Some Problems of Data Analysis in Comparative Research

By

Klaus R. Allerbeck

Universität Bielefeld

Paper prepared for Conference on Comparative Social Research,
Ann Arbor, Michigan, October 3-8, 1977

What is special about data analysis in comparative research? Are there particular problems that occur if one analyses data from several nations? One might answer that problems of data analysis are indeed the same, whether data are collected in one nation or several nations - except that in comparative research they tend to be greater, and there are more of them.

I. Common models of data analysis

Comparative one-variable findings. The most common findings from widely published comparative research are one-variable findings. What % Americans, Germans, Swedes etc. (asked by interviewers) consider themselves "happy"? Or the wealth of nations may be measured, say, by GNP/capita and compared. The only trouble with such seemingly simple comparisons is the identity of what is being measured and counted (not to speak of thorny issues of statistical inference given different sampling procedures, refusal rates etc.). Even where translation is straightforward, the extent to which the shades of meanings may differ cannot be determined exactly: unfortunately, we may never find out whether all respondents who would classify themselves as "happy" would also classify themselves as "glücklich" and vice versa. Identical terms in English may be associated with different connations in the US and the UK. To take but one example: the mean thermometer score for sympathy towards the "Womens Liberation Movement" was 48.4 in the US, 35.4 in the UK, 37.8 in the Netherlands, 59.3 in Germany and 59.9 in Austria. Does this indicate that attitudes towards the feminist movement were most favorable in Austria? We used the terms "Dolle Minnas" in the Netherlands and "Bewegung zur Durchsetzung der Gleichberechtigung der Frau" in Germany and Austria. I would hesitate even to compare the US and the UK means in attempting to draw any substantive conclusions.

Are comparisons using objective data, such as income, really simpler? We can, of course, use official exchange rates to convert Dollars into Deutschmarks, or vice versa (or everything into Gold or SDRs, to use an international standard of the IMF). In that case do we have to take the exchange rate at the date of interview (given recent experiences on the foreign exchange markets, fluctuation is substantial)? Exchange rates, floating or fixed or pegged to something, reflect many things besides purchasing power. We could, of course, use exchange rates calculated on the basis of purchasing power, as defined by official agencies. The question then is, however, which market basket to use. To take a case in point: using US formula for the cost of living index, a German mark was worth 30 cents in 1975. Using the German cost of living index, a mark was worth 40 cents. The 1975 average exchange rate was 41 cents.

While it is possible without too much difficulty to compare individuals within each nation according to their score on some variable - happiness, female liberation, income or whatever -, to compare individuals from different nations on those variables takes so crucial assumptions that the outcome is rarely robust, unless our variables are limited to identically defined elementary events.

The examples given thus far are cases where issues of operationalization appear to be non-problematic. This is, of course, rare in comparative social research. Issues of operationalization tend to become very difficult in comparative perspective. The consensus in the literature maintains that it is pointless to pursue literal identity of questions or indices. Questions may be certified as literally identical by elaborate processes of translation and re-translation but could still mean very different things in different cultures. Obvious examples are Bogardus type social distance scales or measures of political participation. Scheuch has pointed out how different the social distance implications of terms such as neighborhood, are in different countries. Kim, Verba and Nie show in detail how measures of political participation have to be somewhat differently defined in various nations to represent the concept properly in each.

Functionally equivalent indicators of the same concept have the consequence that one-variable comparative findings are not meaningful (unless one is willing to make so many assumptions that the result of the comparison is free of empirical content). This problem is the same even for literally identical indicators, except of the most elementary type.
There are a few variables where one-variable findings are of interest and not invalidated by too generous use of assumptions. Examples are provided by demography. If one deals with birth, death and migration, cross-national comparisons are simplified.

Sophisticated one-variable comparisons. Valid one-variable findings may sometimes be misleading, however. Differences and similarities tend to be ascribed to the nations as characteristics. Mortality or fertility rates could appear as saying something about the respective nation as an entity. A plausible alternative explanation of any difference is to assume composition effects (age composition of a population in the case of mortality). An old tradition in demography and official statistics allows us to take care of composition effects: the age distribution of a nation is adjusted to some standard. Age-specific mortality rates are then combined and weighted to correspond to the standard age composition. Thus we have a special case of a one-variable finding: residual one-variable findings after removing the effect of another variable. This procedure is, of course, not limited to demographic variables. If we assume that willingness to participate in unconventional political activities can be meaningfully compared, we can apply a similar procedure. We found that the Dutch were far more prone to participate in these ways than the Germans. But the Dutch respondents tend also to be somewhat younger than German respondents (partly due to the fact that the Dutch age composition is indeed different from the German one, and partly due to Dutch interviewers being more successful interviewing younger respondents). So we can try to impose the German age distribution on the Dutch data in order to find out if there still would be a difference if age distributions were identical. Such an adjusted participation rate is purely hypothetical, of course. But the procedure allows to get rid of spurious national characteristics: if high Dutch protest potential was mostly due to the respondents being younger rather than due to their being Dutçh.

Problems of inference from sample surveys. It is worth noting that even the comparison of proportions or means across countries is not quite as simple as it seems, if the intention is to decide about similarities or differences between countries. It is possible, of course, to compute confidence limits for sample estimates of these parameters to see whether these overlap. Or differences between estimates could be tested for significance. As simple random samples are rarely used for the kinds of surveys utilized in cross-national research, the computation of confidence limits or significance tests is sufficiently complicated to make the use of computer programs for this purpose necessary. It would be wrong simply to look these statistics up in some table that is based on the assumption of simple random samples, while the actual sampling used was a multistage probability sample. A rule of thumb correction such as "the sampling error for a multistage sample of N= a should roughly correspond to the sampling error of a simple random sample of size b" will not be appropriate in most cases. The size of administrative units that are part of the multi-stage sampling process differs across countries; so does residential segregation according to social class; therefore, effects of clustering on sample estimates of variability will differ between countries. Consequently, the "design effect" (Kish) of the various national surveys will differ. Even with such precautions, however, the decision whether some countries should be considered different or similar with regard to some properties estimated by surveys cannot be made on the basis of such statistical procedures alone. In cross-national survey research, the impact of sampling errors is likely to be smaller than that of possible non-sampling errors.
Response rates in the various countries differ. If non-response is related to the variable being compared, significance tests would be grossly misleading. The composition of the interviewing

staff of fieldwork organizations in different countries may be very different. Effects of interviewer bias may very well operate differently in the countries where the fieldwork was done. It must be shown that such factors as non-response and interviewer bias do not invalidate the comparison of means or marginals across countries. Only then can statistical procedures for inferences about the populations be safely used. The situation is even worse if quota samples are being used instead of probability samples. While the same problems exist, there is no way to assess the magnitude of biases such as non-response.

These remarks should not be taken to mean that the comparison of survey results from different nations is of little value. The point is that such comparison require the judgement of the researcher, his knowledge of the subject and his careful analysis of possible sources of bias. Significance tests alone would not be sufficient as decision rules. Nor will a quick glance at the marginals be sufficient to decide whether certain countries are really different with regard to some variable.

Comparisons of relationships. In social research, more attention is paid to the associations among variables. There is a wide variety of measures of association and models of data analysis - linear, log-linear and otherwise. There is little point in discussing these in detail here, as models and measures are not necessarily tied to the origin of data, national, cross-national or whatever. Some comments are, however, in order, as cross-national designs force us to make explicit a number of assumptions usually taken for granted or ignored. Before doing this I will discuss a type of analysis that has been proposed especially for cross-national research: pooled analysis.

Pooled analysis. The point of this procedure is to make nation enter the analysis explicitly as a variable. There are no separate data files for each nation. Instead, all data are merged into one large data file.

An additional variable is added to each record of the individuals interviewed, representing the nation where the interview took place. Nation is a nominal variable. So it is transformed into a number of dummy variables for the purpose of multivariate analysis using the general linear model. Each dummy takes the value 1 if an individual is interviewed in a particular country and 0 otherwise. To avoid singularity of the correlation matrix, there will be n-1 dummy variables if there are n nations represented in a particular data set.

A regression analysis for pooled data from three nations with one independent individual-level variable would involve a regression equation such as

$$y = a + b_1x + b_2D_1 + b_3D_2 + e$$

where y is the dependent variable, x the independent variable, b_1 its regression coefficient and D_1 and D_2 stand for the dummy variables representing nations. D_1 takes the value i if a case is from country A and the value zero with regard to other countries. D_2 takes the value i if a case is from country B and takes the value 0 if it is not. Cases from country C have a value of 0 on both D_1 and D_2.
The analysis then proceeds in the usual fashion, the only difference being the presence of this set of dummy variables. This procedure is supposed to tell us more about the way in which nation is important. Does the dummy variable for a particular country enter the regression equation in a stepwise regression? Does a nation dummy variable load on a particular factor in a factor analysis? This procedure, then, should give the researcher information about the variables to which nation is related, the "importance" of nation as a predictor of some dependent variable, etc.

There is a crucial assumption of "pooled analysis". It is assumed that the relationship of the variables used for factor analysis, regression or correlation analysis are identical or at least very similar for all nations. Regression analysis in which nations are introduced as dummy variables assumes that the slopes of within-country regressions are parallel. The partial regression coefficients for independent variables other than country are identical or so similar that the expression of these relationships as average is meaningful. In the equation

$$y = a + b_1x + b_2D_1 + b_3D_2 + e$$

b_1, the regression coefficient for the first variable, is the same for all three countries.

The only difference between countries this model allows for is a difference in the intercept, which will be "a" for cases from country C, $(a+b_2)$ for cases from country A and $(a+b_3)$ for cases from country B. If b_1 is the same for cases from the countries A, B and C, this model is appropriate. If b_1 differs from country to country when the data are analyzed separately, pooled analysis would be misleading. It would return one value for b_1 which would not necessarily be characteristic of any nation, but just stand for a weighted average. An average may give a true impression or a false impression, as is well known and illustrated by the old joke of a layman having dinner with a statistician. The statistician has two steaks, the layman none. On the average, they consumed one steak each.

This assumption of parallel slopes may be correct. It may not be taken for granted. It may very well be that nation operates as a specifier variable in Lazarsfeld's sense, that the relationships between variables are conditional and the form and strength of the relationships depends on the particular country in which they are measured. If pooled analysis is not to present grossly misleading averages across countries that are characteristic for none of these countries, the analyst first has to analyze the data separately for each nation to check whether the assumptions necessary for pooled analysis are true. Since the separate analysis has to be performed anyway, it is hard to see that pooled analysis offers any kind of advantage for the analyst who wants to come to rapid conclusions. Instead of saving time and effort, pooled analysis done carefully adds costly and time-consuming steps to the analysis of cross-national survey data.

Perhaps more important than the feasibility of pooled analysis is its desirability as a model of cross-national survey analysis. What is the implied model of social processes that suggests pooling individual-level data gathered in several nations?

The implied model of social processes suggests that the crucial level of interest is that of the individual who is being interviewed. It makes no provision for influences of the social context of the individual other than those mediated through his own properties. Pooled analysis of individual level data assumes - and is appropriate only if this assumption is made - that the context, be it nation or region, has an impact only on the levels of values of particular variables, and not on the social processes which are reflected in the interrelationships of variables.

In the regression equation, b_2 and b_3 may be different from zero. That is, the means for countries A, B and C are allowed to be different, but b_1, the slope, has the same value for all countries. The relationship of the continuous independent variable and the dependent variable has to be the same for all countries, though it may take place on different levels. In other words, the slopes for all countries have to be parallel. This postulate of parallel slopes is a postulate of uniformity of social processes.

Whether social processes in different countries are the same or not is an empirical question. It can be answered by cross-national survey research. The more interesting questions, it seems, are those that seek to assess the impact of the characteristics of nations - such as party systems, systems of social stratification, etc. - on the relationship of variables.

It should be noted that in "pooled analysis", regression analysis is conceptually equivalent to Analysis of Covariance.

Formulating the question explicitly in ANOCOVA terms would probably increase the researchers awareness that crucial assumptions have to be tested or at least need to be considered.

It should also be clear that in order to be applicable, pooled analysis requires identical indicators everywhere, not just functionally equivalent measures. Otherwise mean differences that are expressed by the regression coefficients for nation dummies would not make sense.

Stepwise procedures. The help of the computer is indispensable for cross-national data analysis, of course. How much help we should demand for which tasks is an open question, however. Sometimes the computer is asked to do not only the calculations for a given model, but also to select a model or the variables that should enter it. Stepwise multiple regression or tree analysis are such examples.

The computer selects the independent variables that "work" from a set of candidate variables. It may do this according to some criterion (such as increase in R^2, adjusted R^2 or whatever). Usually, such procedures add one variable in each step. Does the order of entry into a regression (or a "tree analysis") indicate the relative importance of independent variables? It does, provided the independent variables are orthogonal.

If there is multicollinearity the order in which independent variables are entered says very little about the relative importance of particular independent variables. If two correlated independent variables could become part of this equation, the search procedure will select the one that minimizes the unexplained variance in this particular selection step; if the difference in predictive capacity of these two independent variables is small, chance fluctuations will determine which variable is selected. The second independent variable is not likely to be included in some later step, as its additional contribution to the explanation of variance will be small once

the other variable with which it is correlated has been included. Whether such procedures then shows that it variable is "important" compared to other possible independent variables depends to a considerable extent on chance fluctuations on which such automatic procedures tend to capitalize. The problem of multicollinearity does not invalidate stepwise regression procedures in general. If the purpose is only to find the regression equation providing the best fit for one sample, such procedures may be quite adequate. It should be noted, though, that forward selection procedures do not necessarily give the same results as backward selection procedures. If two variables are entered or dropped at a time the results can differ from results of the cheaper and more customary approach to enter or drop only one variable at a time. This problem becomes critical if the independent variables and the order in which they enter are interpreted in substantive terms.
Fortunately, there are now procedures that are capable of finding the "best" (according to some user-specified criterion) subset of regressions without actually calculating all possible regressions (which would be prohibilitively expensive). This technique is quite feasible for up to twelve independent variables. The main advantage of these procedures is that they make it clear that there is a choice of best-fitting subsets. This choice is not easy to make, in most cases, and cannot be delegated to the computer.

Where stepwise regression has been used in cross-national survey research, the results have been disappointing for this very reason: researchers did not realize that there was a choice, and that this choice was theirs. The results of stepwise regression tend to be orders of entry that differ wildly from country to country. Besides some very expected results (the most powerful independent variable to explain performance on a test in French is - surprise - whether the student had a course in French!), stepwise multiple regression does not produce simplicity but merely more confusion.

Marginal constraints on correlation coefficients. The most frequently used measure of a relationship is Pearson's product-moment correlation coefficient. Avoiding the pitfalls of pooled analysis, one might compute R's within each nation and then compare the strength of the correlation of, say, social class and political participation.

Both variables may be (and probably have to be) functionally equivalent measures, but not identical indicators. This presents no problem. Absolute comparisons of levels across nations are not made. Implicitly in the computation of r, both variables are standardized (zero mean, unit variance) within each nation.

If r is higher in country A than in country B, does this mean that the relationship of class and participation is stronger in A? The answer is yes, given assumptions about marginal distributions.

If we are not concerned with statistical inference, there is no need, of course, to require that the data come from bivariate normal distributions. We have to realize, however, that r can reach its upper limit (+1) or lower limit (-1) only if the marginal distributions of both variables are identical. This is of no great concern if our variables are additive indices combining a fair number of single items. Then we may be reasonably confident that marginal distributions do not disturb the result. The normal distribution should be approximated sufficiently well in this case.

We have to consider the effects of marginals, floor and ceiling effects on r, if we are dealing with single questions used as variables (such as education or income). The implicit definition of relationship is such that r does not consider a relationship perfect unless we can predict the value of Y_i for every observation i exactly, given X_i - which means identical marginal distributions.

We might, however, be interested in comparing the strengths of relationships, given that the marginals differ and might want to somehow remove the effect of differing marginals.

Given different marginal distributions, one might want to use measures of association that can reach unity in the absence of identical marginal distributions. Goodman and Kruskal's is such a measure. Note that the definition of "association" changes, as one moves from, say, Kendall's τ_b to Goodman and Kruskal's γ. There is no simple general solution here, but there is a need to spell out the implications and assumptions of such measures being used.

Adjusting marginals in contingency tables. To separate the components of a table into two parts: a) a nucleus of association and b) a set of marginals seems to be an old idea. Yule already adjusted margins of a table to display the nucleus of association, Mosteller (1968) drew attention to the procedure in his survey paper. It is of interest to note that Mosteller (1968) used an example of comparative research, Levine's comparison of British and Danish mobility tables (1967) to illustrate the method.

This procedure was used a good deal in the Octopus study. Its effect is to make the similarity of structures of relationships - like between levels of ideological thinking and education - more apparent. Here we imposed hypothetical equal marginals on all tables. One might, of course, take the marginal.distribution from country A and impose it on the table from country B.

Hierarchical log-linear models. The particular advantage of log-linear models (such as Goodman's) for comparative research is that they force the researcher to take all possible effects into account: margin effects, association effects, interaction effects. The model directs the analyst's attention to all of them, whereas regression analysis and thus path analysis focus on associations only and make assumptions about marginals (equality) and interactions (absent).

The application of log-linear models to cross-national data is not a simple, automatic matter either, of course. Significance tests provide consistent decision rules only if samples from different countries have the same size (frequently, they do not).

Data-analytic strategies. There is no question that data analysis in a comparative framework tends to become a rather complex untertaking. It may become even more problematic due to the nature of cooperative efforts. Here it is even less possible to keep a questionnaire from getting too long, as no one can be forced to give up one of "his" questions. Obviously, problems of data analysis follow because at some point selection has to be made and information has to be reduced.

The "natural" response of cooperative procedures seems to be the delegation of conceptual work to the computer. Unfortunately, as was pointed out, what is missing in conceptual clarity will definitely not come from stepwise regression, tree analysis or whatever. Throwing a massive data set into the multivariate grinder will not give results that are meaningful and can be communicated. Yule and Kendall's warning against "arithmetical enthusiasm not tempered by common sense" certainly applies to cross-national survey analysis.

Instead, analysis should be focussed on few important variables and their relationships. Theory and/or common sense have to be used in their selection. Data reduction by means of index construction is frequently a necessary first step.

In comparative research one has to strive harder than in research within one nation to achieve simple models. The need for simplicity is greater due to the cognitive complexity of dealing with several national data sets. It should be clear that simplicity of analysis models does not come easy. Usually, it only comes from a lot of effort.