1. Introduction

Measurement is the cornerstone of science and if social science is to justify its name it must aim for standards of measurement which bear comparison with those of natural science. In some cases there is little problem in meeting this requirement. For example, birth rates, life expectancy and average hospital waiting times are well-defined concepts, which can be directly measured from readily available data. However, there is another class of concepts which are regarded as quantitative and yet which cannot be directly and unambiguously measured. Quality of the environment is such an example. We use the term in social debate in a way that implies it is something which one can have more or less of. If pressed to justify statements of this kind we would point to a whole collection of directly observable quantities such as levels of atmospheric pollutants, water quality, noise levels, contaminants in food and so on. It thus appears that the term 'quality of the environment' is a shorthand for something to which a constellation of other observable variables are assumed to contribute. It is, in short, a collective property of a set of variables. The problem is how to extract, in some sense, and then to combine into a single measure what each variable is contributing. The statistical problem is to provide a theoretical framework within which this can be done.

This is not a new problem and work on it goes back to the beginning of the 20th century. In the educational and psychological fields the modern subject of psychometrics represents the most highly developed strand of this tradition. Recently there has been a growing interest in measurement in sociology and, especially, in the relationships which exist between such measures. What has been lacking until recently is a unified approach to this class of measurement problems. This has meant that development in various fields has failed to benefit from closely related work, with the result that there has been duplication of effort and, in one case at least, inconsistent conclusions have been reached.

The purpose of this paper is to draw attention to recent work on a unifying framework and then to show how its use illuminates a number of obscure and controversial topics in the measurement of social variables.
2. The Latent Variable Framework

To focus our thinking we take a very simple and familiar example from educational testing. Children vary in their ability to perform tasks such as basic arithmetic. This gives rise to the notion of 'arithmetical ability' as something which individuals have in varying degrees. In order to measure this ability tests are administered in which, for example, a child might be asked to do 20 simple sums which are scored as 'right' or 'wrong'. The total number correct seems to be a sensible summary measure of ability and it is commonly used in this way. On reflection we might wonder whether some other summary measure, such as the geometric mean, might not be better. Or, recognizing that the test items might not be of equal difficulty, whether some items should be given more weight than others. It is questions of this kind to which we might expect a theoretical framework to provide answers.

The essence of the approach is to add a random variable, representing the unobservable quantity, to the set of those corresponding to the observed variables (also known as indicators). The additional random variable is known as a latent variable. Once this is done the question of how to measure the latent quantity can be posed in terms of its conditional distribution. This approach, in rudimentary form, goes back at least to Dolby (1976) but is given in its most complete form to date in Bartholomew (1996) and Bartholomew and Knott (1999). Here we shall briefly summarize the basic essentials sufficient for what follows.

Let there be $p$ observable random variables denoted by $x' = (x_1, x_2, \ldots, x_p)$ and one latent variable denoted by $z$. (Much of what follows holds if $z$ is vector-valued and, later, we shall need this extension). An important feature of our general formulation is that the $x$'s and $z$ may be categorical, discrete or continuous or, in the case of $x$, a mixture. We shall henceforth use the term metrical to include continuous and discrete variables.

Since $z$ is unobservable the only distribution we can make inferences about is $f(x)$, the joint distribution of $x$. This may be expressed as

$$f(x) = \int f(x | z) dF(z)$$

(1)

where the integral is over the range of $z$. It is immediately clear that one cannot uniquely determine either $f(x|z)$ or the prior distribution of $z$. Hence we cannot determine the posterior distribution of $z$ which is given by

$$dF(z | x) = f(x | z) dF(z) / f(x)$$

(2)

We thus appear to be at an impasse but further progress could be made if $f(x|z)$ happened to have the form

$$f(x | z) = a(X, z)b(x)c(z)$$

(3)
where \( X \) is a scalar function of \( x \). In that case we would be able to write

\[
dF(z \mid x) = \frac{a(X,z)b(x)c(z)dF(z)}{b(x)\int a(X,z)c(z)dF(z)} = \frac{a(X,z)c(z)dF(z)}{e(X)} \quad (4)
\]

This would imply that the conditional distribution of \( z \) given \( x \) depended on \( x \) only through the scalar quantity \( X \). In that sense, therefore \( X \) would then contain all the information in \( x \) relevant to \( z \). It follows that any 'measure' of \( z \) should be a function of \( X \). If \( z \) is metrical one could predict \( z \) given \( X \) by \( E(z \mid X) \) or some other measure of location. If \( z \) were a binary \((0/1)\) variable coding two latent classes then \( E(z \mid X) \) would be the posterior probability of being in the class coded 1. In a certain sense \( X \) may be described as 'sufficient' for \( z \); like a sufficient statistic it contains all the information in \( x \) about \( z \).

For practical purposes the usefulness of the foregoing analysis depends on whether one can justify the choice of a model for which the factorization of (3) is possible. It turns out that this can be done in a wide enough range of circumstances to cover most practical needs. In particular if the \( x \)'s are conditionally independent, that is if

\[
f(x \mid z) = \prod_{i=1}^{p} f_i(x_i \mid z) \quad (5)
\]

and if \( f_i(x_i \mid z) \) is a member of the one-parameter exponential family \((i=1,2,\ldots,p)\).

The conditional independence of (5) is necessary because it implies that no latent variable, other than \( z \), is required to account for the dependencies among the \( x \)'s. If (5) did not hold we could infer that at least one other latent variable was needed. The membership of the exponential family, in addition to (5), then ensures the factorization given in (3). Since the Bernoulli, multinomial and normal distributions are all included within this family, categorical and normally distributed variables are included. We emphasize again that the \( x \)'s do not need to have the same distribution.

If \( z \) is categorical we have a latent class problem and the posterior distribution will tell us the probability that an individual with a given \( x \) will fall into the latent class indicated by \( z \). This represents what is often called a nominal scale of measurement. If \( z \) is continuous there is no empirical basis for anything higher than an ordinal level of measurement. This is because the calculation of any measure of location of the posterior distribution depends on the prior \( f(z) = dF(z)/dz \) which is, as we noted, indeterminate. We can only estimate it if we make some assumption about the form of \( f(x_i \mid z) \) and, even then, it may be difficult to estimate \( f(z) \) with any precision. Different choices of \( f(z) \) will lead to different choices of the function of \( X \) to be used. It may be shown that \( E(z \mid X) \) is a monotonic function of \( X \), whatever \( f(z) \), so that all \( f(z) \)'s will lead to the same estimated rankings of the individuals. This is, in fact, what we
should expect since if \( z \) cannot be directly observed there is no natural scale to calibrate it.

The formulation we have given includes virtually all existing models. As we have noted, if \( z \) is an indicator variable we have latent class analysis. When \( z \) is continuous and when the conditional distributions of \( x_i \) given \( z \) are normal we have the factor model. If the \( x_i \)s are binary the latent trait model emerges. Here we shall not look at these important special cases because our purpose is to emphasize results which apply to all members of the class.

3. Measurement Issues in a General Perspective

Here we take a number of measurement issues on which new light is thrown by the approach outlined above.

3.1 Factor Scores

There is a longstanding and highly controversial literature in the factor analysis field on what are called factor scores. A recent and illuminating example of the debate was initiated by Maraun (1996). The problem is that of how to measure \( z \); that is to find some function of the \( x_i \)s to 'estimate' or 'predict' \( z \). Within the framework we have adopted here the matter is entirely straightforward. After \( x \) is observed our knowledge about \( z \) is contained in the conditional distribution of \( z \) given \( x \). The value of \( z \) is indeterminate in the sense that there is no single value that we can assign to it. Rather our uncertainty is expressed by a probability distribution. The best we can do is to use some measure of location of the distribution as a score for \( z \). The expectation, \( E(z|X) \), is an obvious choice and this happens to coincide with one of the traditional factor scores derived by other methods. But, as we have seen, it depends on the arbitrary choice of a prior distribution. In many cases it is intuitively appealing to use \( X \) itself since it often turns out to be a linear combination of the \( x_i \)s.

It is interesting that there has been no such controversy in the case where \( z \) is categorical. This illustrates how little cross-fertilization there has been between the various branches of latent variable modeling. 'Measurement' here consists of identifying a latent class to which an individual belongs. This has always been done by finding the posterior probability distribution over the latent classes, which is precisely what our general strategy indicates. The general approach is not, however, restricted to these special cases. It works whatever kinds of variables are included among the \( x_i \)s and, incidentally, whether or not their conditional distributions belong to the exponential family.

3.2 Reliability

Since we cannot determine \( z \) precisely we need some way of indicating the uncertainty associated with our prediction. This is what a measure of reliability
is designed to do. If our knowledge of $z$ is conveyed by its posterior distribution, then our uncertainty about it can be measured by the posterior variance, $\text{var}(z|X)$, or standard deviation. As we conventionally take the prior to have unit variance it follows that $\text{var}(z|X)\leq 1$. Hence we may take 

$$r^* = 1 - \text{var}(z|X)$$

as a measure of reliability since reliability is complementary to dispersion. The smaller $\text{var}(z|X)$, the larger the reliability and conversely. Like the mean, $r^*$ depends upon the choice of prior but it may be shown that the dependence is slight if $p$, the number of items, is reasonably large. In practice it appears that the posterior distribution is usually close to normal, so the mean and variance together provide a complete description of the distribution. Further details will be found in Bartholomew (1996).

This measure is not the same as that which has been traditionally used in those parts of latent variable analysis which have used the concept. The usual procedure is to base the measure on the distribution of $X$ given $z$. Thus if $S$ is the chosen function of $X$ we would compare $\text{var}(S|z)$ with $\text{var}(S)$, the argument being that if $S$ is a good predictor of $z$ then fixing $z$ will reduce the variance of $S$. Consequently we could measure reliability by $\{1-\text{var}(S|z)/\text{var}(S)\}$. Like $r^*$ this would be 1 if $S$ determines $z$ exactly (when $\text{var}(S|z) = 0$) and is 0 if $S$ conveys no information about $z$. In the form we have defined it $z$ is unknown so the reliability could not be determined. Instead, therefore, we have to take the average value given by

$$r = 1 - \frac{E\text{var}(S|z)}{\text{var}(S)}$$

(6)

Our proposed measure, $r^*$, is a function of $X$ but this is known. In practice it turns out that the dependence on $X$ is slight except possibly for extreme values of $X$.

Rather remarkably the two measures $r$ and $r^*$ can be shown to be close and, in one important case, are identical. Nevertheless $r^*$ has a more convincing rationale in that it is calculated on the basis of what is known, namely $X$. For many purposes the posterior standard deviation $\sigma(z|X)$ is more directly interpretable and is, of course, equivalent to $r^*$. Here again the general approach serves to unify and generalize an important measurement concept.

### 3.3 The Adequate Treatment of Categorical Data

In much social research, especially arising from sample surveys, the observed variables are a mixture of continuous or categorical. Hitherto there have been two ways of handling this situation. One is to downgrade the level of measurement by reducing all metrical variables to categorical, often binary, variables. This provides a valid but less efficient method because information is
lost. The other method is to upgrade the categorical variables to metrical variables. This can only be done by adding an assumption which is, usually, that the categorical variables have resulted from categorizing a continuous variable by dividing its range into segments and merely recording into which segment a sample member falls. In other words a categorical variable is regarded as an incompletely observed continuous variable.

If we make the further assumption that all pairs of these continuous variables have normal bivariate distributions the product moment coefficients can be estimated as tetrachoric, polychoric or bi-serial correlations according to the level of measurement of the variables involved. This enables us to carry out a factor analysis although there may be technical problems if, for example, the correlation matrix happens not to be positive definite. Any attempt to compute standard errors will be invalid. Other problems with this approach are discussed in Lee and Pooη (1999) in relations to the software package LISCOMP. The new approach allows us to treat the data as they are, without loss of information on the one hand, or the introduction of arbitrary information, about the underlying distribution, on the other. Further details will be found in Bartholomew and Knott (1999, chapter 7) and Moustaki (1996).

4. Comparisons

In all of the discussion so far we have been concerned with constructing a scale of measurement for a single latent variable. Having constructed such a scale we may wish to compare different populations. In particular, to compare the mean values of \( z \) in two populations, say. This often arises in practice when comparing average levels of attainment in different schools or age groups. This poses formidable problems. We recall from Section 2 that any one-to-one transformation of \( z \) in (1) leaves \( f(x) \) unchanged. In particular any linear transformation of \( z \) will have this result. Hence the mean and the variance of the prior distribution is arbitrary and we are free to fix them as we will. As a matter of convention we used a standard normal distribution with mean zero and variance one. Our \( z \)-scores were therefore determined relative to that prior distribution. In computing \( z \)-scores for two populations, both sets are based on a population mean of zero, and therefore there is no means of comparing the populations since we have removed any existing difference before calculating the scores. If there actually is a difference between the population means, then that difference will be absorbed into the parameters of the conditional distributions \( f(x|z) \). If we are prepared to assume that the value of \( x_i \) observed for a given value of the unstandardized latent variable is the same in both populations, then any difference in the location of the populations will be reflected in the estimates of these parameters. Differences observed in them can then be interpreted as the effects of population differences.

Although this result is perfectly general it is most easily seen in the case of the linear factor model. We write this as
\[ x_i = \mu_i + \alpha_i Z + e_i \quad (i = 1, 2, ..., p) \]

where \( Z \sim N(n, \sigma^2) \) and \( e_i \sim N(0, \psi_i) \). The \( e \)'s and \( Z \) are mutually independent.

Suppose, now, we have two populations distinguished by the additional subscripts 1 and 2. Then let

\[ z_j = (Z - n_j)/\sigma_j \quad (j = 1, 2). \]

For the two populations we shall then have

\[
\begin{align*}
x_{i1} &= \mu_i + \alpha_i(n_1 + \sigma_1 z_1) + e_{i1} \\
x_{i2} &= \mu_i + \alpha_i(n_2 + \sigma_2 z_2) + e_{i2}
\end{align*}
\]

(7)

Our assumption that \( Z \) has the same effect on the \( x \)'s in both populations implies that \( \mu_i, \alpha_i \) and \( \psi_i \) are the same in both and so do not need suffixes. But when we fit the model to each population with a standardized \( z \) the new ‘\( \alpha \)'s' will be \( \alpha_i \sigma_1 \) and \( \alpha_i \sigma_2 \) which will hot be the same unless \( \sigma_1 = \sigma_2 \). Any difference is therefore indicative of a difference in population standard deviations. To get at the difference of population means we need to know that \( X \) for this model is

\[
X_j = \sum_{i=1}^{p} \frac{\alpha_i}{\psi_i} (x_{ij} - \mu_j) \quad (j = 1, 2)
\]

(8)

It is clear from (7) that

\[
E(x_{i1} - x_{i2}) = \alpha(n_1 - n_2).
\]

Hence if \( X_1 \) is significantly greater than \( X_2 \) we may infer that \( n_1 > n_2 \).

It appears, therefore, that comparisons can be made of the means and dispersions of two populations but only if we are prepared to make an empirically unverifiable assumption. Without that assumption any apparent population differences will be confounded with differences in the way the \( x \)'s are related to the latent variable in the two populations.

5. References

Received: December 1999