and since both response modalities are used to judge the same stimulus, we obtain:

$$\Phi = \frac{R_1}{k_1}^{\frac{1}{x_1}} = \frac{R_2}{k_2}^{\frac{1}{x_2}} \tag{2}$$

When taking the logarithms and omitting the constants, which depend on units of measurement, equation (2) becomes a linear function with the slope equal to the fraction of both exponents:

$$\log R_1 = \frac{x_1}{x_2} \log R_2 + c \tag{3}$$

The latter equation can serve to evaluate the quality of the judgments. This can be done by obtaining a close match between the theoretical and empirically obtained ratios between the two exponents of the response measures (Lodge 1981). Since the exponent in the Power Law is empirically well established for most modalities, one can predict the slope of equation (3), it should be equal to x_1/x_2. An example will illustrate this: suppose we ask a person to judge the attractiveness of 10 individuals by means of numerical estimation and line length production, i.e. assigning a number and drawing a line. Since we know that for numerical estimation (x_1) and for line length production (x_2), the exponents equal 1 we are able to validate the responses made by the respondent. When the respondents' numerical estimates are plotted as a function of his/her line length productions, the slope of that function should equal 1 (or x_1/x_2 should equal $1/1$). The advantage of the cross-modality principle is "that the validation of the magnitude scale is primarily a

function of the response modalities themselves, not the stimuli" (Lodge 1981:30). Since the stimuli used in this study, namely verbal qualifiers, do not have an "objective magnitude", cross-modality matching is particularly suitable here.

As we have shown in the preceding paragraphs, the method of cross-modality matching allows assessing whether changing the orientation of the verbal qualifiers on a rating scale impacts on the perceived intensity of these verbal qualifiers. We expect an impact of the orientation on the perceived intensity of the verbal qualifiers, meaning a difference in perceived intensity for the same verbal qualifier when the orientation is inverted. Subsequently, a survey experiment tests the occurrence of a primacy-effect when inverting the rating scale. In line with the predictions of the satisficing theory and with the results of Krosnick and Alwin (1987), we expect a verbal qualifier placed at the left side of the scale to be selected more often than the same verbal qualifier placed at the right side of the scale. The research thus consists of two experiments: a cross-modality-matching lab-experiment and a survey-experiment. Both experiments are discussed separately.

Lab-experiment

Method

Participants in the lab-experiment were 36 university students, 5 men and 31 women with a mean age of 23.50 (SD=6.20). None of the participants took part in psychophysical scaling experiments before. The experiment was run on identical personal computers in a computer classroom.

Because of the inexperience of the participants with psychophysical scaling experiments, the first condition comprised calibration exercises. In this condition, we had the participants practice magnitude estimation and line length production, the two response modalities used in the remainder of the experiment. As we explained in the introduction, this calibration will allow us to validate the respondents' responses. Participants were instructed to express their perception of each of 10 line lengths into a number (numeric estimation task or NE), and next they adjusted a line's length to each of 10 presented numerals (line length production or LLP) (see Figure 9 and Figure 10 in appendix). All stimuli were presented in random order.

The actual experiment started with condition 2. Here, stimuli consisted of an agreement scale with 5 verbal qualifiers: "fully agree", "rather agree", "neutral", "rather disagree", and "fully disagree".[2] This scale was chosen as it is one of the most frequent kinds of itemized rating scales. Respondents were required to express their subjective appraisal of the intensity of each verbal qualifier into a number (NE) or a line length (LLP). For reasons of standardization of responses among respondents, in each condition, the qualifier "agree" served as a reference stimulus and was linked to an arbitrary standard magnitude (90 for the NE condition, 404 pixels for the LLP condition) (see Figure 11 and Figure 12 in appendix). In both conditions, the presentation order of the stimuli (the 5 verbal qualifiers) was randomized. In order to collect several measures for all qualifiers, both conditions (NE and LLP) were repeated 3 times consecutively (Figure 1). In order to avoid memory effects, LLP and NE were presented alternately. This ensures at least 5 trials with a different response modality between two consecutive presentations of the same verbal qualifier presented with the same response modality.

LLP [a] ► NE [b] ► LLP ► NE ► LLP ► NE

[a] LLP 'Agree'=404 pixels
[b] NE 'Agree' = 90

Figure 1. Experimental design for conditions 2 and 3

In condition 3 of the experiment, the verbal qualifiers were presented in the context of a category rating scale where the qualifier to be judged was printed in red (see Figure 13 in appendix). Each verbal qualifier was judged in the context of a decremental and an incremental scale. When the verbal qualifier "fully agree" is presented leftmost on the scale, we speak of a decremental scale (see Figure 2), while with an incremental scale this qualifier is presented rightmost (see Figure 3). As in condition 2, the stimuli were presented consecutively, 3 times and randomly (see Figure 1).

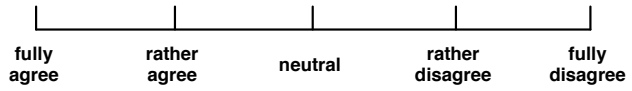


Figure 2. Decremental scale

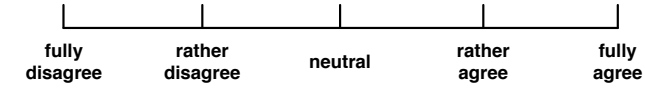


Figure 3. Incremental scale

Results

The object of the first results is the psychophysical validation of the responses for each participant in both calibration conditions. This means that we assess to what extent the participants are able to estimate the objective stimuli in the way predicted by Stevens (1951). In accordance with Stevens (1969), who found that the exponent x in the power law (eq. (1)) equals 1 for both line lengths and numeric estimation, scatters of NE-responses and LLP-responses plotted against the objective stimuli (respectively the line lengths and numbers), should not significantly depart from a straight

[2] In fact the scale was presented in Dutch, the actual verbal qualifiers were, respectively "helemaal akkoord" (Dutch for "fully agree"), "eerder akkoord" (rather agree), "neutraal" (neutral), "eerder niet akkoord" (rather disagree) and "helemaal niet akkoord" (fully disagree).

line with a slope equal to 1. Since the predicted exponent for both NE and LLP equals 1, the power law simplifies to: $R = k\Phi$, thus the response should be a linear function of the stimulus intensity. Remember that we asked the participants to express their perception of each of 10 line lengths in a number and next they adjusted a line's length to each of 10 presented numerals. For each participant we have two equations: one where the numbers are the responses and the line lengths are the objective stimuli (Φ) and one with the line lengths as responses and the numbers as stimuli (Φ). A linear model fitted our data well for 34 of the 36 participants, and this for both NE and LLP (see Table 4 in appendix). For 2 participants a non-linear relationship was observed. For a third participant, the relationship seemed more or less linear, but the data points were widely scattered around a straight line: the coefficient of determination (r^2) for this participant for the numeric estimation task was low (.46 for NE). This suggests that the participant's estimations of the objective line lengths by means of numbers were inaccurate. In this research, a coefficient of determination higher than .70 is considered as satisfactory. Because of the non-linear relationship or the low coefficient of determination, 3 participants were excluded from further analyses. Doing so, coefficients of determination (r^2) (for NE and LLP) ranging from .71 to .99 were obtained (see Table 4 in appendix). This far we know that 33 of the 36 participants could accurately judge the objective stimuli with both response modalities.

Concerning the psychophysical validation of the responses for the social stimuli, the verbal qualifiers, we also checked for a linear relationship for each participant. The difference with validating calibration conditions is that we do no longer compare the responses to the intensity of the objective stimuli. In this case we rather compare NE to LLP responses for all consecutive verbal qualifiers. Since the predicted exponents for both NE and LLP equal 1, both power laws simplify to linear functions and therefore a linear model of this form should result: $R_1 k_2 = R_2 k_1$. Theoretically, this technique is equivalent with the technique proposed by Stevens (1969), which predicts the slope of the line which results when equating the logarithms of the two power functions, to be 1 (see eq. (3)). A linear model will fit only when both exponents equal 1, because under these conditions also the slope is 1 and then a linear function results when both power laws, in this case linear functions, are equated. For 32 of the remaining 33 participants, a linear model fitted significant (see Table 4 in appendix). The participant whose data did not comply with a linear model, was excluded from further analysis as it seemed that this person had responded inattentively in this part of the experiment. However, for another 3 participants the coefficient of determination (r^2) was less than .70. It seems that these 3 participants did not achieve good consistency when estimating the stimuli by means of NE and LLP. These 3 participants were no longer included in the analysis. The remaining 29 participants showed coefficients of determination (r^2) between .70 and .98, when answering to the same stimulus by means of the two response modalities (see Table 4 in appendix). For the calibration, the exponent x in equation (1) was computed as the arithmetic mean of the 29 individual exponents obtained from the included participants. We found $\bar{x}_{NE} = .99$, and $\bar{x}_{LLP} = .99$, which do not differ significantly from 1 (respectively ($t(28) = -.03$, ns) and ($t(28) = -.11$, ns)). Also, for the evaluation of the verbal qualifiers (the second and third condition), both slopes (.97 for x_{NE}/x_{LLP} and .94 for x_{LLP}/x_{NE}) do not differ significantly from the expected value 1 (respectively ($t(28) = -.54$, ns) and ($t(28) = -1.24$, ns)).

This far we know that the remaining participants used the response modalities as predicted by the Power Law. By means of the psychophysical validation it has been confirmed that the 29 remaining participants gave ratio estimates with both NE and LLP. The next step in the analysis is to compute one value for the perceived magnitude of each verbal qualifier and this for each participant separately. Since each response can be expressed as a power function of the intensity of the stimulus, and since each response modality has its own particular exponent, we cannot interpret the raw responses. A correction for the exponent of the modality is needed and therefore we use the following formula to compute the perceived magnitudes (see Lodge 1981):

$$\Psi = \left(R_{NE}^{\frac{1}{x_{NE}}} R_{LLP}^{\frac{1}{x_{LLP}}} \right)^{\frac{1}{2}} \tag{4}$$

Moreover, reconciling both measures (NE and LLP) provides more degrees of freedom in the statistical analysis and a higher reliability of the data (Han 1999). For each participant we use the individual exponents found in the calibration. From this point on, the data are prepared for further analyses.

When reviewing the data, a special result deserves our attention. One group of participants seems to associate the lowest perceived intensity with the verbal qualifier "fully disagree" ($n = 21$), this is illustrated in Figure 5, presenting the results for participant 1. However, the other group of participants ($n = 8$) associated the lowest perceived intensity with "neutral", located in the middle of the scale. To illustrate this, the results obtained from participant 4 can be seen in Figure 4. From now on, we will refer to participants who seem to adhere to either one of these strategies with extreme null-point, and middle null-point respectively. As subjects followed either one of these 2 different approaches, we decided to analyze the data for both strategies separately. These two strategies were also observed in other studies, supporting the validity of our findings (Cools et al. 2004, Hofmans et al. 2005).

In order to assess the effect of the orientation of the scale, we performed a factorial ANOVA, with the verbal qualifier and the orientation of the scale as within-subject factors. The analysis revealed that, for both strategies, only the main effect for the verbal qualifiers was significant ($F(4, 4) = 41.07$, $p < .05$ for middle null-point and $F(4, 17) = 45.36$, $p < .05$ for extreme null-point). A repeated measures ANOVA, comparing the average magnitudes for the consecutive verbal qualifiers in the three conditions (no context, incremental scale, and decremental scale), revealed no significant differences. A different picture appears when comparing the between-subjects standard deviations of the ratings. All anal-

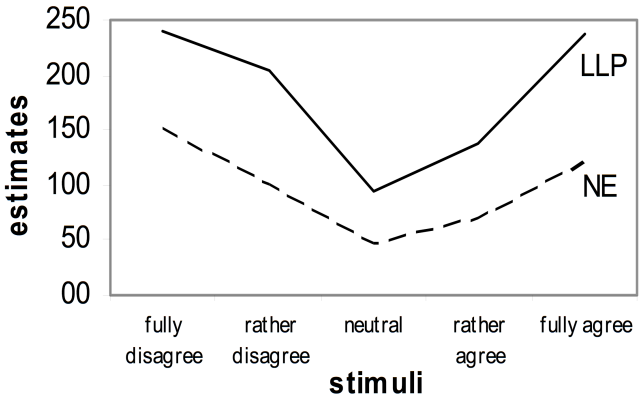


Figure 4. Ratings made by participant 4 with a 'middle null-point'

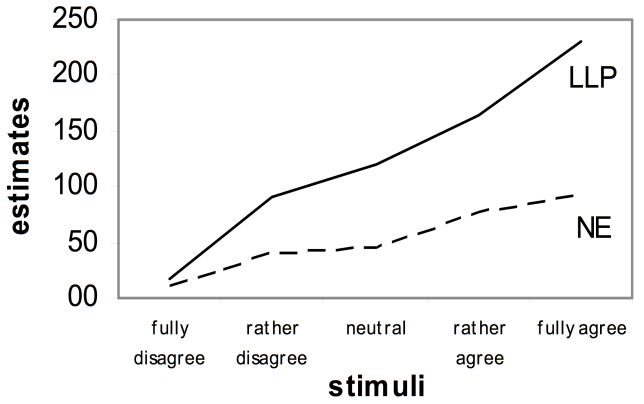


Figure 5. Ratings made by participant 1 with an 'extreme null-point'

yses concerning the standard deviations are carried out on the between-subjects standard deviations since these reflect the amount of disagreement about the intensity of the verbal qualifiers.

For the group with the extreme null-point ($n = 21$), all standard deviations obtained with the incremental scale were significantly smaller than the standard deviations for the decremental scale at the .05 level. The standard deviations for the no-context-condition and for the incremental scale do not differ, only for the verbal qualifier, "neutral", the standard deviation in the no-context-condition is significantly lower than the standard deviation of the incremental scale ($F(20, 20) = 2.28$, $p < .05$). For all qualifiers, the standard deviations in the no context condition are significantly lower than the standard deviations in the decremental scale condition.

The middle null-point group ($n = 8$), associated their perceived null-point with "neutral". For the three central verbal qualifiers, namely "rather disagree", "neutral" and "rather agree", the standard deviations for the three conditions do not differ significantly. For the two extreme qualifiers, however, significant differences are found between both context conditions (incremental scale and decremental scale) and the no context condition. These differences are, for the qualifier "fully agree", significant on the .1-level, and not on the .05-level, because of the small sample size ($n = 8$).

Discussion

Surprisingly, the impact of the orientation of the rating scale on the appraisal of the verbal qualifiers in rating scales seems to depend on the relative position of the perceived null-point of the participant. Concerning this null-point, 2 strategies are discerned (Cools et al. 2004; Hofmans et al. 2005). In the first strategy (extreme null-point), participants seem to locate their lowest perceived intensity near the extreme end "totally disagree". Adherers to the second strategy (middle null-point) seem to locate their lowest perceived intensity in the middle of the scale, near "neutral". Since both strategies are unpredicted and by inference data-driven, we recommend to consider the results of this experiment as being exploratory. In order to strengthen the conclusions, replication of the phenomenon is required. However, assuming that both strategies are valid, i.e. not an artefact of the experimental demands, some conclusions with respect to the effect of the orientation of the rating scale may be formulated.

In the extreme null-point-group, it makes no difference for the subjective intensity of a verbal qualifier whether the verbal qualifiers are put on an incremental scale or whether the verbal qualifiers are presented in absence of a scale (no context). Comparing both conditions (incremental scale and no context) to a decremental scale, we find some differences. The between-subjects standard deviations within the decremental scale-condition are higher than the between-subjects standard deviations in the other two conditions. This means that participants who follow the extreme null-point strategy seem to agree less about the intensity of the verbal qualifiers when they are located on a decremental scale than when located on an incremental scale.

For the middle null-point-group, a very different picture appears. First of all, no differences in subjective intensity were found for the three central verbal qualifiers, namely "rather disagree", "neutral" and "rather agree". The between-subjects standard deviations for the verbal qualifiers "fully agree" and "fully disagree" were higher in the no-context-condition when compared to the incremental and decremental scale. Thus, whether the verbal qualifiers are part of an incremental or a decremental scale makes no difference concerning the amount of disagreement about the intensity of the verbal qualifiers, but putting them on a scale or in a context seems to result in more agreement about the intensity of the response options.

This means that the orientation of the scale seems to have a greater impact on the perceived intensity of the verbal qualifiers for the extreme null-point-group than for the middle null-point-group. In the middle null-point-group, the presence of a scale results in a context that supports better consistency in the intensity ratings of the verbal qualifiers. In the extreme null-point-group, an incremental scale seems most supportive regarding the agreement about the intensity of the verbal qualifiers. Altogether, based on the prelimi-

Table 1: Means and standard deviations for each verbal qualifier for extreme null-point group

	Scale context					
	No context		Incremental scale		Decremental scale	
Verbal qualifier	**M**	**SD**	**M**	**SD**	**M**	**SD**
Fully agree	8.78	11.59	6.30	9.19	10.67	23.56
Rather disagree	52.06	14.83	48.69	14.87	65.77	52.24
Neutral	78.21	16.87	82.55	25.52	101.66	68.70
Rather agree	110.40	30.93	110.70	33.85	149.26	140.00
Fully agree	179.00	36.61	184.70	49.50	228.34	156.10

Table 2: Means and standard deviations for each verbal qualifier for middle null-point group

	Scale context					
	No context		Incremental scale		Decremental scale	
Verbal qualifier	**M**	**SD**	**M**	**SD**	**M**	**SD**
Fully agree	276.20	154.20	195.77	89.46	206.90	91.91
Rather disagree	79.46	28.13	84.15	37.62	84.94	31.10
Neutral	83.34	48.58	85.93	49.49	89.01	52.88
Rather agree	82.83	21.35	78.99	19.94	80.02	20.29
Fully agree	277.20	179.90	235.57	58.82	227.90	49.81

nary results of this experiment, we conclude that incremental scales are the scales of choice.

It must be mentioned, however, that in the English-speaking literature decremental scales are used more frequently (Belson 1966; Chan 1991). To our knowledge, in Dutch, incremental scales are used most. Moreover, the results should be interpreted with the necessary cautiousness since the interpretations concerning the perceived null-point are data-driven. It is an empirical question whether studies with other types of stimuli will find similar results, thereby replicating and validating both response tendencies. Another limitation of this study is the composition of the sample, for the largest part being female, higher educated women. For these reasons, it is suggested to replicate the research with a more diverse sample, in other languages and with other scales.

Web survey experiment

Method

A quota sample[3] of 156 Belgian, Dutch speaking higher educated participants (63 men and 91 women, for 2 respondents the sex is missing) with ages varying between 21 and 64 took part in a web survey experiment. Quotas were set according to 3 variables: sex, age and place of residence (province). The distribution of the sample in terms of educational level is shown in Table 3.

Apart from this sample, from now on called the quota sample, the participants from the aforementioned lab-experiment also took part in the web survey experiment and served as a second, distinct, sample. A soliciting e-mail with a link to the web survey was sent to all participants. Participants of the lab-experiment received the mail one month after completing the lab-experiment. In the instructions, participants were told that they were participating in a study on team roles; the approximate duration of the survey (about 20 minutes) was mentioned and participants were asked to answer as truthful as possible. Participants were able to win a price when participating. Because the survey was concerned with team roles, the sample was limited to higher educated people, thereby maximizing the applicability of the items. Furthermore, this allows a better match with the respondents from the lab experiment, where only university students participated.

Table 3: Sample distribution regarding the educational level

Educational level	**Nr. of participants**	**% of sample**
High School	16	10.3
College	99	63.5
University	28	17.9
Post-University	12	7.7
missing	1	0.6
Total	156	100

The questionnaire was based on the Dutch version of the Self-Perception Inventory[4] (SPI) (Belbin 1986). The items were rearranged into 8 subsets and the orientation of the scale was manipulated: the items in subsets 1 and 4 appeared with a decremental scale and the items in subsets 5 and 8 were

[3] The survey experiment was carried out in cooperation with Insites, a marketing research company specialized in online research. The quota sample was selected from 60000 voluntarily Dutch panel members, being a representative sample of the Dutch speaking Belgians. The survey stopped once 150 respondents participated.

[4] This inventory evaluates the perception of one's own functioning within a team, thereby differentiating between 8 team roles. The original survey is subdivided in 8 blocks consisting of 8 statements or items. In the original Self Perception Inventory, the respondent has to distribute 10 points over the items of each block, thereby distributing 80 points over the entire survey

scored on an incremental scale.[5] Two randomly selected items from each of subsets 1, 4, 5, and 8 were repeated in subset 5, 8, 1, and 4 respectively so that these items were filled-out twice but with reversed scales. The repeated questions can be found in Table 5 in the appendix. To minimize memory effects, there were always 3 complete subsets between the first presentation of the question and the second presentation of that same question (with a reversed scale). The subset with which the survey started was random, but the order of the subsets is preserved because this procedure is expected to minimize memory effects.

Each subset consisted of 8 items to be rated on a 5-point Likert-scale with the following verbal qualifiers: "fully agree", "rather agree", "neutral", "rather disagree" and "fully disagree".[6]

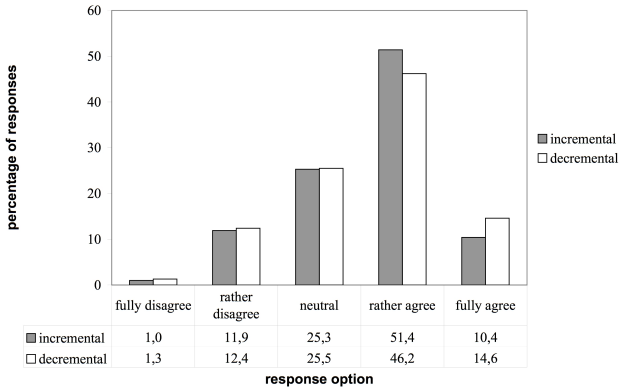
Results

All analyses are within-subjects comparisons for the eight items filled out twice. First, a $2 * 8$ orientation $*$ item factorial ANOVA was performed on the responses for the 8 repeated questions of the quota sample ($n = 156$). The main effect of orientation, the effect of interest, is non-significant ($F(1, 136) = 1.79$; ns), which means that there is no impact of the orientation of the scale on the average values of the ratings.

Next, for the eight repeated questions, the relative frequency of each response option was computed and expressed as a percentage. This was done for both the decremental and the incremental scale, again for those respondents who participated in the web survey experiment only (the quota sample). The relative frequency distribution for the decremental and the incremental scale, summed for the eight repeated questions, is presented in Figure 6.

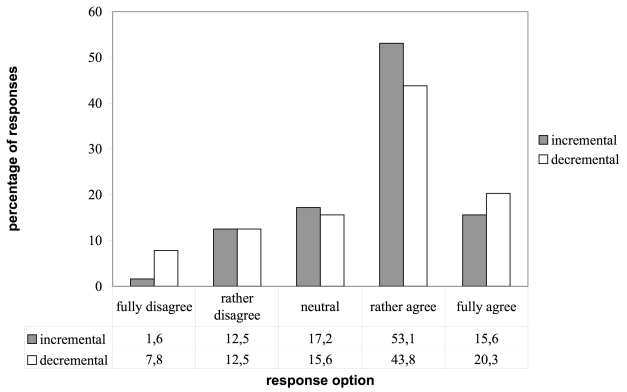
A significant difference is found between the frequency distribution of the decremental and incremental scales, suggesting an impact of the orientation of the scale on the chosen response options ($\chi^2(4, n = 1248) = 12.71$, $p < .05$). When comparing the proportion responses for each response option between the decremental and the incremental scale, by means of pairwise z-tests, significant differences were found for the verbal qualifiers "rather agree" ($z(1247) = 3.71$, $p < .05$) and "fully agree" ($z(1247) = 4.41$, $p < .01$). For the other verbal qualifiers, no significant differences were found.

The relative frequency distributions for the respondents who participated in the lab experiment were analysed for the extreme null-point group and the middle null-point group separately. As Figure 7 and Figure 8 show, similar distributions to the frequency distribution for the quota sample was found for both groups. However, for the extreme null-point group as well as for the middle null-point group, the analysis showed no significant differences between the frequency distributions of the decremental and the incremental scale (respectively ($\chi^2(4, n = 168) = 8.86$, ns) and ($\chi^2(4, n = 64) = 1.26$, ns). This lack of differences is probably due to the lower power of these tests, as compared to the test for the quota sample.



	fully disagree	rather disagree	neutral	rather agree	fully agree
incremental	1,0	11,9	25,3	51,4	10,4
decremental	1,3	12,4	25,5	46,2	14,6

Figure 6. Relative frequency distribution of the responses for the decremental and the incremental scale, summed for the eight repeated items. The results are for the respondents who only participated in the web survey experiment ($n = 156$)



	fully disagree	rather disagree	neutral	rather agree	fully agree
incremental	1,6	12,5	17,2	53,1	15,6
decremental	7,8	12,5	15,6	43,8	20,3

Figure 7. Relative frequency distribution of the responses for the decremental and the incremental scale, summed for the eight repeated items. The results are for the extreme null-point group who also participated in the lab experiment ($n = 21$)

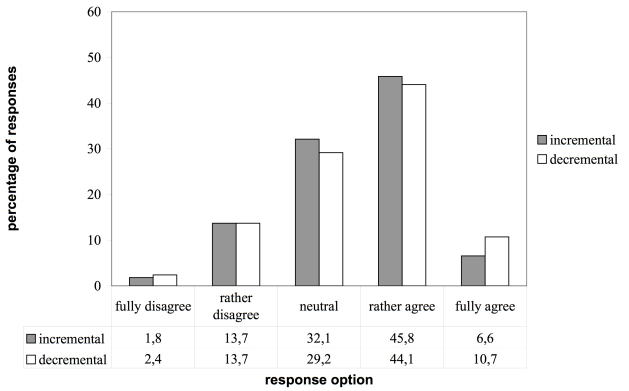
Discussion

The results for the quota sample indicate that the average values of the SPI ratings are not affected by changing the orientation of the scale. However, when reviewing the frequency distributions in this sample, some differences were revealed. When the verbal qualifier "fully agree" is presented leftmost on the scale, thus with a decremental scale, this verbal qualifier is selected more often than when it is presented rightmost, with an incremental scale.

The general primacy effect account would also predict the qualifiers "fully disagree" and "rather disagree" to get

[5] In subsets 2, 3, 6 and 7 the absence or presence of a midpoint on the rating scale was manipulated. Since this manipulation is outside the scope of this paper, these results are not discussed.

[6] Again, the scale was presented in Dutch, the actual verbal qualifiers were, respectively "helemaal akkoord" (Dutch for "fully agree"), "eerder akkoord" (rather agree), "neutraal" (neutral), "eerder niet akkoord" (rather disagree) and "helemaal niet akkoord" (fully disagree).



	fully disagree	rather disagree	neutral	rather agree	fully agree
incremental	1,8	13,7	32,1	45,8	6,6
decremental	2,4	13,7	29,2	44,1	10,7

Figure 8. Relative frequency distribution of the responses for the decremental and the incremental scale, summed for the eight repeated items. The results are for the middle null-point group who also participated in the lab experiment (n = 8)

greater endorsement on an incremental scale than on a decremental scale while "rather agree" should be picked more in a decremental scale than in an incremental scale. Since no significant differences were found for "fully disagree" and "rather disagree", a general primacy effect seems unable to explain our findings. Moreover, the significant difference found for "rather agree" invalidates the general primacy effect account since this qualifier is picked more often in an incremental scale, the opposite of what would be predicted by the general primacy effect account.

Belson (1966) concluded that a general primacy effect account cannot explain the effect of changing the orientation of a scale and he suggested a more differentiated influence on the qualifiers. He noticed that the end items are especially prone to orientation-effects and that they tend to get more endorsement when presented first, we came to the same conclusion for "fully agree", but not for "fully disagree". In line with his findings, we found that the intermediate items were less endorsed when they were presented in the first half of the scale. This finding was confirmed for "rather agree", and not for "rather disagree". Belson's (1966) third finding, that the effect of scale orientation on the centre items was very small or non-existing was also confirmed by our research.

We found similar frequency distributions for the participants of the experiment, but the differences between the frequency distributions were not significant. Although the results do not allow strong conclusions about these two groups, we expect them to show results similar to the stratified sample when increasing the sample size and by inference the power of the tests. This however deserves to be investigated in future research.

As can be seen in Figure 6, 7 and 8, the impact of the orientation of the rating scale on the frequency distributions is rather small in magnitude. This means that the orientation effect may be a statistical matter, having little practical significance because of the small effect sizes. Moreover, the quota sample consists of higher educated people only. This kind of sample maximizes the overlap between the quota sample and the sample from the lab experiments but on the other hand, it limits the generalizability of the results. Additionally, the quota sample is a self-selected sample where people participate in the survey because they are for some reason motivated. This violates the assumption of random sampling for the statistical tests, which may yield standard errors that are too conservative or too large, as demonstrated with the t-test (Reichardt and Gollob 1999). Because of this reason, replication of this study with a random sample is important.

General discussion

In this study we suggested that, besides the position of the response option by itself, also a change in perceived intensity of the verbal qualifier resulting from another position on the scale can account for the appearance of orientation effects in rating scales. Overall, the results tend to indicate that the orientation of the scale does not impact on the average perceived intensity of the verbal qualifiers. Yet another phenomenon was found: participants follow either one of two strategies and these strategies seem to impact on the perception of the verbal qualifiers. For the middle null-point group, the orientation of the scale has no effect on the perception of the verbal qualifiers. Participants in the extreme null-point group in contrast showed more agreement when evaluating the qualifiers on an incremental scale compared to their appraisals of the qualifiers on a decremental scale. Thus although the orientation of the scale has no impact on the average intensity ratings of the verbal qualifiers, using an incremental scale results in more agreement about the intensity of the scale qualifiers for a specific group of people. The existence of both groups has also been demonstrated with an evaluation-scale, containing labels as 'bad', 'average' and 'good' (Hofmans et al. 2005). A possible explanation is that people in the extreme null-point group perceive the scale as one-dimensional while people in the middle null-point group perceive the scale as bidimensional, saying that dimensionality of the scale is at least partly in the eye of the beholder Hofmans et al. (2005). However, this hypothesis is only one of the many possible explanations and deserves to be studied in further research. Since the results of the lab-experiment were unpredicted, and data-driven, they should be interpreted carefully. Replication of the 2 strategies along with the different orientation effects for both groups is needed in order to strengthen the conclusions.

A conclusion drawn from previous research is that the position of a response option on a rating scale has an impact on the frequency distribution of the ratings (Belson 1966; Bishop and Smith 2001; Krosnick and Alwin 1987). As in the research of Weng and Cheng (2000) and Belson (1966), a general primacy effect account was not supported by our data. No differences were found between the averages of the ratings for the questions with an incremental and with a decremental scale. It seems that a more differentiated influence, where the effect depends on the relative placement of the response option fits our data better. The failure in our research to find orientation effects for the response options "fully disagree" and "rather disagree" is probably due

to the sort of questions. The questions in this survey assessed the behaviour of the respondents in a team, thereby probably causing a social desirable response-pattern. This can account for the fact that the positive alternatives are picked more often, and this should be better balanced in further research. However, other confounders, like misrepresentations due to the non-random sample of higher educated people, can also impact on the results of this experiment. Given a number of possible conditional effects in a rather small sample, combined with a lack of specific predictions based on some conceptual model, replication research is essential. However, the results of this study are interesting in that they indicate that the ratings are not independent from the orientation of the scale, a finding which confirms is supported by previous research (Belson 1966; Bishop and Smith 2001; Krosnick and Alwin 1987; Weng and Cheng 2000).

When constructing a survey, the researcher should be aware of the impact of the orientation of the scale on his or her results. In this study the average scores for the survey items were not affected by the orientation of the scale. On the other hand, the orientation of the scale had an impact on the distribution of the responses. If the positive alternatives were presented leftmost, respondents were more inclined to select the extreme option than when presented rightmost. Of course, the sample did not include lower educated people and was non-random, suggesting the need for further research with a random and more representative sample. Furthermore only one type of scale, a 5-point likert scale measuring agreement is tested. It would be interesting to evaluate the impact of scale orientation on other scales.

Acknowledgements

Supported by Grant OZR1041BOF of the Vrije Universiteit Brussel. We are grateful to Frederik Van Acker and two anonymous reviewers for their constructive comments on earlier versions of the manuscript.

References

Belbin, R. M. (1986). *Management teams: Why they succeed or fail*. London: Heinemann.

Belson, W. A. (1966). The effect of reversing the presentation order of verbal rating scales. *Journal of advertising research*, *6*, 30–37.

Birnbaum, M. (1999). How to show that 9 > 221. *Psychological Methods*, *4*, 243–249.

Bishop, G., & Smith, A. (2001). Response-order effects and the early gallup split-ballots. *Public Opinion Quarterly*, *65*, 479–505.

Breakwell, G. M., Hammond, S., & Fife-Shaw, C. (2000). *Research methods in psychology*. London: Sage Publications.

Chan, J. C. (1991). Response-order effects in likert-type scales. *Educational and Psychological Measurement*, *51*, 531–540.

Cools, W., Hofmans, J., Baekeland, S., & Theuns, P. (2004). The numeric estimation of verbal qualifiers measuring level of agreement: differences in response patterns. In A. M. Oliveira, M. Teixeira, G. F. Borges, & M. J. Ferro (Eds.), *Proceedings of the twentieth annual meeting of the international society for psychophysics* (pp. 344–349). Portugal: Coimbra.

Cools, W., Hofmans, J., & Theuns, P. (2006). Context in category scales: is "fully agree" equal to twice agree? *European Review of Applied Psychology*, *56*, 223–229.

Falmagne, J. (1985). *Elements of psychophysical theory*. New York: Oxford University Press.

Garland, R. (1990). A comparison of three forms of the semantic differential. *Marketing Bulletin*, *1*, 19–24. (Retrieved June 29, 2004, from http://marketing-bulletin.massey.ac.nz/article1/V1Article4.pdf)

Gescheider, G. A. (1988). Psychophysical scaling. *Annual reviews Psychology*, *39*, 169–200.

Han, S. H., Song, M., & Kwahk, J. (1999). A systematic method for analyzing magnitude estimation data. *International Journal of Industrial Ergonomics*, *23*, 513-524.

Hofmans, J., Cools, W., Verbeke, P., Verresen, N., & Theuns, P. (2005). A study on the impact of instructions on the response strategies evoked in participants. In J. Monahan, S. Sheffert, & J. Townsend (Eds.), *Proceedings of the twenty first annual meeting of the international society for psychophysics* (pp. 119–124). USA: Traverse City: International Society for Psychophysics.

Krosnick, J., & Alwin, D. (1987). An evaluation of a cognitive theory of response order effects in survey measurement. *Public Opinion Quarterly*, *51*, 201–219.

Lodge, M. (1981). *Magnitude scaling. quantitative measurement of opinions*. London: Sage Publications.

Mellers, B. A., & Birnbaum, M. H. (1982). Loci of contextual effects in judgment. *Journal of Experimental Psychology: Human Perception and Performance*, *8*, 582–601.

Myers, K. (2002). Ten-year review of rating scales. I: overview of scale functioning, psychometric properties, and selection. *Journal of the American Academy of Child and Adolescent Psychiatry*, *41*, 1–22.

Poulton, E. C. (1989). *Bias in quantifying judgments*. London: Lawrence Erlbaum Associates Publishers.

Reichardt, C., & Gollob, H. (1999). Justifying the use and increasing the power of a t test for a randomized experiment with a convenience sample. *Psychological Methods*, *4*, 117–128.

Rohrmann, B. (2002). *Verbal qualifiers for rating scales: Sociolinguistic considerations and psychometric data.* (Project for University of Melbourne)

Stevens, S. S. (1951). Mathematics, measurement and psychophysics. In S. S. Stevens (Ed.), *Handbook of experimental psychology.* New York: Wiley.

Stevens, S. S. (1969). Sensory scales of taste intensity. *Perception & Psychophysics*, *6*, 302–308.

Strack, F., & Martin, L. L. (1987). Thinking, judging and communicating: a process account of context effects in attitude surveys. In H. J. Hippler & S. Schwarz N.and Sudman (Eds.), *Social information processing and survey methodology.* New York: Springer Verlag.

Weng, L., & Cheng, C. (2000). Effects of response order on likert-type scales. *Educational and psychological measurement*, *60*, 908–924.

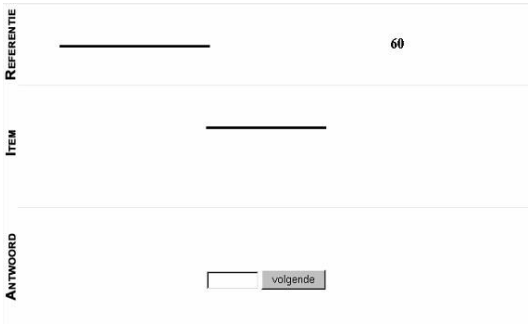


Figure 9. Example of a trial of the Numerical Estimation task in the calibration condition

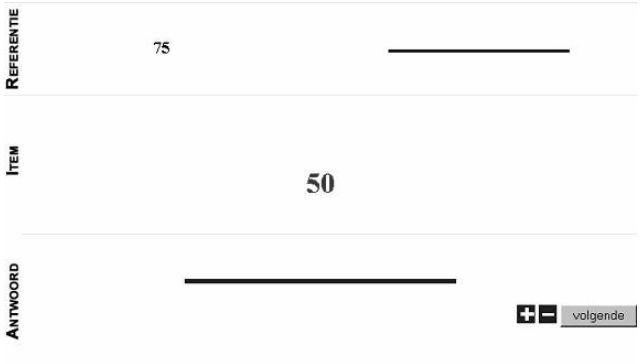


Figure 10. Example of a trial of the Line Length Production task in the calibration condition

Appendix

In the example illustrated by Figure 9, the participants are told that the line in the upper left corner has a value of 60. The task of the participants is to judge the line in the centre relative to the line in the upper left corner. This can be done by assigning a number relative to the reference number (=60). If, for example, a participant perceives the line in the centre as two third of the reference line, he/she should assign the number 40 to this line .

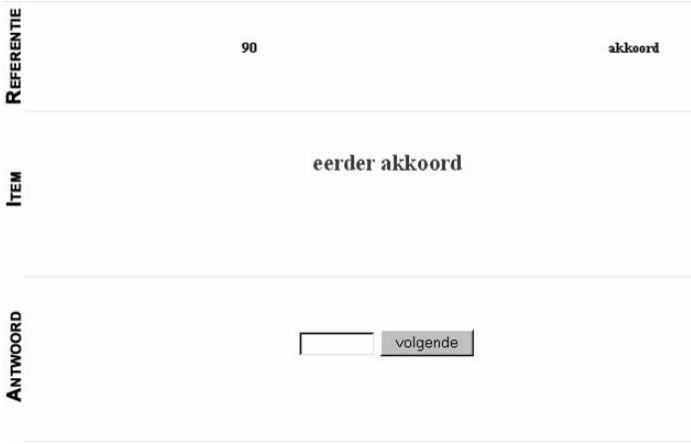


Figure 11. Example of a trial of the Numerical Estimation task in condition 2 (no context)

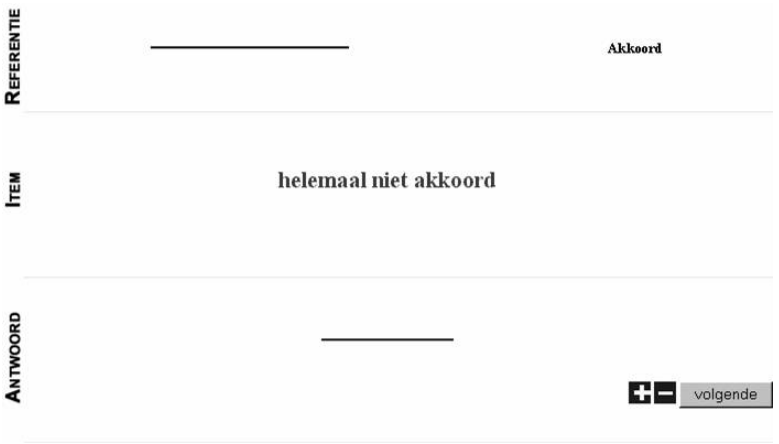


Figure 12. Example of a trial of the Line Length Production task in condition 2 (no context)

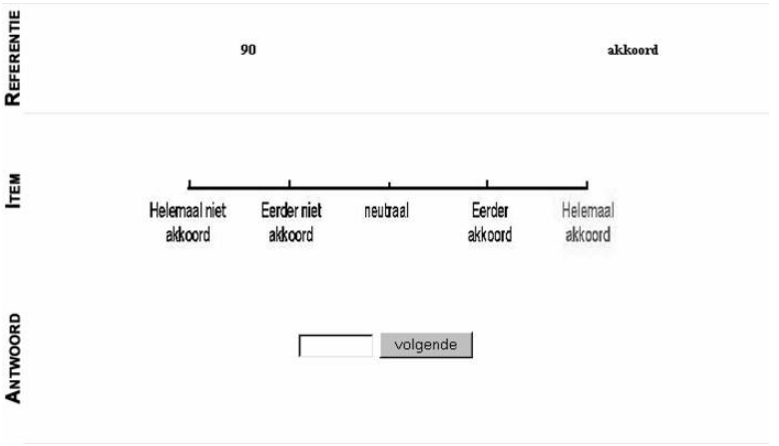


Figure 13. Example of a trial of the Numerical Estimation task in condition 3 (with incremental scale)

In the example illustrated by Figure 10, the participants are told that the number in the upper left corner (75) has a value that equals the length of the line in the upper right corner. The task of the participants is to judge the number in the centre (50) relative to the number in the upper left corner (75). This can be done by adjusting the length of the line relative to the reference line (in the upper right corner). In this example, the participants should draw a line with a length two third (= 50/75) of the reference line.

In the example illustrated by Figure 11, the participants are told that the verbal qualifier in the upper right corner ('akkoord', Dutch for 'agree') has a value of 90. The task of the participants is to judge the verbal qualifier in the centre ('eerder akkoord', Dutch for 'rather agree') relative to the verbal qualifier in the upper right corner ('akkoord', Dutch

for 'agree'). This can be done by assigning a number relative to the reference number (=90). If, for example, a participant perceives the intensity of 'eerder akkoord' as half the intensity of 'akkoord', he/she would assign the number 45 to 'eerder akkoord'.

In the example illustrated by Figure 12, the participants are told that the verbal qualifier in the upper right corner ('akkoord', Dutch for 'agree') has a value that equals the length of the line in the upper left corner. The task of the participants is to judge the verbal qualifier in the centre ('helemaal niet akkoord', Duch for 'fully disagree') relative to the verbal qualifier in the upper right corner ('akkoord', Dutch for 'agree'). This can be done by adjusting the length of the line relative to the reference line (in the upper left corner). If, for example, a participant perceives the intensity of 'helemaal niet akkoord' as one tenth of the intensity of 'akkoord', he/she would adjust the line until it is one tenth of the reference line (in the upper left corner).

In the example illustrated by Figure 13, the participants are told that the verbal qualifier in the upper right corner ('akkoord', Dutch for 'agree') has a value of 90. The task of the participants is to judge the verbal qualifier marked red (in this case grey) ('helemaal akkoord', Dutch for 'fully agree') relative to the verbal qualifier in the upper right corner ('akkoord', Dutch for 'agree'). This can be done by assigning a number relative to the reference number (=90). If, for example, a participant perceives the intensity of 'helemaal akkoord' as twice the intensity of 'akkoord', he/she would assign the number 180 to 'helemaal akkoord'.

The same principle as in condition 2 (no context) was kept up for the Line Length Production. One of the verbal qualifiers is printed in red and the participant is asked to adjust a line length relative to a reference line.

Table 4: r^2 and linear model fit for the calibration conditions

	Calibration NE				**Calibration LLP**				**Calibration social stimuli**			
	r	**df**	**F-value**	**Sign.**	**r**	**df**	**F-value**	**Sign.**	**r**	**df**	**F-value**	**Sign.**
part1	.943	8	131.56	$p < .001$	.731	8	16.34	$p < .010$	.902	18	157.24	$p < .001$
part2	.962	8	203.84	$p < .001$	.842	8	42.48	$p < .001$	**.034**	**18**	**0.60**	***ns.***
part3	.972	8	274.53	$p < .001$	.885	8	62.55	$p < .001$	.816	18	75.18	$p < .001$
part4	.959	8	187.59	$p < .001$	.964	8	215.93	$p < .001$	.891	18	139.49	$p < .001$
part5	.949	8	147.89	$p < .001$	.908	8	78.68	$p < .001$	.942	18	277.77	$p < .001$
part6	.948	8	146.20	$p < .001$	.991	8	866.62	$p < .001$	.817	18	76.00	$p < .001$
part7	.958	8	183.37	$p < .001$	.985	8	517.74	$p < .001$	.798	18	67.27	$p < .001$
part8	.974	8	294.21	$p < .001$	.972	8	278.30	$p < .001$	.970	18	548.09	$p < .001$
part9	.961	8	198.71	$p < .001$	.710	8	19.57	$p < .005$	.896	18	147.05	$p < .001$
part10	.954	8	167.53	$p < .001$	.992	8	977.49	$p < .001$	.962	18	435.86	$p < .001$
part11	.966	8	230.50	$p < .001$	.998	8	4854.31	$p < .001$	.967	18	501.84	$p < .001$
part12	**.000**	**8**	**6.3*10-5**	***ns.***	.978	8	363.92	$p < .001$	.794	18	65.34	$p < .001$
part13	.975	8	310.19	$p < .001$	.989	8	725.15	$p < .001$	**.328**	**18**	**8.31**	***p* < .050**
part14	.714	8	19.96	$p < .005$	.970	8	256.91	$p < .001$	.897	18	147.85	$p < .001$
part15	.947	8	144.26	$p < .001$	.990	8	767.33	$p < .001$	**.596**	**18**	**25.09**	***p* < .001**
part16	.983	8	461.00	$p < .001$	.975	8	317.30	$p < .001$	.932	18	233.60	$p < .001$
part17	.955	8	168.62	$p < .001$	.982	8	448.76	$p < .001$	.969	18	538.96	$p < .001$
part18	.873	8	55.01	$p < .001$	.971	8	271.98	$p < .001$	**.451**	**18**	**13.97**	***p* < .005**
part19	.974	8	295.86	$p < .001$	.997	8	2370.78	$p < .001$	.942	18	275.58	$p < .001$
part20	.988	8	633.78	$p < .001$	.970	8	257.65	$p < .001$	.946	18	296.12	$p < .001$
part21	.988	8	668.26	$p < .001$	.971	8	267.08	$p < .001$	.856	18	100.72	$p < .001$
part22	.978	8	363.22	$p < .001$	.987	8	605.05	$p < .001$	.704	18	40.46	$p < .001$
part23	.794	8	30.80	$p < .005$	.992	8	933.54	$p < .001$	.773	18	57.92	$p < .001$
part24	.768	8	26.48	$p < .005$	.958	8	181.47	$p < .001$	.803	18	69.40	$p < .001$
part25	**.461**	**8**	**6.84**	***p* < .050**	.795	8	31.07	$p < .005$	.458	18	14.34	$p < .005$
part26	.923	8	96.27	$p < .001$	.972	8	273.17	$p < .001$	.732	18	41.04	$p < .001$
part27	**no values**				.968	8	243.10	$p < .001$	.436	15	10.84	$p < .010$
part28	.947	8	143.69	$p < .001$	.885	8	61.64	$p < .001$	.932	18	232.57	$p < .001$
part29	.995	8	1469.80	$p < .001$	.993	8	1129.46	$p < .001$	.945	18	294.77	$p < .001$
part30	.901	8	72.47	$p < .001$	.969	8	252.20	$p < .001$	.983	18	989.42	$p < .001$
part31	.992	8	1021.76	$p < .001$	.994	8	1401.00	$p < .001$	.827	18	81.08	$p < .001$
part32	.911	8	71.45	$p < .001$	.922	8	94.62	$p < .001$	.937	18	251.45	$p < .001$
part33	.833	8	39.92	$p < .001$	.964	8	215.31	$p < .001$	.875	18	118.67	$p < .001$
part34	.983	8	469.07	$p < .001$	.967	8	233.94	$p < .001$	.734	18	46.86	$p < .001$
part35	.952	8	157.22	$p < .001$	.976	8	326.39	$p < .001$	.807	18	71.08	$p < .001$
part36	.857	8	48.08	$p < .001$	.964	8	217.09	$p < .001$	.793	18	65.00	$p < .001$

Note: The participants marked in bold were excluded from further analyses

Table 5: The eight items filled out twice with different scale orientations

Pair	**Subset**	**Item**
1	1 & 5	When involved in a project with other people I have an aptitude for influencing people without pressuring them.
2	1 & 5	I gain satisfaction in a job because I can meet people who may have something new to offer.
3	1 & 5	My characteristic approach to group work is that I have a tendency to avoid the obvious and to come out with the unexpected.
4	1 & 5	What I believe I can contribute to a team I think I can quickly see and take advantage of new opportunities.
5	4 & 8	When involved in a project with other people I can be counted on to contribute something original.
6	4 & 8	My characteristic approach to group work is that I think I have a talent for making things work once a plan has to be put into operation.
7	4 & 8	I gain satisfaction in a job because I am interested in finding practical solutions to problems.
8	4 & 8	If I have a possible shortcoming in team work, it could be that my objective outlook makes it difficult for me to join in readily and enthusiastically with colleagues.