On 28th August 2019 The Guardian website published an article headlined 'Children in UK least happy they have been in a decade, says report'[1]. The report that was cited was published by The Children's Society (2019) and presents findings from a number of sources including "... the latest figures for The Good Childhood Index from a survey of almost 2,400 children conducted in June–July 2019." (p. 17). This perhaps seems to be a large sample size, but the research as a whole has "now included over 37,000 children aged 8 to 17" (p. 14). Such large scale studies are enabled by the rationalisation of data collection and analysis that are facilitated by the survey approach allowing a national story to be told about increasing unhappiness amongst children that is clearly of importance and of legitimate interest to a national newspaper and, indeed, to national government and non-governmental organisations. I do not intend to summarise the research or its findings beyond the gloss provided in the newspaper headline; it is certainly complex in terms of structure and technical in terms of methods used, but the report cited here is certainly readable for a lay audience (especially if they omit the footnotes): have a look. My point in introducing the study is to illustrate the key advantages of survey method, which facilitates large sample sizes, but also working in potentially large teams that may be distributed geographically and in time, again because of the rationalisation of methods of data collection and analysis.

Surveys on a smaller scale may also be conducted by researchers working alone or in small groups. Put simply, a survey involves a set of items, often inviting each respondent to choose between a number of predetermined—or 'pre-coded'—responses for each item. The same questionnaire is administered to each respondent verbally, online, or by post or telephone and the responses collated and generally subjected to statistical analysis. This is a very efficient way to carry out research. What is perhaps apparent is that the analysis of the setting cannot wait until after the responses have been received. Before data collection has even begun there is analysis to be done in constructing the questionnaire, that is, in formulating the items and pre-coding responses; decisions will also have to be made about sampling the population of the setting and about the administration of the survey instrument—the questionnaire—and about ethical issues. I shall start with the production of questionnaire items.

Deciding what to ask

A frequent area of interest for survey research is the measurement of attitudes to cultural issues, practices, technologies and so forth, 'attitude' might be defined as:

> ... 'the amount of affect for or against an object', in accordance with Ajzen and Fishbein ..., who argue that affect is determined by a person's beliefs and behavioural intentions and so attitude is sensory-cognitive in character. (Sanderson, 2008; p. 475)

Patricia Sanderson (2000, 2001, 2008) was interested in schoolchildren's attitudes to the aesthetic dimensions of dance, aesthetic as distinct from dance for the purposes of fitness or as entertainment. This defines the general **Setting** of the study, which was an area that had received little attention in educational research. As a sole researcher she might have

[1] https://www.theguardian.com/society/2019/aug/28/childhood-happiness-lowest-level-in-decade-says-report

carried out a small-scale interview-based study that would have enabled her to explore her subjects' feelings and experiences in depth. She wished, however, to describe a bigger picture and so opted to conduct a survey. A widely used strategy in the measurement of attitudes is to ask individuals to respond to statements concerning the topic of interest. Two contrasting statements about ballet that were used by Sanderson (2008; p. 476) were:

'ballet is just jumping around in a pair of tights'

'ballet can look so beautiful'

These and other statements were initially collected from groups of school students aged between 11 and 16 from six schools, Sanderson having prepared a video 'showing different styles and types of theatre dance' (2000, p. 93) as a stimulus for these discussions. The statements that were collected were reviewed by Sanderson herself and 'an expert judge' (ibid) for the purposes of clarification and the removal of ambiguities. The intention was to include the 70 statements that emerged from this process on a questionnaire to be given to schoolchildren who were to indicate their strength of agreement or disagreement with each statement on a 5-point Likert-type scale (Likert, 1932): strongly agree, agree, neither agree nor disagree, disagree, strongly disagree. Responses for each positive statement (such as the second example above) would be scored 5 for strongly agree, 4 for agree, 3 for neither agree nor disagree, 2 for disagree, and 1 for strongly disagree and the other way around for negative statements such as the first example above (see NOTE 2.1). The scores would be totalled to give an aggregate score for each item or 'measured variable'.

NOTE 2.1

The questionnaire was administered to 368 pupils. This resulted in a distribution of aggregate scores for the 70 measured variables. In principle, the correlation between attitudes and responses to other questionnaire items relating to, for example, gender (2 values) or social class (5 values) could be calculated. This would be a very complex analysis and in need of simplification. So the researcher makes what I am referring to as an **Epistemological** assumption that the 'variance' (see NOTE 2.2) within the distribution of attitude measurements could be largely explained by a far smaller number of 'common factors' that are presumed to constitute the structure of attitudes to dance. These common factors would be identified using a statistical process known as 'Exploratory Factor Analysis' (EFA). I shall not go into the mathematical details of this process, which, in any event, would usually be carried out by computer, but it is worth spending a little time on its general principles. The intention here is not to provide instruction to a level that will enable the reader to conduct EFA, but to present sufficient description of the approach that will enable them to make sense of research reports, such as those by Sanderson.

So we have distributions for 70 measured variables. I need to introduce a few terms in order to proceed. I will not be saying very much about each term, but don't worry: familiarity with the terms is important, a full grasp of their meanings is not necessary for present purposes. Firstly, the *variance* for each variable (see NOTE 2.2) is a measure of the spread of the distribution. The *communality* for a variable is a measure of the amount of its variance that is accounted for by all of the other variables. Communalities can be estimated using a

process called *multiple regression analysis*. These estimates for the communalities may be used as the starting values for the *extraction of factors*, using one of a number of alternative processes that proceed iteratively. Jason Osborne (2014) recommends using *Maximum Likelihood* extraction (ML), where the variables can be assumed to be distributed *normally*, which is to say their graphs are bell-shaped curves that are symmetrical about their means (eg the curves in the first diagram in NOTE 2.2), and *Principal Axis Factor* extraction (PAF) otherwise. Whichever method is used, the number of factors produced is the same as the number of measured variables. Since the whole point of the proces is to reduce the number of variables, this is not much of an achievement, so far!

NOTE 2.2

Each factor will have a correlation with each measured variable; this is called the factor *loading* for that variable. Squaring each loading and summing the results for a given factor gives the *eigenvalue* for that factor, which is a measure of how much of the total variance of the distribution is accounted for by that factor. The factors having the larger eigenvalues are worthy of greater attention by the analyst than those having smaller eigenvalues. As I have indicated, the purpose of EFA is to reduce the number of variables, the researcher will retain for the purposes of the analysis, those factors having the larger eigenvalues. Exactly where to draw the line is a decision that is made by convention. Commonly, those factors having eigenvalues greater than 1 are retained: this is referred to as the Kaiser rule. One alternative rule is to arrange the factors in order of their eigenvalues and plot them on a graph—called a 'scree plot' as shown in the diagram in NOTE 2.3. The factors to be retained are those to the left of the elbow in the graph. In the fictitious case shown in NOTE 2.3 two factors are retained. In this case applying the Kaiser rule gives the same result.

NOTE 2.3

Sanderson's initial analysis produced 8 factors, but the 70 measured variables did not cluster neatly around these factors. An 8-dimensional space is difficult if not impossible to imagine visually, so NOTE 2.4 illustrates a fictional 2-factor solution. The measured variables—shown by the points on the graph—are in two clusters, but these are not aligned with the factors, each of which load on each of the variables. So, it is not possible to offer a clear interpretation of each cluster, each representing a combination of the two factors. The solution to this is to perform a rotation of the axes (the factors) as illustrated in the second diagram in NOTE 2.4. Now, the measured variables cluster around the each of the two factors, so an inspection of the variables—which variables cluster around which rotated factor—will enable an interpretation of the factors. Sanderson's original eight factors accounted for 73.1% of the total variance of the measured variables. After rotation, the variables clustered around five of these and accounted for 82.9% of the 73.1% of the variance accounted for by the initial eight, which is to say, 60.1% of the total variance of the 70 measured variables.

NOTE 2.4

The final stage of the factor analysis was to test the internal consistency of the five surviving factors by calculating *Cronbach's alpha coefficient*, which is a measure of internal

consistency or how strongly the measured variables within each factor are related to each other. Setting an arbitrary lower limit of 0.6 resulted in one of the five factors being eliminated so that Sanderson was left with four factors or 'scales' for attitudes to dance, these were: ballet (10 items); dance (9 items); male dancers (7 items); and dance performance (6 items)—comprising a total of 32 items.

All of this work—the initial student group discussions, review of the outputs of these discussions reducing these to 70 items, the preliminary survey of 368 subjects, factor analysis of the results or this survey producing, initially, eight factors that were reduced to five after rotation and subsequently to four following calculation of the internal consistency of the factors—was preparatory to the main survey that was to follow. The extent, complexity and statistical sophistication of this activity may come as a surprise to anyone who may have thought that questionnaire design is a simple matter of putting together a collection of statements that seem to exhibit *face validity* as indicators of the concept to be measured. There are also other considerations such as the controversy regarding whether a Likert scale can appropriately be regarded as interval level of measurement (see NOTE 2.1) and it is perhaps important to note that this quantitative analysis—as with all research—involves interpretation, which is sometimes wrongly thought (by an ill-informed public) to be exclusive to qualitative research.

The 32 items, *pre-coded* as 5-point Likert scales together with other pre-coded items that included measures of social class and gender were to be administered as a *closed* questionnaire—ie one on which responses are pre-coded—by teachers to a representative sample of school students.

Deciding whom to ask

There are fundamental differences between the kind of survey research conducted by Sanderson and most approaches to qualitative research. One key distinction concerns the issue of generalisation. Sanderson's research aims to produce an analysis of young people's attitudes to the aesthetic dimensions of dance that is capable of generalisation from her sample—she clearly can't ask everyone—to a wider population, notionally, the population of 11-16-year-olds in England at the time of the research. In order to achieve this the sample that she draws needs to be *representative* of this population and, indeed, this is her claim (Sanderson, 2008). Much if not most qualitative research does not (or, at least, should not) claim to generalise its findings in the same way. Rather, qualitative research may generalise firstly through the generation of theory and/or through the accumulation of cases. Neither of these forms will enable prediction beyond the sample studied in the research, so what is its point? Well, not all research is about prediction. Think about this kind of research as analogous to the exploration of an unfamiliar geographical region, the results of which do not say very much directly about territories beyond the region itself. The findings may, however, expand our understanding of that is possible and, in doing so, develop the conceptual apparatus that we use to speak geographically. Of course, we may identify features that recall others discovered in previous studies enabling, perhaps, the explorer to recruit extant geographical language, always taking care to note discontinuities with and so enrich this language. Qualitative research can generalise in much the same way, but more on this in later chapters.

Returning to the quantitative approach, Sanderson wanted a sample that was representative of the population of 11-16-year-olds in England, which is to say, the sample should comprise distributions of the various characteristics that differentiate between members of this population such that the distribution of these characteristics in the sample was in equal proportions to that in the population. Obvious examples of such characteristics are: age, gender, socioeconomic background, educational background and achievement, ethnicity, family background, race, health, diet, fitness, physical appearance, experience with dance, media, sports (of different kinds), and so on. We don't have to think very hard to expand this list potentially indefinitely. In the final analysis, everyone is an individual, but we may feel that not all of the possible differences are relevant to one's attitude to the aesthetic aspect of dance, but which? If it were possible to identify the crucial categories, then we might attempt to generate a representative sample by setting a suitable quota for each characteristic. This, however, would be to ignore possible interactions between these characteristics so that, to be truly representative, the sample would have to match the population with equal distributions of each possible combination of characteristics. To hark back to my geographical metaphor, such a sample would possibly be the social research analogue of Borges' fictional map of a scale of 1 mile to the mile (1975 edn).

In experimental work where it is believed that the key characteristics can be identified, are independent, and manageable in number, then it is possible to construct a representative sample using this quota method and this method is frequently used in market survey research. In social research more generally, however, it is more usual to generate a sample by 'random' selection from the relevant population. A random sampling strategy—see NOTE 2.5—is defined as one in which each member of the population has an equal probability of being selected in the sample.

NOTE 2.5

Sanderson describes her sample as follows:

> The sample for analysis comprised a total of 1298 pupils, aged between 11 and 16 years, 44% males, 56% females, drawn from 19 mixed secondary schools located in the five main geographical areas of England. Each area included inner city, suburban and semi-rural schools. Sampling was purposive in that efforts were also made to include schools that offered GCSE dance (signalling some interest in the subject) among the 19, as well as those schools with a particular focus on the arts. The aim was to try to achieve a representative sample of the age group in terms of dance experience, awareness and interest, as well as gender and social class. Although ethnicity was not included as a variable, the procedures followed in the selection of the schools ensured that the final sample included children from various ethnic backgrounds. (Sanderson, 2008; p. 474)

The sampling strategy is described here as 'purposive', which means that particular criteria—mentioned in the extract above—were deployed in selecting respondents. This is not a random sampling strategy and, indeed, just how the 19 schools were chosen is not revealed here nor is the method that was used to sample within the schools. Nevertheless, Sanderson is claiming that her sample was representative of the relevant age group. Insufficient information is provided on either the sample or the represented population to enable a confident assessment of this claim to be made. As with all research, there is

necessarily a point at which the reader must place trust in the researcher: there is, not can there ever be any certainty in research findings.

The questionnaire was delivered by post to the schools and administered by teachers who were following instructions from Sanderson.

Handling the Results

The questionnaire results will consist of distributions of Likert scale scores for each of the four factors (ballet, dance, male dancer, dance performance) and these can be separated into distributions for each of the four categories of social class and again for each of the two genders as well as the distribution for the whole sample: seven distributions in total. The mean score and the variance can be calculated for each of these distributions. The question that can now be put to the results is as follows: are these distributions *significantly* (NOTE 2.6) different. Sanderson chose to address this question by performing a *two-way Analysis of Variance* (ANOVA) test, two-way because there are two *independent variables*, social class and gender, and there is one *dependent variable*, Likert scale scores for the four factors. In fact, the two-way ANOVA tests for three null hypotheses: i) there is no difference between the distributions for the different social classes; ii) there is no difference between the distributions for the two genders; iii) there is no difference between the distributions for the combined sample. ANOVA calculates the *F-statistic* for each case. The *F-statistic* is the ratio of the total varianc *within* samples to the variance *between* samples. This is a fairly complex calculation that will, in all likelihood be performed by a computer application such as SPSS-X. The *F-statistic* is compared with the *critical value* to determine the level of *significance*: the ANOVA application will also return the *p-value* (NOTE 2.6). Sanderson's results were *significant* at the 0.01 level for both social class and gender for all four factors. Only the ballet factor proved to be significant (again at the 0.01 level), however, indicating a link between social class and gender for the ballet factor only.

NOTE 2.6

Sanderson also conducted *t-tests* between pairs of social class distributions. The *t-test* compares the *means* of distributions in the same kind of way as ANOVA compares *variances*. Sanderson does not specify which *t-test* she used, but since the distributions for each social class had different *variances* and different sample sizes, she should have used the *unequal variances* or *Welch's t-test* rather than the more familiar *Student's t-test*.

Summary

I shall not engage in further presentation of Sanderson's, but I encourage you to read the research cited here. I wish, though, to have illustrated that survey research, carried out properly, is a highly complex and technical business that demands a high level of expertise at all stages of the study from the design of the survey instrument, through sampling decisions, statistical analysis and interpretation of results. This quantitative method does indeed involve decisions and interpretations, for example, in factor retention and naming, the latter constituting the conceptual structure of the attitude scales. As I have mentioned, there is also a decision to be made as to whether a Likert scale can legitimately be

interpreted as at the *interval level of measurement*. Statistical tests generally place requirements on the nature of the data to which they can be applied. *T-tests*, for example, are *parametric*, which assumes that the data form a *normal distribution*, but this requirement seems often to be treated rather loosely. Did Sanderson assume that the *variances* of her distributions had equal *variances* and use the *Student's t-test* or did she opt for the more robust *unequal variances* test?

More fundamentally, perhaps, just what does one think one is doing when conducting *Exploratory Factor Analysis*? Are we simply reducing the number of variables to produce a meaningful story, or are we claiming to identify an underlying structure to our measured variables or even an underlying structure to our setting? We might describe the former as an *interpretivist* position and the latter as *realist*. The distinction, in my view, is important only in respect of the marketing of one's work and the language that one choses to present it. Provided that one is clear about one's methods it is ultimately of no relevance whether one believes that these methods have constructed the outcome or that one has identified something fundamental about the social world other than to those who regard interpretations as fictions and only discoveries as worth attending to. But fictions are also possibilities!