Using League Table Rankings in Public Policy Formation: Statistical Issues

Harvey Goldstein

Centre for Multilevel Modelling, University of Bristol, Bristol BS8 1TX, United Kingdom; email: h.goldstein@bristol.ac.uk

Annu. Rev. Stat. Appl. 2014. 1:385–99

First published online as a Review in Advance on November 8, 2013

The *Annual Review of Statistics and Its Application* is online at statistics.annualreviews.org

This article's doi:
10.1146/annurev-statistics-022513-115615

Keywords

league tables, institutional performance, multilevel models

Abstract

This article reviews the statistical models that underpin institutional comparisons on the basis of outcome measures for their students. These multilevel models are developed to levels of complexity that match the problems posed. The strengths and limitations of inferences from these models are explored with examples taken from education.

1. INTRODUCTION

One striking feature of public policy over the past three decades has been the growing utilization of quantitative measures of institutional performance in many countries to make judgments and allocate resources. The key feature of these systems is that the resulting rankings of institutions such as schools, hospitals, and police forces are published by government bodies and publicized by the media. These systems are distinguished from "intelligence systems" (Hood 2007) that also produce rankings. However, rather than publish their rankings widely, intelligence systems use their rankings to inform the institutions and those responsible for monitoring them. Such systems operate, for example, in the area of public transport and have also been described in an educational context (Yang et al. 1999). Intelligence systems also have advantages over published systems (Foley & Goldstein 2012) in that they minimize unwanted or "perverse" side effects such as "gaming" to improve ranking position. They also directly address the underlying issues of how institutional performance can be improved, rather than indirectly attempting to address such issues by exposing current performance to public scrutiny. Although the statistical issues associated with design and analysis are similar in many respects between these two systems, in this review I do not consider intelligence systems in any detail, but instead concentrate on a discussion of public rankings. I deal largely with rankings in the area of education, especially school education.

This review addresses the issues associated with the design of ranking systems and with the modeling and interpretation of the results of published rankings. Ideally, the evaluations of the effects of such systems would be discussed, but such evaluations, however desirable, are rare and typically not envisaged when systems are designed. Where attempted, however, these evaluations are noted. Section 2 sets out some basic concepts, and Section 3 addresses different application areas and technicalities. Some examples are then provided, and I conclude with recommendations and areas for further work.

2. CONSTRUCTING RANKINGS

To illustrate the process of ranking construction, consider the case of school education where data are available from individual students attending each school. These data may be generated from, for example, responses to a questionnaire seeking views about satisfaction with teaching or the results of test or examination scores. Typically, data are chosen to represent a particular time point or period, such as the age at which external tests for admission to higher education are taken. At its simplest, a ranking will be formed by calculating the mean value across students and then producing a ranked list of these means. There may be several rankings, for example, for different curriculum subjects; occasionally, rankings may also be averaged into a single index using weights, as is done for rankings of universities (Dill & Soo 2005). In such cases, rankings are based on aggregate measures made at the institutional level. For universities, these would include such things as reputation among peers and measures of total research output. I discuss issues associated with these measures below. Before doing so, I deal with cases where linked information is available on individuals within institutions as is often the case with schools.

Aggregate rankings have been subject to criticism on two broad counts. First, they fail to contextualize the results by taking into account factors over which institutions have little control but which nevertheless have a strong association with the results. A particularly important factor is a selective intake: For example, hospital units may have different case mixes; some have higher risk patients than do others. Another example includes schools that recruit high-achieving students who are expected to have higher test and exam scores irrespective of the quality of the schooling received. In such cases, the ostensible purpose of the ranking, namely to compare schooling quality, will be undermined. A number of authors have discussed this issue and shown how such contextual factors can be incorporated as covariates in a model-based, value-added approach (Bird et al. 2005, Goldstein & Spiegelhalter 1996, Lockwood et al. 2007). The following model captures the essence of such approaches and is elaborated as appropriate.

The basic data structure is that of a two-level hierarchy with students nested within schools (or patients nested within hospitals, etc.). The standard approach to describing such data is via a multilevel or random effects model as follows (Goldstein 2011):

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + u_j + e_{ij},$$

$$e_{ij} \sim N(0, \sigma_e^2), \quad u_j \sim N(0, \sigma_u^2); \qquad 1.$$

for simplicity, mutually independent normal random effects are assumed. For example, an examination score at the end of secondary (high) school for student i in school j is y_{ij}, and x_{ij} is a prior achievement test score designed to capture any selection by prior achievement. In the next section, I discuss this model further, but first, I describe how, with appropriate data, we can derive rankings.

Ignoring the possibility of any missing data, Model 1 is fitted to all those students in the sample, which may be the total number of pupils in each school-year group or cohort. Under the normality assumption, we can obtain parameter as well as posterior estimates for each of the random effects u_j. If maximum likelihood is used, these effects are the usual shrunken residuals, and the equivalent posterior estimates can be obtained from a straightforward Bayesian analysis, for example, using a Gibbs sampler with diffuse priors (Goldstein 2011, ch. 2) (see also **Figure 1**).

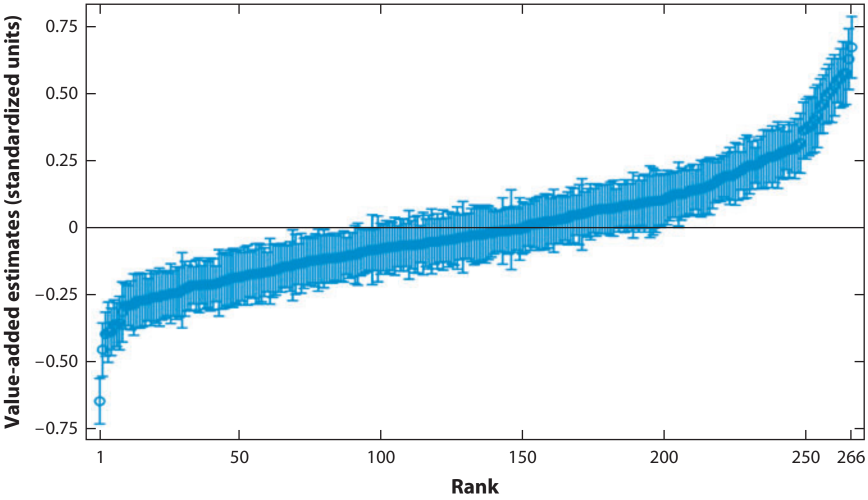

Figure 1

Ranking of a sample of 266 schools in England with a median cohort size of 190 (Leckie & Goldstein 2009). The response is a normalized examination score taken at the end of compulsory secondary schooling in grade 11 and adjusted using a test score taken prior to entry to secondary school as well as ethnic group and various measures of social and linguistic disadvantage. The vertical bars are conventional 95% normal confidence intervals, and approximately half the schools have an interval that includes the population mean. Comparison of two chosen schools in terms of whether their intervals overlap yields an appropriate interval length of approximately 0.7 times that of the intervals displayed (Goldstein & Healy 1995). Similar results can also be displayed directly in terms of rankings, rather than residual estimates, with corresponding intervals (Goldstein & Spiegelhalter 1996).

Such displays as shown in **Figure 1** can provide basic information about school effects that may be useful to inspectors seeking more detailed follow-up of individual institutions. They have also been advocated for use by students' parents as an aid in choosing schools. However, such use is problematic because what is really required is a prediction of school effects several years ahead, although this adds extra uncertainty that makes any statistical separation very difficult (Leckie & Goldstein 2009) and is further explored below. Alternative formulations have been proposed on the basis of a measurement of the effects of individual teachers during a period of schooling (see below).

The use of such adjusted or value-added models is intuitively appealing as an improvement on the use of unadjusted means and has become reasonably well established in the area of schooling. In other areas, however, this approach is more problematic. For university rankings (see, for example, Hazelkorn 2011), it is typically difficult to find sensible adjustment variables, and measuring and linking these together for individuals within universities is also difficult. Furthermore, the purpose of such rankings is often unclear. If the intention is to provide choice for undergraduate applicants, then a measure of degree outcome or satisfaction would be appropriate, but such measures tend to be incomparable both across universities and across disciplines. If the intention is to provide overall measures of research performance, additional difficulties arise with defining these measures in relation to factors such as citation indices, about which there seems to be little consensus (Spiegelhalter & Goldstein 2009). Another difficulty lies with the typical requirement of combining individual indicators into a single measure for presentational purposes, which involves decisions about which weights to use in such a process (for a more detailed discussion, see Foley & Goldstein 2012; also see below).

3. THE IMPLEMENTATION OF ADJUSTED-RANKING MODELS

Within model implementation, the primary area of concern is with the nature of the criterion being used to judge institutional performance. This issue is discussed in relation to university rankings as well as schooling. For the latter, despite being common, the use of test scores raises objections as schools may either overconcentrate on improving test scores at the expense of broader educational measures or promote some students at the expense of others to improve their league table position. Likewise, when measuring aspects of policing, the choice of outcome is debatable. Although such debates are important, they are additional to the technical concerns of this review and are not pursued here (but see Foley & Goldstein 2012).

When adjustments are incorporated into ranking models, they are typically based on measures taken at an earlier time or set of times. They are distinct from scaling adjustments that may be used, for example, to measure university research output per academic, where a measure of total output is scaled by an estimate of the number contributing to it. Such an estimate may not be straightforward to compute, but this is a measurement rather than a modeling problem. Nevertheless, a university dropout rate may need to be adjusted for intake measures to avoid, for example, any manipulation of the results owing to the exclusion of students from underprivileged backgrounds who are more likely to drop out for financial reasons. In the following sections, I consider several relevant issues, including the adequacy of prior measures used, student mobility, differential school effects, missing data, endogeneity, and other aspects of model misspecification.

3.1. Prior Information

Most league table rankings, whether of schools or other institutions, utilize a single measure of prior performance to adjust for selection factors, whether purposeful or haphazard. However, in the case of secondary schooling, information about prior school attended and achievement during that period of school will generally change inferences (Goldstein & Sammons 1997). Nevertheless, other authors (Goldstein et al. 2007) suggest that rankings in primary (elementary) schools are relatively unaffected when a sequence of prior achievement measures, as opposed to just one measure, is used.

3.2. Moving Across Institutions

Most rankings are published using students' school membership at the time at which the outcome measure is taken. In practice, however, many students will change schools throughout their period of schooling so that the contributions of all schools attended should be taken into account. When this is done (Goldstein et al. 2007), the variation attributable to schools generally increases but, again, seems to have little effect on the rankings.

To take account of such mobility, we may use a multiple-membership model that extends Model 1. For simplicity, the following model includes just two schools between which students can move:

$$y_{i\{j_1 j_2\}} = \beta_0 + \beta_1 x_{i\{j_1 j_2\}} + w_{1ij_1} u_{j_1} + w_{1ij_2} u_{j_2} + e_{i\{j_1 j_2\}}$$

$$w_{1ij_1} + w_{ij_2} = 1 \qquad 2.$$

$$e_{ij} \sim N(0, \sigma_e^2), \quad u_j \sim N(0, \sigma_u^2).$$

Model 2 states that the random effects contribution to the response is a weighted combination of the random effects associated with the schools attended. For several schools, contributions come from one or more with associated weights. The weights have to be chosen, for example, such that they are proportional to the time spent at each institution, and in some cases, these weights can be estimated (see discussion below). In practice, sensitivity analyses are carried out with different weighting functions, from which the one that produces the best fit may be chosen, for example, as judged by the DIC (deviation information criterion) statistic in a Bayesian analysis (Spiegelhalter et al. 2002). One consequence of Model 2 is that the total level-2 variance has the form $\sigma_u^2 \sum_h w_{ih}^2 \leq \sigma_u^2$, so that ignoring mobility will lead to an underestimate of the level-2 variance. Further details on fitting such models are given in Goldstein (2011, ch. 13).

3.3. Moving Within Institutions

In the case of schools, especially secondary or high schools, longitudinal data may be available for students at the end of each year of schooling and may be attached to different teachers who will provide separate effects on outcome measures. Multiple-membership models such as Model 2 can be adapted for this situation (Lockwood et al. 2007). Such a model can be written as

$$y_{tij} = (X\beta)_{tij} + u_{tj} + \sum_{t^* < t} \alpha_{tt^*} u_{t^*} + e_{tij}, \quad t = 1, \ldots p, \qquad 3.$$

where u_{tj} is the contribution from the current teacher at the end of year t, and u_{t^*} represents the contributions from all of the different teachers prior to year t. The covariates (X) can include prior attainment as well as socioeconomic indicators and school-level variables. Model 3 is known as the general persistence model ($\alpha_{tt^*} < 1$): A special case is the "complete persistence" model where $\alpha_{tt^*} = 1$, indicating that each previous teacher has the same effect on a future outcome irrespective of how far ahead year t may be. The level-1 (occasion) residuals for a student are also correlated across occasions. Enough movement among groups of students is also assumed from year to year to enable identification of the model parameters.

To illustrate this model in a simple case with just three occasions, we have

$$\begin{aligned}
y_{1ij} &= \beta_{10} + u_{1j} + e_{1ij},\\
y_{2ij} &= \beta_{20} + u_{2j} + \alpha_{21}u_{1j} + e_{2ij},\\
y_{3ij} &= \beta_{30} + u_{3j} + \alpha_{32}u_{2j} + \alpha_{31}u_{1j} + e_{3ij}.
\end{aligned}$$

This basic model can be extended in a number of ways.

- If we have several teachers for each student in any given year, then we can introduce standard multiple-membership weights for each student, thus adding to 1.0 as in Model 2.
- The pupil-level residual covariance matrix can be structured as a function of time to reduce the number of parameters, for example, $e_{tij} = e_{0i} + e_{1i}t + \delta_{tij}$.
- Extra levels, such as those of school or cross classifications, can be introduced.
- Generalized linear models can be used.
- We can accommodate the same teacher in more than one year by modifying the indicator matrix for the teacher random effects.
- A multivariate extension is possible whereby we can model outcomes in more than one curriculum subject.

However, in any given data set, we may be missing the teacher identification for some students for some years. In this case, one approach is to assume that all those students with a missing teacher identification in any given year belong to a new pseudoteacher and to then sample accordingly. An alternative is to assume that the true teacher is one of those for whom data are available and use weights similar to multiple-membership weights corresponding to the observed distribution of students among these teachers.

These models are used in many US state education systems to evaluate teachers. An introduction to a series of papers discussing their strengths and weaknesses can be found in Amrein-Beardsley et al. (2013). A key issue is that the confidence intervals associated with any one teacher tend to be large and sensitive to the assumptions of the model (Lockwood et al. 2007).

3.4. Differential Effectiveness

So far, I have assumed a simple (random) effect for an institution. There is now a great deal of evidence, at least for schooling, that the institutional effect will also depend on a characteristic of the individual, for example, whether girl or boy or whether an initial low or high achiever (Nuttall et al. 1989). Such characteristics can be incorporated using a random coefficient model such as the following in which we allow different random effects for boys and girls, parameterized in terms of an overall school effect as well as for the boy-girl difference:

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + u_{0j} + u_{1j}x_{2ij} + e_{ij}, \qquad 4.$$

or alternatively

$$\begin{gathered}
y_{ij} = \beta_{0j} + \beta_1 x_{1ij} + \beta_2 x_{2ij} + u_{0j} + u_{2j}x_{2ij} + e_{ij},\\
\beta_{0j} = \beta_0 + u_{0j}, \quad \beta_{2j} = \beta_2 + u_{2j},\\
\begin{pmatrix} u_0 \\ u_2 \end{pmatrix} \sim N\left(\begin{matrix}0\\0\end{matrix}, \begin{matrix}\sigma^2_{u0} & \\ \sigma_{u02} & \sigma^2_{u2}\end{matrix}\right), e_{ij} \sim N(0, \sigma^2_e).
\end{gathered}$$

When such a model is introduced, comparisons may alter considerably. Accordingly, Yang et al. (1999) showed that rankings of primary schools can be very different for initially high and low achievers. They also demonstrated that misspecifying the model by ignoring a random coefficient for initial test scores resulted in a spuriously high correlation between the adjusted and unadjusted school effects.

3.5. Endogeneity and Model Misspecification

Some forms of model misspecification can lead to endogeneity, whereby one or more covariates in the model are correlated with the random effects, in particular with the level-2 residuals. For example, in Model 4, if the term $u_{1j}x_{2ij}$ is omitted, then this component of level-2 variation will be absorbed into u_{0j}. As a result, the random effects will, in part, depend on x_{2ij}, thereby inducing a relationship with the covariate in the fixed-effects part of the model. Such an omission will usually lead to misleading inferences for institutional rankings. In general, reporting and basing decisions on estimates, when available, from the full model may be preferable.

In some cases, however, we may wish to estimate the parameters of the simpler model given by Model 1, thus effectively marginalizing Model 4 over the random coefficient effects. By default, such marginalization is typically carried out with respect to the observed sample, assumed to be representative of a suitably defined population. Using maximum likelihood, which assumes Model 1 is correct, researchers can easily show how a standard estimation will provide biased estimates for the required marginal model.

The issue, however, is not straightforward because we may choose to marginalize with respect to a different population structure—for example, by standardizing the within-school distribution of gender to adjust for different proportions of male and female students that are considered irrelevant when making comparisons. In such a case, we may choose to marginalize with respect to having equal numbers within each school in an effort to apply equal weights to the male and female effects for each school. In another example, the coefficient of prior achievement may be varied randomly across schools, and marginal estimates may be desired for a standard distribution of prior achievement so that schools could be compared directly after adjusting for prior achievement. To carry out such marginalizations, for example, using bootstrap methods, we first need to fit the fully specified model. Thus, alternative methods such as GEE (generalized estimating equation) (see Hubbard et al. 2010) that fit marginal methods directly, effectively using the observed sample structure, are generally not appropriate. Ultimately, endogeneity is part of a general concern associated with model misspecification, rather than a narrower concern associated with biased estimates for a particular marginal model.

3.6. Measurement Error

Measurement or category misclassification errors are usually present in measures used as both responses and predictors in models such as Model 1 and more complex models. These errors are rarely taken into account, although they will typically result in biased parameter estimates if ignored. A discussion of the effect of measurement errors is given by Ferrão & Goldstein (2012), who showed that the effects of such errors in one data set can be large if not properly adjusted for.

3.7. Missing Data

A standard procedure for handling missing data is via multiple imputation (Rubin 1987), wherein missingness is assumed to be random, at least conditionally on the basis of other measured variables. In essence, the use of multiple imputation relies on the ability of researchers to sample (impute) from a posterior distribution estimated for the value under consideration. Missing-data values arise essentially in two ways: First is the usual way when the values of a measurement, such as a test score, are unobserved. Second is when an identification, for example, of a teacher or school, is unknown. The latter example is discussed above for the case of missing teacher identification wherein an imputed value for the (unknown) teacher may be obtained according to assumptions regarding the reason for the missing identification (see Section 3.3). Although Hazelkorn (2011, ch. 13) discussed estimation for such models, this area has been little explored and further empirical data would be useful.

3.8. Multivariate Models

The extension to multiple outcomes of interest is relatively straightforward, and apart from computational considerations, few new issues occur. Jointly modeling several outcomes at multiple levels has the advantage that the relationships between different types of institutional effects can be studied; such study may be important for certain kinds of judgments. Computationally, a new issue arises when the outcomes are of different types, such as a mixture of binary, normal, and ordered responses. In such cases, a latent normal model can be fitted for which ordered and binary variables are treated as deriving from underlying normal variables, extending the simple probit analysis model. This model can also be extended to unordered categorical variables and count data (Hazelkorn 2011, ch. 7). These models, therefore, allow the joint modeling of data such as exam passes, ordered exam grades, and continuously distributed test scores alongside, for example, attitude ratings (for an example of a simple joint model exploring the results of mathematics and English examinations, see Goldstein et al. 1993).

3.9. Modeling Where Outcomes Are Measured at Higher Levels

So far, I have described cases where the outcome of interest is measured at the lowest level of the data hierarchy, such as students in schools. Now consider, for example, the case of policing, for which the interest lies in comparing police forces or policing areas in terms of crime rates of different types. Although the rate for an area is essentially an aggregation of individually reported incidents, often few, if any, measurements at the level of the individual incident are relevant for adjustment purposes. Thus, if the interest is in how efficiently a police force is tackling burglary, taking account of the vulnerability of the properties where burglaries were reported may be relevant because this measurement may differ among areas. Yet, such information may often be available only in aggregate form at the area level. We may still use aggregate-level variables for adjustment, but in general, these will be less efficient. In addition, if the number of areas is small, care will be needed to avoid overfitting as a result of a large number of correlated covariates.

In other cases, data may be defined only at the institutional level. For example, rankings of reputation for universities are typically based on responses from individuals asked to rate institutions (Baty 2010). In this example, we assume a number of responses to be aggregated for each institution. Unlike the case of school examination results, however, we do not have independent responses across institutions because each rater provides a measure for each university, for example, on a simple rating scale. Formally, this can be modeled as a cross classification of raters by institutions, and a simple model, assuming normality, can be written as

$$y_{\{j_1 j_2\}} = (X\beta)_{\{j_1 j_2\}} + u_{j_1} + u_{j_2} + e_{\{j_1 j_2\}}, \qquad 5.$$

$$e_{\{j_1 j_2\}} \sim N(0, \sigma_e^2), \quad u_{j_1} \sim N(0, \sigma_{u1}^2), \quad u_{j_2} \sim N(0, \sigma_{u2}^2).$$

In this model, adjustment variables (*X*) may be obtained from the raters or measured at the institutional level. If uncertainty intervals are required for such rankings, then these intervals need to account for the data structure as derived from Model 5 and will typically be larger than naive estimates that treat the responses as independent.

Adjusting for rater characteristics will be especially important because such rankings are often derived using convenience samples such as those derived from databases of journal authors. Thus, for example, larger institutions will tend to produce more authors and therefore have greater representation in the samples used. In this case, careful consideration needs to be given to the possibility of weighting respondents to obtain what may be considered a representative sample, the determination of which remains a matter of debate. Additionally, the samples obtained in such surveys often have response rates as low as approximately 10%, which raises additional concerns about bias (P. Baty, personal communication).

4. EXAMPLES

I now explore two examples that applied some of the above models and show how the results are relevant within an educational accountability framework. The first example is a data set on a cohort of 5,748 students in 66 secondary (high) schools in inner London. The students entered their secondary schools in 7th grade (ages 11 and 12 years) and took school-leaving examinations in 11th grade (for further details, see Goldstein et al. 1993). Cases with any missing data (35%) were excluded from this analysis, and the authors report that this did not suggest any serious biases among those cases with complete data. Ideally, a more efficient analysis would use multilevel multiple imputation procedures (Goldstein et al. 2009), but for simplicity of illustration, I present only the complete case analysis.

Because secondary schools differ in the mean academic achievements of incoming students, I employ a value-added model for which the principal variable used to adjust for intake achievement is the reading test score provided by the London Reading Test (LRT) taken during the year before entry. The basic aim of the analysis was to explore the extent to which schools can be held accountable for the examination results of their students after adjusting for selection factors (prior achievement scores). The analysis also looked at differences between curriculum subjects. The response (y) and the LRT score (x_1) were both transformed to have standard normal distributions. Other predictors were student gender (x_2 with boys as the reference category), school gender [girls school (x_3), boys school (x_4), and mixed-gender school (the reference category)], and school denomination [Church of England (CE, x_5), Roman Catholic (RC, x_6), and state maintained (the reference category)]. In addition, the analysis included the results of a verbal reasoning test that is taken prior to entry where students are grouped into three categories representing approximately 25%, 50%, and 25% of the distribution: The three categories are VR1 (x_7), VR2 (x_8), and VR3 as the reference category.

The final fitted model with the normalized examination score as the response is as follows:

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_1^{(2)} x_{1ij}^2 + \beta_2 x_{2ij} + \sum_{k=3}^{4} \beta_k x_{kij} + \sum_{k=5}^{6} \beta_k x_{kij} + \sum_{k=7}^{8} \beta_k x_{kij} + u_{0j} + u_{1j} x_{1ij} + u_{2j} x_{5ij} + e_{ij},$$

$$\begin{pmatrix} u_{0j} \\ u_{1j} \\ u_{2j} \end{pmatrix} \sim N(0, \Omega_u), \qquad e_{ij} \sim N(0, \sigma_0^2 + \sigma_{01} x_{1ij}). \qquad 6.$$

Thus, at the school level we have an overall (intercept) effect that varies across schools. This effect includes a linear component of the relationship with LRT that varies across schools and a difference between the VR1 and combined VR2 + VR3 categories that also varies across schools, and these variables are assumed to have a three-variate joint normal distribution. At the pupil

Table 1 Analysis of total examination score

Fixed coefficients	Estimate (standard error)		
β_0	−0.53		
β_1	0.37 (0.02)		
$\beta_1^{(2)}$	0.035 (0.008)		
β_2	0.13 (0.03)		
β_3	0.07 (0.06)		
β_4	0.09 (0.07)		
β_5	−0.04 (0.13)		
β_6	0.20 (0.06)		
β_7	0.70 (0.04)		
β_8	0.31 (0.03)		
Between-school variation (Ω_u) (correlation)			
	u_0	u_1	u_2
u_0	0.055		
u_1	0.012 (0.75)	0.0046	
u_2	0.013 (0.40)	0.009 (0.97)	0.019
Between-student variation			
σ_0^2	0.55		
σ_{01}	0.046		

level, we assume a normally distributed residual with a variance that is a linear function of the LRT score. **Table 1** gives maximum likelihood estimates for this model (for details, see Goldstein 2011, ch. 2).

As expected for the fixed coefficients, we see large and statistically significant (at the 5% level) effects for the LRT score with a quadratic relationship indicated; for the VRT category; for gender, with girls on average scoring higher than boys; and for attendance at a Roman Catholic school. At level 2, statistically significant variation is associated with the LRT score ($\chi_3^2 = 24.7$, $P < 0.001$) and the VR1 category ($\chi_3^2 = 11.0$, $P = 0.008$): In each case, the null hypothesis is that the variance term and two associated covariances are zero. The chi-bar test statistic is used (Goldstein 2011, ch. 2) because the variance term is constrained to be nonnegative. At level 1, there is a significant positive relationship of the variance with the LRT score ($\chi_1^2 = 66.0$, $p < 0.001$). Based on these model estimates, **Figure 2** (corresponding to **Figure 1** above) shows the residual or school effects for the reference categories and the mean value (0) of the LRT.

We again see the marked uncertainty associated with comparisons among schools.

Because the school effect is also a function of the LRT score and the VR band, we can estimate the effect at different values. Thus, we can compare two extreme groups: for example, the low-achievers at intake with an LRT score of −2 (approximately the lower 2.5 percentile) and in the VR2 or VR3 category and the high-achievers at intake with an LRT score of 2 (approximately the upper 97.5 percentile) and in the VR1 category. **Figure 3** shows a scatterplot of the values from these two groups estimated from the model as $(u_{0j} - 2u_{1j})$ and $(u_{0j} + 2u_{1j} + u_{2j})$, respectively. Despite a moderate correlation between these values, the schools appear to vary in terms of how different types of students perform.

An additional analysis was undertaken by Goldstein et al. (1993) in which separate examination scores for English and mathematics were analyzed jointly in a bivariate two-level model. This allows us to estimate a school effect for both mathematics and English, and **Figure 4** shows the relationship between these estimated residuals. The estimated correlation between the effects of

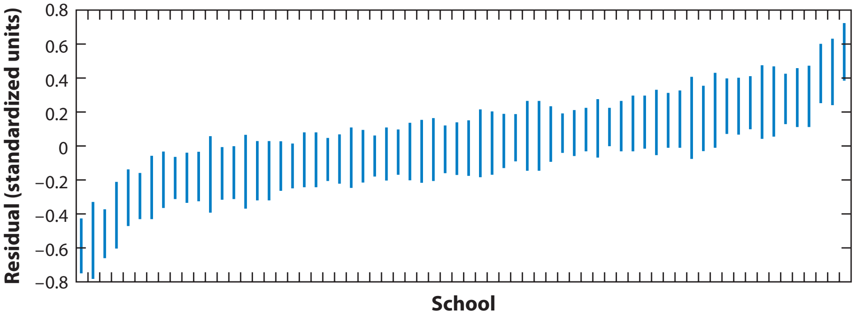

Figure 2

Residual estimates with 95% confidence intervals for examination score data.

the mathematics and English scores is only approximately 0.1 (see **Figure 4**). Together with the differential effectiveness as a function of intake achievement and the large amounts of uncertainty, any use of such data in the form of overall (unidimensional) league tables is highly problematic. Nevertheless, for screening purposes to identify schools that may be performing unexpectedly poorly or well, the use of such plots may be helpful (for a more detailed discussion, see Foley & Goldstein 2012, Yang et al. 1999).

The second example further examines school comparisons within the specific context of school choice: It is a national data set (the National Pupil Database) that contains longitudinal performance data on all students within maintained (state-funded) schools in England (for details, see Leckie & Goldstein 2009, 2011). The students were allocated a unique identification when they entered the system, and events such as school changes as well as test and exam scores, limited demographic information, and data on the schools the students had attended were recorded. This database is used both for research purposes and to produce annual league tables of schools based on test and examination scores, both unadjusted and adjusted. Within an accountability context, these tables may be used to assist parents in choosing schools, especially secondary schools, for

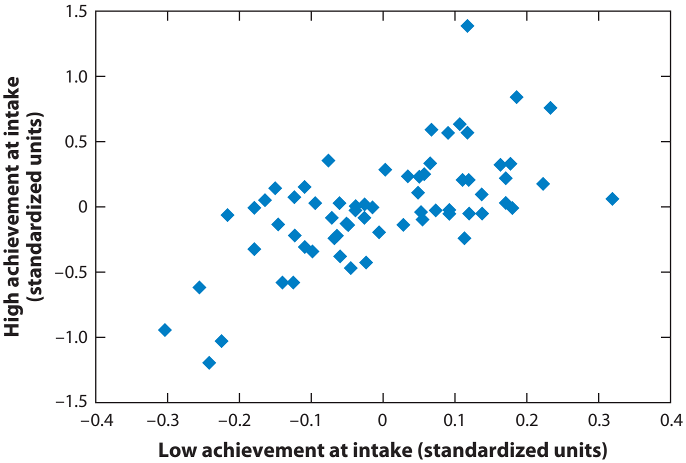

Figure 3

School residual estimates for low versus high achievement at intake.

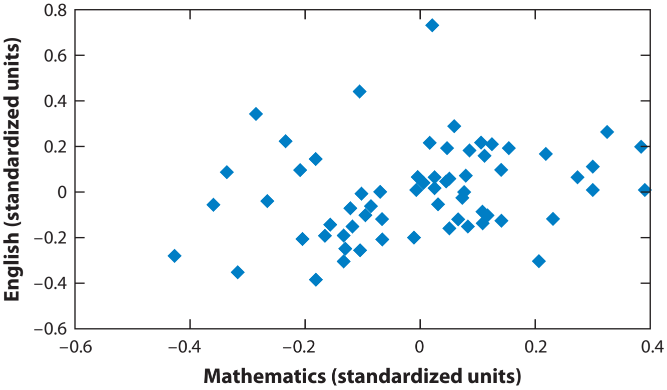

Figure 4

Estimated residuals for English and mathematics examination scores.

their children because the rankings process favors the schools that are "good" in terms of how they promote student achievements. In addition, parents who decide to base their choice of secondary school, at least in part, on such a school ranking will generally have available, at best, results that apply to the previous year's cohort. Their interest, however, is in the school's future performance in five or six years' time when their child will take the equivalent examination. The problem thus becomes how to predict a future set of school effects using a current set of school effects. Leckie & Goldstein (2009) utilized General Certificate of Secondary Education (GCSE) (school-leaving) examination data for two cohorts over a six-year period (2005 and 2010) for which prior achievement data on both cohorts are available.

The relevant model for the two cohorts of students can be written as

$$\begin{gathered} y_{ij}^{(1)} = \beta_0^{(1)} + \beta_1^{(1)} x_{ij}^{(1)} + u_j^{(1)} + e_{ij}^{(1)}, \\ y_{ij}^{(2)} = \beta_0^{(2)} + \beta_1^{(2)} x_{ij}^{(2)} + u_j^{(2)} + e_{ij}^{(2)}, \end{gathered} \qquad 7.$$

$$\begin{bmatrix} u_j^{(1)} \\ u_j^{(2)} \end{bmatrix} \sim N(0, \Omega_u), \qquad \Omega_u = \begin{bmatrix} \sigma_{u1}^2 & \\ \sigma_{u12} & \sigma_{u2}^2 \end{bmatrix},$$

$$\begin{bmatrix} e_{ij}^{(1)} \\ e_{ij}^{(2)} \end{bmatrix} \sim N(0, \Omega_e), \qquad \Omega_e = \begin{bmatrix} \sigma_{e1}^2 & \\ 0 & \sigma_{e2}^2 \end{bmatrix},$$

where superscripts 1 and 2 denote cohort 1 and cohort 2. Hence, $y_{ij}^{(1)}$ is the GCSE score for the ith pupil in the jth school in cohort 1 (2005), and $y_{ij}^{(2)}$ is the GCSE score for the ith pupil in the jth school in cohort 2 (2010). In general, the level-2 school residuals will be correlated. The level-1 residuals for the two responses are modeled as independent because a pupil can belong to only one cohort. Hence, this is a bivariate model where the bivariate structure is at level 2, rather than in the traditional multivariate multilevel model where it is at both levels.

From this model, having obtained the parameter estimates, we can obtain estimates of the school effects predicted for cohort 2 as functions of the terms in Ω_u, Ω_e, and $y_{ij}^{(1)}$. For school j,

$$\frac{\rho_{u12} n_j^{(1)} \sigma_u^2}{(n_j^{(1)} \sigma_u^2 + \sigma_{e1}^2)} \tilde{y}_j^{(1)},$$

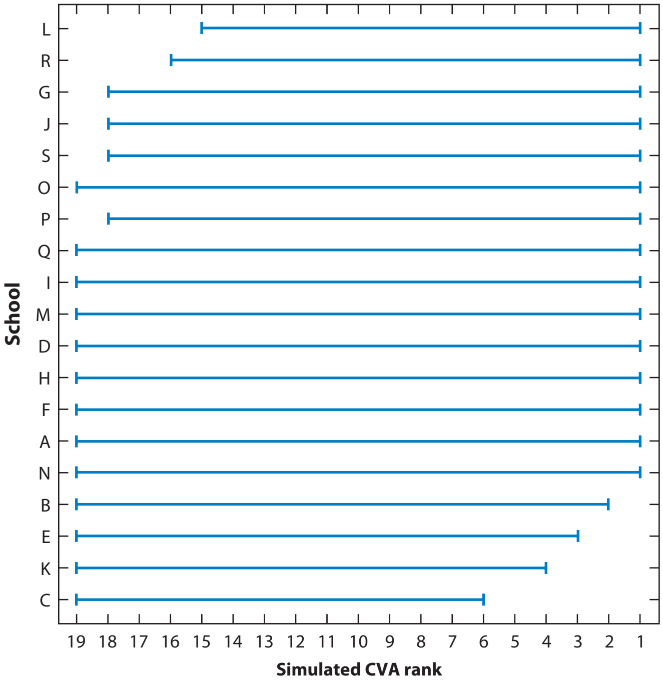

Figure 5

Predicted effects for 19 Bristol secondary schools with 95% confidence intervals. Adapted from Leckie & Goldstein (2011), with permission. Abbreviation: CVA, contextual value added.

where $\tilde{y}_j^{(1)}$ is the mean of the raw residuals for the *j*th school in cohort 1, with a corresponding term for the estimated standard error. From this we can construct a caterpillar plot that ranks the predicted school effects together with interval estimates (illustrated in **Figure 5**). Based on the results for 19 secondary schools in Bristol with 95% confidence intervals, the effects shown in **Figure 5** are striking with all intervals overlapping, thus providing no reliable separation.

Leckie & Goldstein (2011) presented school comparisons in terms of probabilities: For any given set of schools forming a choice set, any particular school will have the largest or smallest predicted effect. **Figure 6** illustrates this result for three chosen schools, none of which has a better than even chance of being ranked first. Such displays can convey the extensive uncertainty in ways that are clearly intelligible to nontechnical audiences.

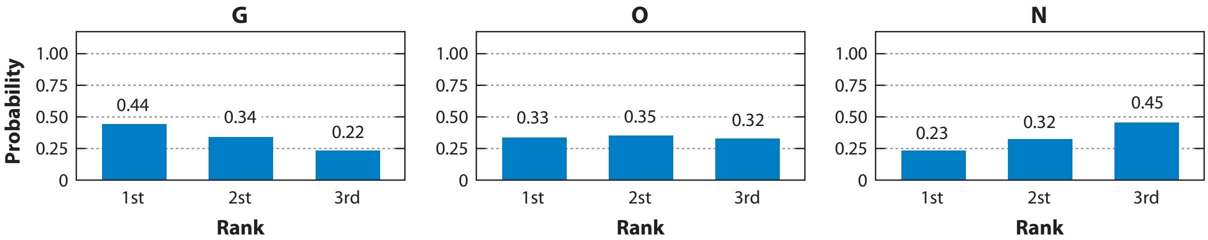

Figure 6

Probability that schools G, N, and O may rank first, second, or third. Adapted from Leckie & Goldstein (2011), with permission.

5. CONCLUSIONS

As pointed out in Section 1, I have concentrated on the statistical issues, providing illustrative examples, and I have not discussed side effects including "perverse incentives" for institutions to behave in ways that may not serve the best interests of those whom they are meant to serve, be they students, patients, or the general public. All of these issues, however, are both important and researchable, and the producers of league tables need to do more to encourage such research (for a fuller discussion, see Foley & Goldstein 2012). The evidence, where suitable data are available, indicates that rankings of institutions have large measures of uncertainty attached to them, even when appropriate adjustments for selection effects have been made. Perhaps the most effective uses of institutional rankings are as screening instruments that can suggest where problems may be occurring, rather than diagnoses of what the problems are.

I am not suggesting that league tables should never be published. Quantitative data that bear on performance are a useful tool for addressing the clear need for accountability from public (and other) institutions. When such data are reported publicly, however, their quality and reliability need to be displayed so that users of the data are not misled about what can be inferred. To withhold information about the uncertainty of rankings is to deprive users of information to which they are entitled.

Although useful work may need to be done to develop the models described here further, a more pressing need is to find ways to enhance data quality and, especially, to prevent unrealistic inferences from being drawn from any oversimple presentation of the results. For example, it is perfectly possible to develop software for sites that host institutional databases to provide information similar to that shown in **Figure 6**. This could be done in real time and could display bespoke comparisons among institutions. By displaying the real uncertainty surrounding institutional comparisons, such data would help users make properly informed judgments.

DISCLOSURE STATEMENT

The author is not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

I wish to acknowledge the helpful comments of anonymous referees and editors.

LITERATURE CITED

Amrein-Beardsley A, Collins C, Polasky SA, Sloat EF. 2013. Value-added model (VAM) research for educational policy: framing the issue. *Educ. Policy Anal. Arch.* 21:4

Baty P. 2010. Measured, and found wanting more. *Times High. Educ.*, July 8

Bird SM, Cox D, Farewell VT, Goldstein H, Holt T, Smith P. 2005. Performance indicators: good, bad, and ugly. *J. R. Stat. Soc. A* 168:1–27

Dill D, Soo M. 2005. Academic quality, league tables and public policy: a cross-national analysis of university ranking systems. *High. Educ.* 49(4):495–537

Ferrão ME, Goldstein H. 2012. Adjusting for differential misclassification in multilevel models: the relationship between child exposure to smoke and cognitive development. *Qual. Quant.* In press. doi:10.1007/s11135-012-9765-5

Foley B, Goldstein H. 2012. *Measuring Success: League Tables in the Public Sector*. London: British Acad.

Goldstein H, Burgess S, McConell B. 2007. Modelling the effect of pupil mobility on school differences in educational achievement. *J. R. Stat. Soc. A* 170(4):941–54

Goldstein H, Carpenter J, Kenward M, Levin K. 2009. Multilevel models with multivariate mixed response types. *Stat. Model.* 9(3):173–97

Goldstein H, Healy MJR. 1995. The graphical presentation of a collection of means. *J. R. Stat. Soc. A* 581(1):175–77

Goldstein H, Rasbash J, Yang M, Woodhouse G, Pan H, et al. 1993. Multilevel analysis of school examination results. *Oxf. Rev. Educ.* 19(4):425–33

Goldstein H, Sammons P. 1997. The influence of secondary and junior schools on sixteen year examination performance: a cross-classified multilevel analysis. *Sch. Eff. Sch. Improv.* 8(2):219–30

Goldstein H, Spiegelhalter DJ. 1996. League tables and their limitations: statistical issues in comparisons of institutional performance. *J. R. Stat. Soc. A* 159:385–443

Goldstein H. 2011. *Multilevel Statistical Models*. Chichester, UK: Wiley. 4th ed.

Hazelkorn E. 2011. *Rankings and the Reshaping of Higher Education: The Battle for World-Class Excellence*. New York: Palgrave MacMillan

Hood C. 2007. Public service management by numbers: Why does it vary? Where has it come from? What are the gaps and puzzles? *Public Money Manag.* 27(2):95–102

Hubbard AE, Ahern J, Fleischer NL, Van der Laan M, Lippman SA, et al. 2010. To GEE or not to GEE. *Epidemiology* 21(4):467–74

Leckie G, Goldstein H. 2009. The limitations of using school league tables to inform school choice. *J. R. Stat. Soc. A* 172:835–51

Leckie G, Goldstein H. 2011. Understanding uncertainty in school league tables. *Fisc. Stud.* 32:207–24

Lockwood JR, McCaffrey DF, Mariano LT, Setodji C. 2007. Bayesian methods for scalable value-added assessment. *J. Educ. Behav. Stat.* 32(2):125–50

Nuttall DL, Goldstein H, Prosser R, Rasbash J. 1989. Differential school effectiveness. *Int. J. Educ. Res.* 13:769–76

Rubin D. 1987. *Multiple Imputation for Non-Response in Surveys*. Chichester, UK: Wiley

Spiegelhalter DJ, Best NG, Carlin BP, Van der Linde A. 2002. Bayesian measures of model complexity and fit. *J. R. Stat. Soc. B* 64:583–640

Spiegelhalter DJ, Goldstein H. 2009. Comment: citation statistics. *Stat. Sci.* 24:21–24

Yang M, Goldstein H, Rath T, Hill N. 1999. The use of assessment data for school improvement purposes. *Oxf. Rev. Educ.* 25(4):469–83

Annual Review of Statistics and Its Application

Volume 1, 2014

Contents