DOCUMENT RESUME

ED 106 308 TM 004 414

AUTHOR Larsson, Bernt
TITLE Frequency Words and Frequencies: A Pilot Study on Relations Between Differently Anchored Scales. Didakometry; No. 44, November 1974.
INSTITUTION School of Education, Malmo (Sweden). Dept. of Educational and Psychological Research.
PUB DATE Nov 74
NOTE 16p.
JOURNAL CIT Didakometry; n44 Nov 1974

EDRS PRICE MF-$0.76 HC-$1.58 PLUS POSTAGE
DESCRIPTORS Measurement Techniques; *Rating Scales; *Response Mode; Semantics; Statistical Analysis; *Testing Problems; Test Reliability; Test Validity; *Transformations (Mathematics)
IDENTIFIERS *Scales (Measurement); Sweden

ABSTRACT

Subjects are asked to answer six questions, partly with a frequency and partly by marking a verbally anchored scale with five categories. Some univariate and multivariate analyses are performed to elucidate the relations between variables with the two different modes of response. Although there are similarities in results for the two types of variables they cannot be regarded as interchangeable. The frequency spread for a given category is often far from negligible. (Author)

FREQUENCY WORDS AND FREQUENCIES: A PILOT STUDY ON RELATIONS BETWEEN DIFFERENTLY ANCHORED SCALES

Bernt Larsson

Larsson, B. Frequency words and frequencies: A pilot study on relations between differently anchored scales. Didakometry (Malmö: School of Education), No. 44, 1974.

Subjects are asked to answer six questions, partly with a frequency and partly by marking a verbally anchored scale with five categories. Some univariate and multivariate analyses are performed to elucidate the relations between variables with the two different modes of response. Although there are similarities in results for the two types of variables they cannot be regarded as interchangeable. The frequency spread for a given category is often far from negligible.

Keywords: Measurements, transformations, stability

INTRODUCTION

A measurement procedure may be regarded as a process with three stages. First comes the definition of the concept, then the selection of the measuring instrument, and finally the allocation of numbers to the possible outcomes. In educational research the definitions are often quite loose. Because the definitions may be far from unequivocal, different researchers, who verbally profess themselves to the same concept, can construct dissimilar instruments. As a not very farfetched consequence, different results may then arise which are not easy to interpret.

While opinions can be rather different about selecting an instrument for a certain concept, a pseudo agreement often exists about the allocation of numbers. There is seldom much in the educational measurement procedure that prescribes how to match numbers with outcomes. Nevertheless most researchers use successive integers as their allocation rule. However, other, more or less opportunistic rules exist and are described in papers, found under such keywords as 'transformation'.

In some papers, Larsson (1973, 1974a, 1974b), I have treated the stability of results due to different scale transformations. This report intends to elucidate the variability on the second stage. I have chosen to compare frequency words with frequencies by asking people how often they do or experience certain things. They have responded to the questions both by giving a frequency statement (the number of days per year) and by (indirectly) marking a category of a verbally anchored scale (almost never, seldom, sometimes, often, almost always).

According to my point of view, the two types of response mode can imply differences in the very first stage. When you answer e.g. '150 days per year', it is your honest attempt to determine the frequency, but when you answer e.g. 'often' you also evaluate 150 by some frame of reference. So I maintain that one may, right from prior considerations, expect differences between the two scales, because they do not measure the same thing. But, judging from their discussions, some researchers apparently believe that they measure frequencies with the above type of verbally anchored scale.

THE STUDY

The subjects of this pilot study are 44 persons, most of whom work at the departments of education in Lund and Malmö. (Age between 20 and 70, 24 men and 20 women.) The six questions asked concern how often they 1) watch TV, 2) go to the pictures, 3) wake up rested, 4) have a headache, 5) are stressed and 6) feel expectant. The questions are expected to comprise different degrees of agreement as to what is asked for. I imagine that most people agree on what the activity 'go to the pictures' implies, but feeling expectant is probably a very subjective experience. The respondent has partly answered with the relevant number of days per year, partly given lower and upper limits (again with the unit 'number of days') for the verbal statements almost never, seldom, sometimes, often and almost always. The order of the questions has been random for every subject.

The subjects were instructed in the following manner for every question:

Consider the statement at the top of the paper and fill in the number of days per year, which is true for you. Remember, for instance, that once a week is about 50 days per year. Then go on to 'almost never' and give me your opinion about its frequency meaning for this statement. Which interval on the scale 'number of days per year' do you think is correctly described by 'almost never'? Write down the lowest and the highest possible number in your opinion. Then proceed with 'seldom' and do the same thing as for 'almost never', then 'sometimes' and 'often', and finally 'almost always'. Let the upper limit of a verbal expression be equal to the lower limit of the next expression. For instance, the upper limit of 'sometimes' equals the lower limit of 'often' Notice that the upper limit of 'almost always' cannot exceed 365.

For each question eight variables has been punched on cards. These are the frequency statement, the verbal category, the lower limits of almost never, seldom, sometimes, often and almost always, and the upper limit of almost always. Only six limits are coded, because the subjects were instructed to let the upper limit of a category be equal to the lower limit of the next category. The verbal category is coded 1(1)5 and indicates to what category the frequency statement corresponds. (For some cases a statement is equal to a limit. The frequency is then randomly assigned to one of the two possible categories. Frequencies as extreme as or more extreme than the lowest and highest limits are assigned to almost never and almost always, respectively.) It can be added that nonresponse does not exist.

The main purpose of the study is to compare frequencies and categories. This is performed in several ways. For instance, frequency statements are correlated with categories, partly when the latter are coded 1(1)5 and partly with an optimal coding which maximizes the correlation. One may further ask if the correlation structure of the six frequency statements is the same as for the six verbal variables. Also, we will investigate if there are any relations between the frequency statement and the limits for a question and if a limit is determined in the same way for all six questions.

RESULTS

We first present some simple descriptions of the data for the six questions. Table 1 shows the means and standard deviations for the frequency statements and distributions of the categories. The frequencies have very different averages with relatively high spread (the coefficient of variation is in three cases above 1). The distributions seem to be in accordance with the means, for instance the rank correlation between the means and the median classes is about 0.97. If we let these classes characterize the group, we can say that they often watch TV, almost never go to the pictures, often wake up rested, seldom have a headache, and sometimes are stressed and feel expectant.

Table 1. Means (m) and standard deviations (s) of the frequency statements and distributions of the categories

	Frequencies		Categories				
	m	s	1	2	3	4	5
Questions 1	181.0	105.4	6	4	6	19	9
2	8.4	9.4	26	13	4	1	0
3	187.9	117.5	3	7	9	12	13
4	31.9	40.7	21	8	8	6	1
5	88.6	74.6	4	7	19	12	2
6	109.2	116.1	4	6	16	8	10

Another simple description is given in table 2, which presents the average frequency breadth of the categories for the six questions. With two exceptions, the breadth becomes larger and larger: 'almost never' has the smallest breadth and 'almost always' the largest one. This seems strange to me. In my opinion, 'sometimes' is the broadest category, followed by 'seldom' and 'often', while the extreme categories

have the smallest breadth, because they have a more specific meaning to me.

Table 2. The average frequency breadth of the categories

Questions	Categories 1	2	3	4	5
1	25.7	37.6	79.0	114.9	86.7
2	7.5	14.1	31.9	65.2	144.7
3	30.1	40.1	90.0	98.9	76.6
4	10.8	17.8	42.0	92.4	151.9
5	17.8	26.8	62.0	102.0	121.4
6	15.7	24.9	56.3	105.9	120.3

One possible hypothesis is that your discrimination is best for that part of the scale where you have your own position. This would mean that categories will be broader and broader the further they are from your position. While this seems to hold for question 2, the hypothesis is certainly not true e.g. for question 1 (compare with the first column of table 1). On the other hand, the ranks (over the questions) for the average frequency and category breadth are related: 0.94 for categories 1, 2 and 3, 0.66 for category 4 and -0.94 for category 5. However, tables 1 and 2 present rather coarse results, about which we should not speculate too much.

We now turn to the relations between the frequency statements and the verbal statements (or categories). For this purpose we can consult table 3 and start by looking at the first column. It shows the squared linear correlations between the frequency statements and the verbal categories when coded 1(1)5. Column 2 presents the maximal squared correlations for optimal coding of the verbal categories (restricted to monotonic transformations). The general procedure for obtaining maximal or minimal values of some indices of result is given in Larsson (1974b). In this simple case, the optimal scale values (also given in table 3) is any linear function of the average frequencies for the categories, provided that they are ordered in the same way as the categories. In fact, the difference between the values of column 2 and 1 can be regarded as a measure of the degree of nonlinear (but monotonic) regression for frequency on category, when the latter is coded 1(1)5.

Table 3. Squared correlations (r^2) between frequency statements and verbally anchored scales

		r^2	max r^2	Optimal scale values			
Questions	1	0.753	0.755	1.00	1.96	2.75	3.93	5.00
	2	0.457	0.532	1.00	2.08	2.08	4.00	-
	3	0.719	0.745	1.00	1.00	2.49	3.89	5.00
	4	0.607	0.627	1.00	1.57	3.21	4.61	5.00
	5	0.401	0.421	1.00	1.70	2.44	3.93	5.00
	6	0.584	0.689	1.00	1.25	1.52	3.40	5.00

Let us take a closer look at question 1, which has the highest correlation, according to both column 1 and column 2. The average frequencies for the categories are 10.7, 78.5, 134.2, 218.0 and 293.3 and for the coding 1(1)5 these means almost form a line. This implies that the correlation of column 1 cannot be substantially improved and we see that the optimal scale values are similar to 1(1)5. The plot reveals rather much heteroscedasticity: the standard deviations for the categories are 7.6, 23.8, 28.9, 69.4, 47.7, thus meaning that 'almost never' has a fairly strict frequency definition, while 'often' can comprise quite different frequencies. It can be added that the (weighted) average standard deviation of the above type is 52.2. (This is equal to the standard deviation about the best regression curve, which also can be calculated as the standard deviation of the frequency variable $x\sqrt{1 - \max r^2}$, the maximum being calculated without scale restrictions.) The lack of a perfect relation can also be exemplified by the fact that there are persons having the same frequency but who are distributed over three different categories. We may conclude that although the correlations of question 1 are pretty high, there is, at least for some categories, an equivocal frequency meaning.

The same pattern as for question 1 is on the whole also obtained for the other questions. Heteroscedasticity pervades all the plots and is typically asymmetric: with few exceptions 'almost never' has the smallest frequency spread and 'often' or 'almost always' has the largest one. We find again the same picture as for the frequency breadths of table 2, and again it puzzles me: why this asymmetry? Why is the (frequency) apprehension of 'almost never' so much more the same from person to person than that for 'almost always', and more the same for 'seldom' than for 'often'?

For all but two cases the frequency means are monotonically increasing. For question 2 'sometimes' has a slightly lesser mean than 'seldom' and for question 3 'seldom' has a lesser mean than 'almost never', but again the difference is small. This is reflected in the optimal scale values of table 3: the two categories for each question get the same value. (Notice that it has not been possible to determine the value for 'almost always' of question 2, since there is nobody in this category.) By comparing columns 1 and 2 of the same table, we see that an optimal coding can raise the correlation, although not very much. The greatest difference is obtained for question 6, which is 0.105, and perhaps not altogether negligible. The possibility of predicting frequency from the categories is different for the six questions, but in no case do I think that the categories can replace the frequency statements

Table 4. Some ANOVA results for frequencies and categories

		Freq.	Cat.
Questions	1	0.800	0.426
	2	0.491	0.365
	3	0.804	0.582
	4	0.469	0.598
	5	0.607	0.233
	6	0.510	0.472

Let us take the following simple case as another instance of the relations between frequencies and categories. For each question, define group 1 as those subjects who are below the median of the frequency variable and let group 2 be the rest. We then perform an ANOVA, partly with the frequency variable and partly with the categories (coded 1(1)5) as the dependent variable. Table 4 presents the result in the form av Hays' ω^2, which here is nothing but the squared point biserial correlation between the binary group variable and the dependent variable. I imagined that the difference between the ω^2 values (for the two dependent variables) should be smaller the higher the correlation between the variables (see column 1 of table 3) and that ω^2 of the frequency variable should be higher than that of the verbal variable. As you can see, neither the first nor the second supposition is correct, although the second one is nearly so. Further, the values of the two columns in table 4 are fairly independent of each other: the relative position of the effect size indices according to frequency is not repeated for the categories.

The above presentations of the relations between frequencies and categories have treated one question at a time. We can also treat them simultaneously and ask ourselves if the structure of the six frequency variables is the same as the structure of the six variables composed by the verbal categories, when coded 1(1)5. Several methods can be used to answer this question about structures. One way would be to perform a restricted canonical correlation analysis with coefficient vectors constrained to be the same for a pair of factors. As a computer program for such an analysis is not available to me, an unrestricted analysis has been made instead. This is shown in table 5, which gives coefficients (correlations between factors and questions) for the first three factors, communalities (h^2), proportions of total variance (second last row) and all six squared canonical correlations (last row). Most canonical correlations are rather high and the coefficients for frequencies and categories are quite similar for all factors. It can be added that the two sets are also similar in the sense of the Stewart-Love total redundancy measure (see Cooley & Lohnes, 1971, pp. 170-173): we have 0.665 for frequencies and 0.634 for categories.

Table 5. Canonical correlation analysis of frequencies and verbal categories: Structure matrices and squared correlations

	Frequencies				Categories			
	I	II	III	h^2	I	II	III	h^2
Questions 1	-0.690	0.462	-0.139	0.708	-0.642	0.519	-0.228	0.734
2	-0.251	-0.750	-0.275	0.702	-0.134	-0.628	-0.463	0.626
3	0.216	0.732	-0.607	0.951	0.256	0.781	-0.536	0.962
4	-0.496	0.013	0.625	0.637	-0.453	0.061	0.522	0.482
5	-0.401	0.136	0.513	0.442	-0.386	0.441	0.396	0.500
6	0.680	0.119	0.627	0.869	0.573	0.160	0.584	0.695
	0.242	0.224	0.252	0.718	0.196	0.249	0.221	0.666
	0.899	0.814	0.629	0.597	0.390	0.126		

Another way would be to ascertain whether independently performed analyses will reveal the same structure. In this case I have only made two modified component analyses, the modification consisting of varimax rotation of those components for which the eigenvalue exceeds 1. The analyses are presented in tables 6 a and 6 b. Table 6 a presents the coefficients (correlations), communalities and proportions of total variance (last row). From this table, factor I seems to be the same in both analyses but the other two factors can presumably not be regarded

as identical. However, table 6 b shows that the squared correlations between components from the two sets of variables is not very high. As for table 5, I refrain also here from interpreting the factors. (I am reluctant to 'dig' too deep with such a small sample.) We may say that these multivariate analyses have shown some similarities of structure, but not to the extent that frequencies and categories are interchangeable. Notice that canonical correlation analysis would not have been feasible, in case one researcher had used the frequency variables and another had used the verbal variables on different samples of subjects.

Table 6 a. Modified component analysis of frequencies and verbal categories: Structure matrices

		Frequencies				Categories			
		I	II	III	h^2	I	II	III	h^2
Questions	1	0.200	0.842	0.256	0.814	0.201	-0.871	0.172	0.829
	2	0.045	0.049	-0.921	0.853	-0.050	0.245	0.654	0.491
	3	-0.576	0.362	0.526	0.738	-0.384	-0.691	-0.303	0.717
	4	0.796	0.014	-0.151	0.656	0.831	0.156	0.022	0.715
	5	0.886	0.076	0.120	0.805	0.734	-0.355	-0.331	0.774
	6	0.199	-0.783	0.418	0.828	0.090	0.308	-0.793	0.732
		0.305	0.244	0.234	0.783	0.238	0.257	0.215	0.710

Table 6 b. Modified component analysis of frequencies and verbal categories: Squared correlations

		Categories		
		I	II	III
Frequencies	I	0.549	0.022	0.009
	II	0.017	0.528	0.139
	III	0.006	0.307	0.538

Finally, I will show some results about the limits. We first examine whether the determination of the limits is related to the frequency statement. For each question, the six linear correlations are quite small, and so are the squared multiple correlations, which are shown in column 1 of table 7. Consequently, the limits predict the frequency statements badly (as far as the relations are not essentially nonlinear, which I doubt).

Table 7. Analysis of the limit variables: Squared multiple correlations and cumulative proportions of total variance

Questions or limits	r^2	I	II	III
1	0.222	0.584	0.814	0.908
2	0.373	0.538	0.725	0.829
3	0.207	0.529	0.712	0.817
4	0.220	0.612	0.780	0.858
5	0.213	0.577	0.757	0.853
6	0.349	0.749	0.866	0.929

We may also ask ourselves if a limit is determined similarly for all questions. I have investigated this by performing a component analysis for each limit. The results of the analyses are partially displayed in the last three columns of table 7. They present the cumulative proportions of the total variance for the first three principal components. A sterotype determination would mean that the first principal component is sufficient to explain the variables. Although the upper limit of 'almost always' has a rather high value for the first component, one cannot on the whole assert that the determination is fully stereotype. More than one aspect is involved, e.g. the contents of the questions.

FINAL COMMENT

As was anticipated in the introduction, the general result of this study is that the frequency statements and the verbal categories do not mea ure the same things. However, to a certain extent they measure similar things. This correspondence is different for different questions and different analyses. For instance, the relation between frequency and category is not high for question 5 (see table 3), while I imagine that the canonical correlation analysis (see table 5) would yield the same interpretation of the factors for each set of variables.

A drawback of this study is the lack of any reliability estimates. One may maintain that the correlations of table 3 differ from 1.0 only because of unreliable variables. Although this may be true for question 1, I find it hard to believe concerning, for example, question 5. We know from classical reliability theory that the reliability coefficient of a

variable is not less than the squared correlation between this variable and any other one. From this it is reasonable to assume that the average coefficient is not less than the first squared canonical correlation (0.899). Moreover, for each question one finds the maximal squared correlation between limits varying between 0.784 and 0.866. This indicates some other possible lower limits of reliability. Altogether, it seems to me that unreliable variables cannot alone explain correlations are not 1.0. For instance, I find it hard to believe that answers to question 2 (go to the pictures) would be so much more unreliable than answers to question 1 (watch TV): The squared correlations of column 1 of table 3 are 0.457 and 0.753, respectively.

Verbally anchored scales are common in educational research and certainly not only confined to frequencies. Alternative, nonverbal scales may exist, such as 'the number of days per year' of this report. As another example we can take a five category scale, verbally anchored from 'very uncertain' to 'very certain'. The alternative scale can be produced by assessing subjective probabilities, for which a number of techniques exist (see e.g. Staël von Holstein, 1970). The latter scale is usually more expensive to use, but it is presumably more unequivocal. The verbally anchored scale has the drawback of being diffuse, e.g. 'very uncertain' may mean different things to different persons (variation in subjective probability). This is a dilemma: should we choose a cheap scale with dubious properties or a more dependable one which is more difficult to produce? Perhaps it would be easier to answer if we knew anything about the robustness of the verbal scales: how much of the results from analyses of a set of alternative scales are reproduced when analysing a set of verbal variables?

However, verbally anchored scales are not only used for measuring a certain dimension, but also - and perhaps more often - to evaluate points on that dimension, which in effect implies a new dimension. The two purposes are not always easy to distinguish between for a reader of a report. For instance, the verbally anchored scale of this study seems to have been used in both ways, that is, as a substitute for frequency and as an evaluation of the frequency. In some cases

the purpose is clear, e.g. a verbal scale used to assess the difficulty of an achievement test. (Nobody uses such a scale to measure the achievement itself, I hope.)

Also, verbally anchored scales can be devised in different ways. Suppose you want to know if your pupils think that mathematics is difficult. You can give them the statement 'Mathematics is a difficult subject' and ask them to mark one of several words (from 'fully agree' to 'fully disagree'). Another way would be to state 'Mathematics is ...' and then ask them to mark one word in a string ranging from 'very easy' to 'very difficult'. Although intended to measure the same thing, I am convinced that the two scales would not be perfectly related Suppose further that two ways are exactly the same, except for the verbal anchoring. Let us take the following alternatives (for judgment of a certain product):

very poor	terrible
poor	poor
neither poor nor good	fair
good	good
very good	excellent

Not even in this case do I anticipate a perfect relation.

One may sometimes wonder about the order of the verbal categories. For instance, anchoring a middle category with 'don't know' seems, at least under certain circumstances, objectionable to me. Some persons may mark this category when they are not able to answer the question, while for others the category constitutes a real indifference interval of the scale. I have also encountered scales with twodimensional verbal anchorings, e.g. from 'little of both A and B' to 'much of both A and B' I think that a minimal requirement of a scale is that the respondents have the same apprehension about the order of the categories.

This study has been particularly interested in some verbally anchored scales. However, the problem dealt with is not confined to this type of scales. The question whether the same results occur is pertinent in all collections of variables proposed to measure the same properties. I do not think it is necessary to have stability to the point of numerical invariance. 'Practical' invariance will suffice, that is researchers using different collections on the same or similar

measurement objects should reach the same conclusions, etc. We may take table 6 a as an example and ask: would two researchers - one of them using the frequency variables, the other using the verbal variables - interpret the result in the same way (provided that they use the same criteria of interpretation)? I am not sure.

REFERENCES

Cooley, W.W. & Lohnes, P.R. Multivariate data analysis. New York: Wiley, 1971.

Larsson, B. Obtaining maximal correlations by the construction of binary variables. Didakometry, No. 38, 1973.

Larsson, B. The influence of scale transformations: A study of factor analysis on simulated data. Didakometry, No. 40, 1974. (a)

Larsson, B. The stability of results: Some examples of the effects of scale transformations. Didakometry, No. 42, 1974. (b)

Staël von Holstein, C.-A.S. Assessment and evaluation of subjective probability distributions. The Economic Research Institute, Stockholm School of Economics, 1970.

Department of
Educational and
Psychological Research
School of Education
Malmö, Sweden

16

Abstract card

arsson, B. Frequency words and frequencies: A pilot study on relations between differently anchored scales. Didakometry (Malmö, Sweden: School of Education), No. 44, 1974.

Subjects are asked to answer six questions, partly with a frequency and partly by marking a verbally anchored scale with five categories. Some univariate and multivariate analyses are performed to elucidate the relations between variables with the two different modes of response. Although there are similarities in results for the two types of variables they cannot be regarded as interchangeable. The frequency spread for a given category is often far from ne-ligible.

Indexed:
1. Measurements
2. Transformations
3. Stability

Reference card

Larsson, B. Frequency words and frequencies: A pilot study on relations between differently anchored scales. Didakomet (Malmö, Sweden: School of Education), No. 44, 1974.