The use of magnitude estimation in large scale survey research

W. E. Saris, C. Bruinsma, W. Schoots and C. Vermeulen*

Abstract

A rarely mentioned problem in social science research consists of the discrepancy between the measurement level of variables on one side and the mathematical properties and requirements of the techniques with which the data are analyzed and compiled, on the other.

Presumably, the problem is often concealed for practical reasons, especially when empirical arguments, concerning the validity, seem to justify the measurement level of analysis.

Some researchers have tried to measure social science concepts on a ratio scale level by a method called magnitude estimation. In this paper the authors have tested whether magnitude estimation, which has been used mainly in small scale research, can be applied in large scale survey research. It turned out that in case of consensus between respondents as well as in case of lack of consensus, magnitude estimation is a useful technique to improve the measurement level of variables.

1. Introduction

Social sciences have been criticized frequently for the use of techniques of analysis which do not match the measurement level of the variables. Especially in the field of attitude research, measurement is mostly done on ordinal level and analysis on interval level. Some researchers have tried to develop techniques of analysis for ordinal variables (Somers 1968, Davis

* Vrije Universiteit Amsterdam.
We thank R. Hamblin for his useful comment.
1967, Goodman 1972) as a solution for this problem. Others have tried to develop measures for social science concepts on interval or higher level. A number of scaling procedures have been developed, and two different approaches to unidimensional scaling have been developed (Torgerson 1958, pp. 52-58):

1. the variability approach;
2. the quantitative judgement approach.

According to the first approach the respondents make ordinal judgements about differences of stimuli and the researcher reconstructs an interval scale on the basis of these judgements and some assumptions concerning the distributions of the judgements around the true score.

In the quantitative judgement approach the respondent not only makes ordinal judgements but he also has to estimate the psychological distances or ratios between the stimuli. In this case the work is done by the respondent and the researcher only has to summarize the results to correct for individual errors.

The best known scaling procedures using the variability approach are the Thurstone scaling procedures (Thurstone 1928) and category scaling (Torgerson 1958). Ratio estimation and magnitude estimation (S. S. Stevens 1956) are the most important techniques in the quantitative judgement approach. In case the resulting scales show linear relationships, the choice between different procedures is irrelevant. However, in several studies relationships appear to be nonlinear. The Thurstone scale turns out to have a logarithmic relationship with the magnitude scales (Kuennapas and Wilkstroem 1963, Ekman and Kuennapas 1962, 1963a, 1963b, Kuennapas and Sellin 1965). The category scales are related with magnitude scales by a function which is somewhere between a logarithmic function and a power function (Stevens 1961, Galanter and Messick 1961, Stevens and Galanter 1957).

Eisler (1965) and Shinn (1973) have shown that nonlinearity is a consequence of deviations from the assumptions made in the variability approaches. By introducing corrections to these assumptions, the scales were found to be linearly related, while goodness of fit was very high ($r^2 = .99$).

On the basis of these studies Shinn concludes (1973, pp. 153-156):

'We find few situations in which either of the Thurstonian methods would produce satisfactory results and since the experimental work involved in gathering the data is typically much greater than that for other methods, their use is not recommended.'

And concerning the category scales:
The category scale has been shown to be definitely nonlinear relative to perceived magnitude and I would recommend that its use be severely curtailed; in its stead I prefer the routine use of magnitude- or ratio-estimation methods, since they provide data which easily meet the interval level assumption underlying many statistics and provides additional ratio level information which may be useful in some context.

Following Shinn’s recommendation we shall concentrate on the quantitative judgement approach in this paper. The problem raised here is whether these techniques can be applied in large scale survey research or, stating it differently, whether the results of small group experiments done by psychophysicists (Stevens et al.) and social scientists (Sellin and Wolfgang 1964, Shinn 1969, Hamblin 1973) can be replicated in large scale survey research.

Results attained until now are very impressive for two reasons: a. a high goodness of fit, and b. the replicability of the results. We shall summarize these results briefly.

Already in the middle of the last century Weber had found an empirical ‘law’ which has been verified since then by many experiments. It states that the change in stimulus intensity that can just be discriminated \(d\varphi\) is a constant fraction \(c\) of the starting intensity of the stimulus \(\varphi\):

\[
d\varphi = c\varphi \quad \text{or} \quad d\varphi/\varphi = c	ag{1}
\]

In the experiments it turned out that in most cases a correction is necessary for the stimulus values around the zero point (Gescheider 1975).

Another law formulated by Ekman (1956, 1959) states that the psychological size of the just noticeable difference \(d\psi\) is a constant fraction \(k\) of sensation magnitude \(\psi\):

\[
d\psi = k\psi \quad \text{or} \quad d\psi/\psi = k
\]

This law also needs some correction for the lowest values of the stimuli according to the empirical tests (Harper and Stevens 1948, Eisler 1962, 1963).

By equating the second form of both equations and integrating both sides a third law can be derived which is known as the ‘power law’ of Stevens. This law states that the sensation magnitude \(\psi\) is a fraction \(c\) of the magnitude of the stimulus \(\varphi\) to the power \(a\) where the constant \(a\) is specific for each specific field of research:

\[
\psi = c\varphi^a
\]

(3)
In earlier times another equation was derived by Fechner on the basis of Weber's law and an alternative form of equation (2). He assumed that \( d_\psi \) is constant (b) for all values of \( \psi \):

\[
d_\psi = b
\]

Then also follows another form for (3):

\[
\psi = c \ln \varphi
\]

where \( c \) is a constant and \( \ln \) stands for the natural logarithm. In many experiments a power function has been found to be the best fitting curve. Gescheider (1975) gives a summary of the results. An example is presented in Figure 1. The goodness of fit of function (3) with data, measured with the \( r^2 \), is in most instances .99 for aggregated data, the median \( r^2 \) for individual data is mostly .95 (Hamblin 1973, p. 69). According to these experiments zero point corrections are sometimes necessary before a good

**Table 1. Summary results: bivariate attitude experiments**

<table>
<thead>
<tr>
<th>Attitude</th>
<th>Social stimulus</th>
<th>Median c</th>
<th>Median ( r^2 )</th>
<th>No. of investigations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aggression</td>
<td>Decisions leading to losses</td>
<td>0.52</td>
<td>.99</td>
<td>1</td>
</tr>
<tr>
<td>Status</td>
<td>Income &lt; 12,000</td>
<td>0.84</td>
<td>.99</td>
<td>5</td>
</tr>
<tr>
<td>Status</td>
<td>Income &gt; 12,000</td>
<td>0.48</td>
<td>.99</td>
<td>5</td>
</tr>
<tr>
<td>Status</td>
<td>Education</td>
<td>2.46</td>
<td>.98</td>
<td>7</td>
</tr>
<tr>
<td>Power</td>
<td>Population size</td>
<td>0.83</td>
<td>.99</td>
<td>1</td>
</tr>
<tr>
<td>Power</td>
<td>Per capita military budget</td>
<td>0.72</td>
<td>.99</td>
<td>1</td>
</tr>
<tr>
<td>Power</td>
<td>Per capita GNP</td>
<td>0.99</td>
<td>.99</td>
<td>1</td>
</tr>
<tr>
<td>Seriousness</td>
<td>Money stolen</td>
<td>0.17</td>
<td>.99</td>
<td>1</td>
</tr>
<tr>
<td>Dislike of self</td>
<td>Killing the enemy</td>
<td>0.1</td>
<td>.99</td>
<td>1</td>
</tr>
<tr>
<td>Dislike of self</td>
<td>Becoming angry at associates</td>
<td>0.5</td>
<td>.99</td>
<td>1</td>
</tr>
<tr>
<td>Dislike of self</td>
<td>Drinks per week</td>
<td>0.8</td>
<td>-.96</td>
<td>1</td>
</tr>
<tr>
<td>Dislike of self</td>
<td>Pounds overweight</td>
<td>1.0</td>
<td>.98</td>
<td>1</td>
</tr>
<tr>
<td>Dislike of wife</td>
<td>Number of premarital sex partners</td>
<td>0.2</td>
<td>.98</td>
<td>1</td>
</tr>
<tr>
<td>Like</td>
<td>Wages</td>
<td>0.66</td>
<td>.99</td>
<td>2</td>
</tr>
<tr>
<td>Dislike</td>
<td>Wages</td>
<td>2.05</td>
<td>.98</td>
<td>2</td>
</tr>
<tr>
<td>Poverty</td>
<td>Income</td>
<td>-3.8</td>
<td>.99</td>
<td>2</td>
</tr>
<tr>
<td>Suicide rate</td>
<td>Status integration</td>
<td>-3.6</td>
<td>.73</td>
<td>2</td>
</tr>
</tbody>
</table>

*Sources:* All of the data come from investigations reported above except for the 'seriousness-money stolen' data. Those come from Sellin and Wolfgang (1964).
fit is found. More recently the same relationship turned up in several experiments where stimuli such as size of income and magnitude-scaled attitudes of respondents, such as social status, were related with each other. In Table 1 we present the results of these experiments as they have been summarized by Hamblin (p. 113). Table 1 makes clear that the fit of the function to the data is nearly as good as in the psychophysical experiments. Given these results Hamblin says:

'There appears to be a general law which describes the relationship between the magnitude of an attitude (A) and the magnitude of its related social stimulus (Σ) as in equation (6).

\[ A = a \Sigma^e \]  
(6)

It must be recognized that this law should be expected to apply only when the attitudinal variable and the social stimulus variable are both measured at the ratio-level, that is, from the origin of the relationship.'

The last remark refers to the fact that in attitudinal research more often than in psychophysical experiments a zero point correction of the stimulus variable is necessary. In that case the relationship is expressed by equation (7):

\[ A = a(\Sigma - b)^e \]  
(7)

where a, b and c are again constants.

Also the relationship between an attitude and its multiple stimuli has been studied. An equation which seems to fit the data rather well is the multiplicative power function (8):

\[ A = a \Sigma_1^{c_1} \Sigma_2^{c_2} \ldots \Sigma_k^{c_k} \]  
(8)

Evidence for this relationship is summarized in Hamblin, Table 3.9 (pp. 114-115) which we present as Table 2. The goodness of fit of equation (8) with other data sets is again very high. However, the first two topics in this table suggest another equation which we shall discuss below.

The status and power experiments suggest an alternative relationship between a general attitudinal judgement and the attitudinal judgements of the stimuli. If we represent the general attitudinal judgement by A and the
attitudinal judgements related to the stimuli by $A_i$ the following equation has shown to fit rather well for these two cases:

$$A = aA_{l1}^c \cdot A_{l2}^c \cdots A_{lk}^c$$  \hspace{1cm} (9)

As far as has been published, explained variance for this multiplicative power function is 2 to 5 percent higher than the fit of a linear additive function, which seems not much of an improvement; however, the difference was statistically significant (Hamblin, p. 98). The advantage of form (9) is that combining equations (6) and (9) leads to equation (8), assuming that the attitudes concerning the aspects mediate between the stimuli and the general attitude. If a linear additive form (10)

$$A = b_0 + b_1A_1 + b_2A_2 + \ldots + b_kA_k$$  \hspace{1cm} (10)

Figure 1. Magnitude estimations of loudness plotted on logarithmic coordinates as a function of stimulus magnitudes: two experiments
Table 2. Summary results of analyses for multiplicative power functions: multivariate attitude experiments

<table>
<thead>
<tr>
<th>Response variable</th>
<th>Stimulus variable</th>
<th>Median c</th>
<th>Median $R'$</th>
<th>No. of experiments</th>
</tr>
</thead>
<tbody>
<tr>
<td>General status</td>
<td>Income status</td>
<td>0.48</td>
<td>0.38</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Occupational status</td>
<td>0.38</td>
<td>0.35</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Educational status</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>General power</td>
<td>Population power</td>
<td>0.50</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Economic power</td>
<td>0.62</td>
<td>0.98</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Military power</td>
<td>0.39</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Poverty</td>
<td>Income deficit</td>
<td>3.85</td>
<td>1.19</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Family size</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Professorial status, local</td>
<td>Merit of teaching</td>
<td>0.65</td>
<td>0.53</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Professorial demeanor</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Professorial status, professional</td>
<td>Professional age</td>
<td>0.16</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Merit of teaching</td>
<td>0.63</td>
<td>0.97</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Cordiality</td>
<td>-0.47</td>
<td>0.62</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Merit of publication</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fair amount to ante</td>
<td>Money to be won</td>
<td>0.98</td>
<td>0.61</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Probability of winning</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Money anted</td>
<td>Money to be won</td>
<td>0.80</td>
<td>0.92</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Probability of winning</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Sources: All of the data are from investigations summarized above except for professional status; see Hamblin and Smith (1966).

would have been chosen for the relationships between the attitudinal judgements, the form of the relationship between the stimuli and the general attitudinal judgement would also be different from (8). In that case the relationship would be an additive power function as presented in equation (11):

$$A = d_0 + d_1 \sum c^1 + d_2 \sum c^2 + \ldots + d_k \sum c^k$$ (11)

As far as we know this function has never been fitted to the data.

In this study we have tested whether the relationship between the stimuli
and attitudinal judgements is a power function of the form (6) or (7) or that a linear or logarithmic curve like (5) has to be preferred.

Furthermore, we compared the goodness of fit of a linear form (10) and a multiplicative power function (9) for the relationship between different attitudinal judgements.

A comparison is also made between the fit of function (8) and (11) and a linear form for the relationship between an attitudinal judgement and its multiple stimuli.

These comparisons have been made on individual level and, in case of reasonable consensus between the respondents, also on aggregated level. In large scale survey research consensus is, of course, problematic, and aggregation is only meaningful when consensus is high. Then aggregation would have the effect of cancelling out random errors. Without consensus, aggregation could lead to false accuracy. Low consensus does not mean that these techniques would be useless. We will return to this point in the discussion.

2. Research design

A considerable number of measurement instructions can be used. Stevens (1956) used the following instruction in his first experiment:

'The left key presents the standard tone and the right key presents the variable. We are going to call the loudness of the standard 10 and your task is to estimate the loudness of the variable. In other words, the question is: If the standard is called 10, what would you call the variable? Use whatever numbers seem to you appropriate — fractions, decimals, or whole numbers. For example, if the variable sounds seven times as loud as the standard, say 70; if it sounds one-fifth as loud, say 2; if a twentieth as loud, say 0.5, etc.

Try not to worry about being consistent. Try to give the appropriate number to each tone regardless of what you may have called some previous stimulus. Press the standard key for one or two seconds and listen carefully. Then press the "variable" for one or two seconds and make your judgement. You may repeat this process if you care to before deciding on your estimate.'

In this experiment nine sound levels were used covering a range of 70 db. The judgements were made by 36 observers. The judgements were averaged by calculating the median over the 36 judgements for each sound level. The results of this experiment, which has been repeated with a different standard, has been presented in Figure 1.

Hamblin has paraphrased the dicta of Stevens for this kind of measurements in nine points:
1. Use a standard whose level does not impress the observer as being either extremely soft or extremely loud (i.e., use a standard in the middle of the stimulus range).

2. Present variable stimuli that are both above and below the standard.

3. Call the standard by a number like “10” that is easily multiplied and divided.

4. Assign a number to the standard only, and leave the observer completely free to decide what he will call the variable. In particular, do not tell the observer that the faintest variable is to be called “1” or that the loudest is to be called some other number. (If the experimenter assigns numbers to more than one stimulus, he introduces constraints of the sort that forces the observer to make categorical rather than magnitude judgments.)

5. Use only one standard in any experiment, but use various standards in later replications, for it is risky to decide the form of a magnitude function on the basis of data obtained with only one standard.

6. Randomize the order of presentation. With inexperienced observers, it is well, however, to start with stimuli that are not extreme and are, therefore, easier to judge.

7. Make the experimental session short, about ten minutes.

8. Let the observer present the stimuli to himself. He can then work at his own pace and so is more apt to be attending properly when the stimulus comes on.

9. Since some estimates may depart widely from those of the average observer, it is advisable to use a group of observers large enough to produce a stable median. Groups of twenty to thirty observers are typically used in psychological experiments, and these are large enough to obtain parameters which vary plus or minus 5 percentage points.

Although there has been a lot of discussion about these dicta (see for example Marks 1974), we followed them quite strictly.

As we were not using trained observers, we have trained people during the interview with a task in which they had to estimate the lengths of rectangulars. To simplify procedures we first asked them to make an ordinal judgement and next a magnitude estimation. The instruction for the training task was as follows:

'The interviewer will give you seven cards on which figures have been drawn. They are numbered from A to F and one card has an S: compare the figure on card S with the other six figures.

You can compare in *words or signs* by saying: this figure is very much bigger (+ + +), much bigger (+ +), bigger (+), equal (o), smaller (−), much smaller (− −) or very much smaller (− − − −).

Then compare the figures using *numbers*. If we say that the standard (S) is 100, assign a number of 300 if a figure is 3 times as big. If the figure is ¼ as big as the standard assign the number 25.'

After the respondent had read this instruction the interviewer put the standard card in front of the respondent. Then he shuffled the other cards and put the top one next to the standard. The respondent made the verbal comparison and after that a judgement in numbers. These judgements
were written down by the respondent. The interviewer then took away the first stimulus and presented the next stimulus for comparison with the standard, and so on.

When the interviewer had the idea that the respondent understood the procedure he went on to the first real judgements with the following words:

'Now you have to judge in the same way how much education different persons have. You again make a judgement in words or signs and in numbers. The standard (S) is somebody who has finished a primary school and who has done two courses organized by the company.'

In this way several topics were handled. After the first five sets of stimuli, some other questions were asked to give the respondent some rest.

There were two kinds of stimuli which had to be judged by the respondents. One kind was concerned with work classification, the other with willingness to participate in actions of the unions. Work classification is normally done on a basis of sixteen different dimensions (Hazewinkel 1967), but we have used only those which in earlier projects had been discovered to be the most important ones, according to the workers:

— formal education;
— number of subordinates;
— work circumstances; and
— years of experience.

We first presented the respondents eight univariate stimuli concerning formal education, asking them to compare the different amounts of formal education. The same was done with the other stimuli: number of subordinates, work circumstances and years of experience.

Then we combined four stimuli, one from each category, to one multidimensional stimulus and asked the respondents to compare these multidimensional stimuli on the amount of money which should be paid for the work. Thus the fifth set of stimuli was a set of multidimensional stimuli concerning different types of functions.

The second topic was willingness to participate in actions of the workers unions.

We presented as stimuli eleven different possible actions from 'protest meeting' to 'occupation of the factory'.

As these stimuli are multidimensional we asked for judgements of different aspects in the same way as we have described above:

— effectiveness of these means;
— the personal risk involved in participation;
— the acceptability of the actions as means;
— their willingness to participate in such actions.
In this way the respondents were confronted with six different sets of stimuli of which four were unidimensional and two multidimensional. Furthermore, two sets contained metric stimuli, the others contained non-metric or mixed stimuli. Table 3 summarizes the kinds of stimuli used.

Table 3. The kinds of stimuli used in this study

<table>
<thead>
<tr>
<th>Metric</th>
<th>Nonmetric</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unidimensional</td>
<td>Number of subordinates</td>
</tr>
<tr>
<td></td>
<td>Years of experience</td>
</tr>
<tr>
<td>Multidimensional</td>
<td>Kind of work</td>
</tr>
<tr>
<td></td>
<td>Work circumstances</td>
</tr>
<tr>
<td></td>
<td>Kind of work</td>
</tr>
<tr>
<td></td>
<td>Actions</td>
</tr>
</tbody>
</table>

As metric stimuli are necessary for the study of the relationship between stimuli and attitudinal judgements, there are two unidimensional sets of stimuli which could be used for this test. For the study of the relationship between the attitudinal judgements nonmetric stimuli could also be used. We can study this relationship for two topics: kind of work and actions.

The metric stimuli concerning the work have also been used to study the relationship between a general attitudinal judgement and its multiple indicators. Although there were four work stimuli we could only use the two which were metric, which means that one cannot expect as high a fit as has been found in other studies.

The nonmetric unidimensional stimuli could not be used separately. They have only been used for the test of the relationship between the attitudinal judgements.

Respondents in this study were workers in a steel factory in the Netherlands. The sample contains 505 respondents. In contrast to other studies, respondents mostly had no other formal education than primary school. In most other studies, students have been used as judges. Nevertheless, we did not receive more complaints about the kind of questions than usually in survey research. No respondents refused to answer the magnitude estimation questions, and the number of missing values was very low (at most two for one judgement).

The amount of time necessary for each topic was very limited. The total interview with 213 questions took an average of 1¾ hour.

In this study we wanted to compare the fit of different curves to different data sets on individual level and aggregated level.

The program ‘Nonlinear’ of SPSS has been used to fit the curves to the
data. This program computes the least squares solution for the parameters.

A comparison of the fit of the different curves is made in two different ways:

1. The fit of the different curves can be compared on the basis of the explained variance ($R^2$). But as the number of parameters in the different curves is different, one has to correct for this. If necessary we adjusted $R^2$ and printed the adjusted value next to the original values in brackets (Theil 1971, pp. 178-179).

As for the analysis on the individual level, for each individual the same analysis was done and $R^2$ computed. The data were summarized by computing the median, as $R^2$ has a skewed distribution.

2. In case of bivariate analysis the fit of the different curves can also be compared on the basis of scatter diagrams in which the best fitting curves are drawn. If the points are distributed around the curve in a random manner the curve is probably correctly chosen, otherwise the choice was incorrect. To see whether the deviations are systematic or not the relationships are linearized by logarithmic transformations of the data. For example: if one takes the logarithm of (6) the result is

\[
\ln(A) = c\Sigma + \ln(a)
\]

or

\[
y = c\Sigma + a^+
\]

where $y = \ln (A)$ and $a^+ = \ln (a)$, which is a linear function. Such transformations are made in all cases and the linear forms are compared.

This comparison is of course only made on aggregated level.

The aggregation is done by the geometric mean as the data are expected to have a log-normal distribution.

In case there are theoretical reasons to expect that consensus between respondents is very low, aggregation is not advisable. Therefore we have tested for consensus by comparing individual judgements with the mean judgements of the sample as a whole. As this is done for each individual the results are again summarized by the median squared correlation.

3. Findings

We shall start by relating attitudinal estimations to the unidimensional
stimuli; next the relationship between the attitudinal judgements will be discussed; and then the relationship between an attitude and its multiple stimuli will be analyzed. The analysis on individual level is reported before the consensus test and the analysis on aggregated level is performed only in case of high consensus.

3.1. The relationship between an attitude and a unidimensional stimulus

Three metric unidimensional data sets were presented to the respondents. One of them was the test data set where the length of differently sized bars had to be estimated.

The two social science stimuli sets were: a. numbers of subordinates, from which the respondents had to estimate the leadership capacity necessary to lead such a group; b. numbers of years one is working in a job, from which the respondents had to estimate how well the person could do the job. For each topic eight stimuli were presented.

Table 4. The median $r^2$ of the different functions for metric unidimensional data on individual level

<table>
<thead>
<tr>
<th>Function</th>
<th>Stimuli</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Length</td>
</tr>
<tr>
<td>Linear</td>
<td>.976 (.968)</td>
</tr>
<tr>
<td>Logarithm (5)</td>
<td>.908 (.887)</td>
</tr>
<tr>
<td>Power (6)</td>
<td>.976 (.968)</td>
</tr>
<tr>
<td>Corrected power (7)</td>
<td>—</td>
</tr>
</tbody>
</table>

Table 4 shows that the respondents were behaving quite well according to the expectations. There is of course a difference in fit between the psychophysical stimuli and the social science stimuli, however the results with the second and third data set are still very good if the zero point correction is introduced. The consensus among the 505 judges concerning these topics is also very high. We tested for this in two different ways: first of all we tested whether there were background variables in our data which were significantly related to the judgements of the respondents. We could not find such variables.

Next we calculated the correlation between the individual scales and the scale found by aggregation of the data over all respondents. In case of high consensus this correlation had to be high too. In Table 5 we present the results of this analysis.
Table 5. The median r² between the individual and group scale

<table>
<thead>
<tr>
<th>Stimuli</th>
<th>Median r²</th>
</tr>
</thead>
<tbody>
<tr>
<td>Length</td>
<td>.984</td>
</tr>
<tr>
<td>Leadership</td>
<td>.865</td>
</tr>
<tr>
<td>Experience</td>
<td>.866</td>
</tr>
</tbody>
</table>

Table 5 shows that there is quite a high consensus between respondents concerning these three topics. Therefore we suppose it is justified to aggregate the data.

By aggregation, random errors cancel out and the form of the curve can be studied more accurately. Therefore we present the results in a table as well as in scattergrams.

Table 6. The fit of different curves for bivariate aggregated data

<table>
<thead>
<tr>
<th>Function</th>
<th>St⁰imuli</th>
<th>Length</th>
<th>Leadership</th>
<th>Experience</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear</td>
<td>.999 (.999)</td>
<td>.798 (.758)</td>
<td>.927 (.915)</td>
<td></td>
</tr>
<tr>
<td>Logarithm (5)</td>
<td>.943 (.932)</td>
<td>.987 (.984)</td>
<td>.989 (.987)</td>
<td></td>
</tr>
<tr>
<td>Power (6)</td>
<td>.999 (.999)</td>
<td>.988 (.987)</td>
<td>.991 (.990)</td>
<td></td>
</tr>
<tr>
<td>Power (corr) (7)</td>
<td>—</td>
<td>.997 (.996)</td>
<td>.996 (.994)</td>
<td></td>
</tr>
</tbody>
</table>

Table 6 shows that the power function fits the data best in all three cases. Only for the first test the linear function fits as well as the power function, and that is because the exponent in (6) is close to 1.0. In the other two cases the linear form is certainly not the right one, but the log function fits nearly as well as the power function. However, inspection of the scattergrams with the fitted curves, presented in Figures 2 through 5, indicates that the residuals are randomly distributed around the fitted curve in case of the power function (2c) or corrected power function (3c, 4c). In all other cases there is some systematic pattern in the errors which indicates that a wrong equation has been fitted to the data. Thus, the power function with zero point correction (7) seems to be the right equation for relating social stimuli and magnitude estimations of attitudes, as was expected from prior research.

The interpretation of this correction is that the zero point of the scale of the stimuli does not coincide with the zero point of the subjective scale of the respondents. For example in the case of experience b = .56 which
Figure 2A
Test data linear function

Figure 2B
Test data log function (5)

Figure 2C
Test data power function (6)
Figure 3A
Leadership linear function

Figure 3B
Leadership log function (5)

Figure 3C
Leadership power (corrected) (7)
Figure 4A
Experience linear function

Figure 4B
Experience log function (5)

Figure 4C
Experience power (corrected) (7)
Figure 5A  Leadership power function (6)  Figure 5B  Experience power function (6)
could mean that the respondents think that only after about six months a worker is qualified to do the job he stands for.

In case of leadership one could expect that the zero point had to be at the value 1 on the scale of the number of subordinates, as there is no leadership capacity necessary below the value of one. The analysis led to an estimate of the value of \( b \) which was close to this value: \( b = .82 \).

These results are in agreement with findings in other studies. Goodness of fit is also not lower than in small scale studies. Thus there is no reason to believe that our respondents could not perform the magnitude estimation task.

3.2. The relationship between attitudinal judgements

So far, the scaling of two metric unidimensional variables, number of subordinates and years of experience, is reported.

We also scaled the two nonmetric unidimensional variables formal education and work circumstances. Descriptions of jobs varying in the amount of education required were presented as stimuli to our respondents, with one of the descriptions as a standard. The judgement procedure resulted in scaling ‘amounts of education’ on an interval scale.

In the same way, descriptions of work situations varying in work circumstances (dust, noise etc.) were presented as stimuli with one description functioning as a standard. This resulted in scaling work circumstances on an interval scale measuring ‘comfort’.

Finally, a set of multidimensional stimuli derived from the four aspects or dimensions judged before were judged according to the wages men in such positions should earn.

After these five judgements the relationship between the last judgement and the other four judgements of the different aspects were analyzed and the results summarized.

In the same way the different actions were compared on different dimensions: effectiveness, the risk involved, and the acceptability as means. Since the respondents also indicated their willingness to participate in the different actions, the relationship was analyzed between willingness to participate and judgements of different actions.

Starting with the analysis on individual level, Table 7 presents the fit of the multiplicative power function (9) and the linear additive function (10). Table 7 shows that the linear additive function fits the data slightly better than the multiplicative power function.

This finding is in contradiction with the two experiments mentioned by
Table 7. The fit of the multiplicative power function (9) and the linear additive function (10) for attitudinal judgements on individual level

<table>
<thead>
<tr>
<th>Functions</th>
<th>Kinds of work</th>
<th>Actions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiplicative power</td>
<td>.68</td>
<td>.71</td>
</tr>
<tr>
<td>Linear additive</td>
<td>.72</td>
<td>.73</td>
</tr>
</tbody>
</table>

Hamblin where the same kind of relationship had been studied.

Hamblin suggested, in a personal communication, that a possible explanation for this difference in results could be that some respondents slipped into category scaling. Therefore, we checked whether the linear additive form also fitted better than the multiplicative power function for respondents who did the bivariate estimation in all cases extremely well (which means an $r^2 > .95$). But also for this group of 82 persons the fit of the linear additive form was better than the fit of the multiplicative power function for the first topic. The median $R^2$ were respectively .812 and .785.

On the basis of this result we conclude that the difference in results cannot be explained by errors in the measurement, but may be caused by difference of topic (see also Fischer 1975).

Perhaps more important than this slight disagreement with previous research is the relatively high explained variance in both cases. This means that the respondents are rather consistent in their judgements across the different kinds of work and across the different actions.

Concerning the consensus, the results were not as positive as for the data sets discussed before. Table 8 presents the median squared correlation coefficient between the individual scales and the total scales.

Table 8. The median $r^2$ between the individual scales and the scale for the sample as a whole

<table>
<thead>
<tr>
<th>Work classification</th>
<th>Median $r^2$</th>
<th>Actions</th>
<th>Median $r^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Capacity</td>
<td>.86</td>
<td>Personal risk</td>
<td>.53</td>
</tr>
<tr>
<td>Leadership</td>
<td>.86</td>
<td>Acceptability</td>
<td>.41</td>
</tr>
<tr>
<td>Formal education</td>
<td>.86</td>
<td>Effectiveness</td>
<td>.22</td>
</tr>
<tr>
<td>Work circumstances</td>
<td>.65</td>
<td>Willingness to</td>
<td></td>
</tr>
<tr>
<td>General judgement</td>
<td>.66</td>
<td>participate</td>
<td>.43</td>
</tr>
</tbody>
</table>

Consensus concerning the differentiation of the kinds of work on differ-
ent dimensions was quite high even for the nonmetric stimuli. We also
could not find any background variables which were strongly related with
the scoring of the respondents. Therefore we think that aggregation, to
correct for random error, is acceptable for the work classification data.
For the ‘actions’ data, however, the consensus is rather low and it was
possible to find groups which differ considerably in their judgements. This
leads to the conclusion, in this case, that variation is not random and that
aggregation is not advisable.

Given these results, we have only analyzed the aggregated data for the
topic ‘work classification’. Results are summarized in Table 9.

Table 9. The fit of the multiplicative power function (9) and the linear
additive power function (10) on aggregated level*

<table>
<thead>
<tr>
<th>Function</th>
<th>Explained variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiplicative power function (9)</td>
<td>.957</td>
</tr>
<tr>
<td>Linear additive function (10)</td>
<td>.986</td>
</tr>
</tbody>
</table>

* This function has not been fitted to the data by the nonlinear program because
the solution is dependent on the start values. Therefore we have fitted the form
\[ A = a + b_1 \hat{A}_1 + b_2 \hat{A}_2 \]
where \( \hat{A}_i = \bar{c}_i(\Sigma_i - \bar{c}_3) \) and \( \hat{A}_i \) was computed on the
basis of the bivariate analysis.

Although the fit of both functions is quite good, the fit of the linear
additive form, with all exponents assumed to be 1.0, is again a little bit
better than the fit of the multiplicative function.

3.3. The relationship between an attitude and its multiple stimuli

As said before, the relationship between an attitude and its multiple stimuli
can only be tested for one topic, and even in this case the data are not
optimal. All functions were characterized by four stimuli: two metric
stimuli and two nonmetric stimuli. The two metric stimuli were ‘years of
experience’ and ‘number of subordinates’. We have fitted a linear additive
function, a multiplicative power function (8) and an additive power func­
tion (11) to these data using only the metric stimuli. Table 10 summarizes
the results.

As expected, the fit of all curves is not as good as in the other cases. Table
10 also shows that the additive power function fits only slightly better than
the other curves for individual data.
Table 10. The fit of three different curves for the relationship between an attitude and its multiple stimuli on individual level

<table>
<thead>
<tr>
<th>Function</th>
<th>Median</th>
<th>Explained variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear additive function</td>
<td>.387</td>
<td>(.319)</td>
</tr>
<tr>
<td>Multiplicative power function (8)</td>
<td>.526</td>
<td>(.473)</td>
</tr>
<tr>
<td>Additive power function (11)*</td>
<td>.534</td>
<td>(.334)</td>
</tr>
</tbody>
</table>

* See note to Table 9.

Table 11. The fit of three different functions for the relationship between an attitude and its multiple stimuli on aggregated level

<table>
<thead>
<tr>
<th>Function</th>
<th>Explained variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear additive function</td>
<td>.582 (.478)</td>
</tr>
<tr>
<td>Multiplicative power function (8)</td>
<td>.789 (.736)</td>
</tr>
<tr>
<td>Additive power function (11)*</td>
<td>.869 (.813)</td>
</tr>
</tbody>
</table>

* See note to Table 9.

Table 11 shows that taking out random error by aggregation increases the fit considerably, and in that case the additive power function fits much better than the other two curves, as is expected when it is assumed that attitudes mediate between a general judgement and its stimuli.

4. Discussion

Our study presents two more cases where the stimuli and the attitudinal magnitude judgements of the respondents were related by a power function. This function had the best fit on the individual level and on the aggregated level.

The fit was no worse than that found in small scale studies, mostly done with students. Workers of the steel factory, with relatively low education, could evidently understand and perform the task as expected.

On the other hand our study did not confirm that the multiplicative power function is the best fitting equation between the attitudinal judgements of stimuli dimensions and the general attitudinal judgement. Although the multiplicative power function fits quite well for individual data...
and very well for aggregated data, the linear additive function always fits a bit better. This result differs from those found for social status and power (Hamblin 1973), but it corresponds with other findings mentioned by Fischer (1975, pp. 12-13).

Consequently, assuming that attitudes mediate between stimuli and general attitudinal judgements, the additive power function resulted in a better fit of the relationship between the general attitudinal judgement and its multiple stimuli. It is possible that this result is specific for this data set, but this function has not been used yet in analyses of other data sets. Thus further research is desired on this point.

More important than these small differences is the fact that in this large scale survey research goodness of fit between results and theory is very high in comparison with the usual survey research. This agrees with the finding of Rainwather (1971) that magnitude estimation is a technique which can be applied with good results in large scale survey research.

At this point we wish to say something more about further analysis that might be done with scores obtained by magnitude estimates of attitudes, both in case of consensus and lack of consensus.

For the high consensus work classification topic there are good reasons to assume that the weights the respondents used to form their general judgement differ only because of random fluctuation. We could not find any background variable which had a significant relationship with these coefficients.

Furthermore, if the coefficients only fluctuate by chance, the means of these coefficients should be equal to the coefficients derived from the data on aggregated level (except for sampling errors). Table 12 shows that the differences between these estimates is maximal .036 for the topic 'work classification'.

Table 12. The regression coefficients of the different aspects of the work classification on the general judgement

<table>
<thead>
<tr>
<th>Data</th>
<th>Experience</th>
<th>Education</th>
<th>Circumstances</th>
<th>Leadership</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aggregated</td>
<td>.185</td>
<td>.308</td>
<td>.210</td>
<td>.341</td>
</tr>
<tr>
<td>Individual</td>
<td>.186</td>
<td>.280</td>
<td>.174</td>
<td>.365</td>
</tr>
</tbody>
</table>

Results similar to those presented here have been found by Rainwather (1971) in a large scale survey research for the topics 'social status' and 'poverty'.

391
1 = Occupation of the whole factory
2 = Occupation of key positions in the factory
3 = Occupation of parts of the factory
4 = Strike
5 = Refusal of overwork
6 = Go-slow movement
7 = Collect signatures
8 = Protest meeting
9 = Spread of stencils
10 = Make demands in the board
11 = Discussion between unions and direction

Figure 6. Comparison of the mean judgements of 'hawks' and 'doves'
In such cases of high consensus the scales and the weights can be used in further research. For example, in this case the scale values and the different weights could be used to calculate a metric score on a scale for each job from simple (not all metric) functional dimensions. These metric scores could then be used in further analysis.

High consensus is however not always to be expected. For example, in our study of social actions we found differences of opinions between ‘hawks’ and ‘doves’ — people who think that only hard actions can have any result and people who think that only discussion can have any result. On the basis of our data we could trace these two groups. Figure 6 presents the differences in attitudinal judgements between them. This lack of consensus does not mean that the procedure is useless for further analysis. There is little reason to suspect the correctness of the scores, for the consistency of the respondents individually was no less for this than for the topic work classification (see section 3.2). It therefore seems interesting in this case to find variables which could explain the variance in the scoring.

Another source of lack of consensus can be found in the different weights the respondents attach to the judgements of aspects of the stimuli in formulating their general judgement. Some people in our study weighed effectiveness much higher than personal risk or acceptability. Others weighed acceptability more heavily than the other two, and so on. These weights are represented by the regression coefficients in equation (8) or (10). We have attached these coefficients as individual scores to our data file and the variation in these coefficients will be analyzed in the usual way. Similar parameter studies, but with different kinds of data, have been done by Griliches (1957) and Hamblin (1977).

This discussion indicates that in case of consensus, as well as in case of lack of consensus, magnitude estimation can produce metric scores which can be used in further analysis.

For such studies it is important to have many judgements of the respondents in order to achieve more reliable results. Although we have used only a few judgements for each topic, our study suggests that by the use of magnitude estimation the explained variance can be raised considerably.

References


Shinn, A. M., 'Relations between scales', in: H. Blalock Jr. (ed.), Measurement in


Thurstone, L. L., 'Attitudes can be measured', American Journal of Sociology, 1928, 33:529-544.