Statistical Analysis of Performance Indicators in UK Higher Education

David Draper

Department of
Applied Mathematics and Statistics
University of California, Santa Cruz

`draper@ams.ucsc.edu`
`http://www.ams.ucsc.edu/∼draper`

(joint work with **Mark Gittoes**, *Higher Education Funding Council For England*)

VIII International Meeting on Quantitative Methods for Applied Sciences

11–12 September 2006 (Siena IT)

Keywords: Causal inference, calibration, counterfactuals, fixed-effects models, league tables, multilevel modeling, observational study, quality assessment, random-effects models, simulation methods, standardization.

Citation: Draper D, Gittoes M (2004). Statistical analysis of performance indicators in UK higher education (with discussion). *Journal of the Royal Statistical Society, Series A*, **167**, 449–474 (discussion 497–499).

The Main Case Study

“Following recommendations of the *National Committee of Inquiry into Higher Education*, the [UK] Government asked the funding councils to develop suitable **indicators** and **benchmarks** of **performance** in the higher education sector.”

[First report of the *Performance Indicators Steering Group*,
published by HEFCE in February 1999]

“Recognizing the diversity of higher education, the purpose of **performance indicators** is to

- provide better and more reliable information on the nature and **performance** of the UK higher education sector as a whole;
- influence **policy developments**; and
- contribute to the **public accountability** of higher education.”

[*Performance Indicators in Higher Education, 1996–97, 1997–98*,
published by HEFCE in December 1999]

Performance Indicators

Performance indicators (PIs) are outcomes that are thought to be measures of the **quality** with which British universities are carrying out their public mandate.

The *Higher Education Funding Council for England* (HEFCE; `www.hefce.ac.uk`) and the UK *Higher Education Statistics Agency* (HESA; `www.hesa.ac.uk`) have been publishing results for **three** types of such indicators **annually since 1997**:

- Participation of **under-represented groups** (access) (percentage of students at the university (i) from *state-funded* secondary schools, (ii) whose parents' occupation was *skilled manual, semi-skilled*, or *unskilled*, and (iii) whose home area, as denoted by postcode, is known to have a *low proportion* of 18– and 19–year-olds in higher education);

- **Learning and teaching outcomes**, e.g., **efficiency** (*how long it takes students to finish* compared with how long it should take them), and student **non-continuation** (drop-out) (percentage of students who are *absent from the higher education (HE) system* (apart from graduation) in Oct 1998 having started at a new university in Oct 1997); and

- **Research** output (quantity) relative to resources consumed (*numbers of PhD students* and *amounts of research grants and contracts* obtained, per academic staff costs and funding council research allocation).

We will concentrate here on student **drop-out** data from the first HEFCE report in $\boxed{\textbf{1996–97}}$ (actually we work with **progression** = 1 − drop-out).

Statistically Speaking

In the language of **causal inference** and **experimental design**,

- there is an **outcome** variable Y at the student level (a 0/1 indicator of progressing into the second year or not),

- there is a *supposedly causal factor* (SCF) at the university level S (the underlying quality of the university), which is **unobserved**, and

- the process by which students choose universities is a large **observational study** in which the choice mechanism may be confounded with the SCF.

In any observational study it is crucial to identify as many **potential confounding factors** (PCFs) X as possible—these are variables which may possibly be **correlated both with the outcome and with the SCF**.

Failure to adjust for such PCFs will lead to a **biased** estimate of the causal effect of the SCF on the outcome.

Multilevel Data

The data have a two-level character, with **284,399** students having entered **165** universities in 1996–97.

The range of entry class size was from **55** [RCN Institute] to **6,831** [UWE]; Bath had 1,344; Bristol 2,499; Oxford 3,443; median **1,539**.

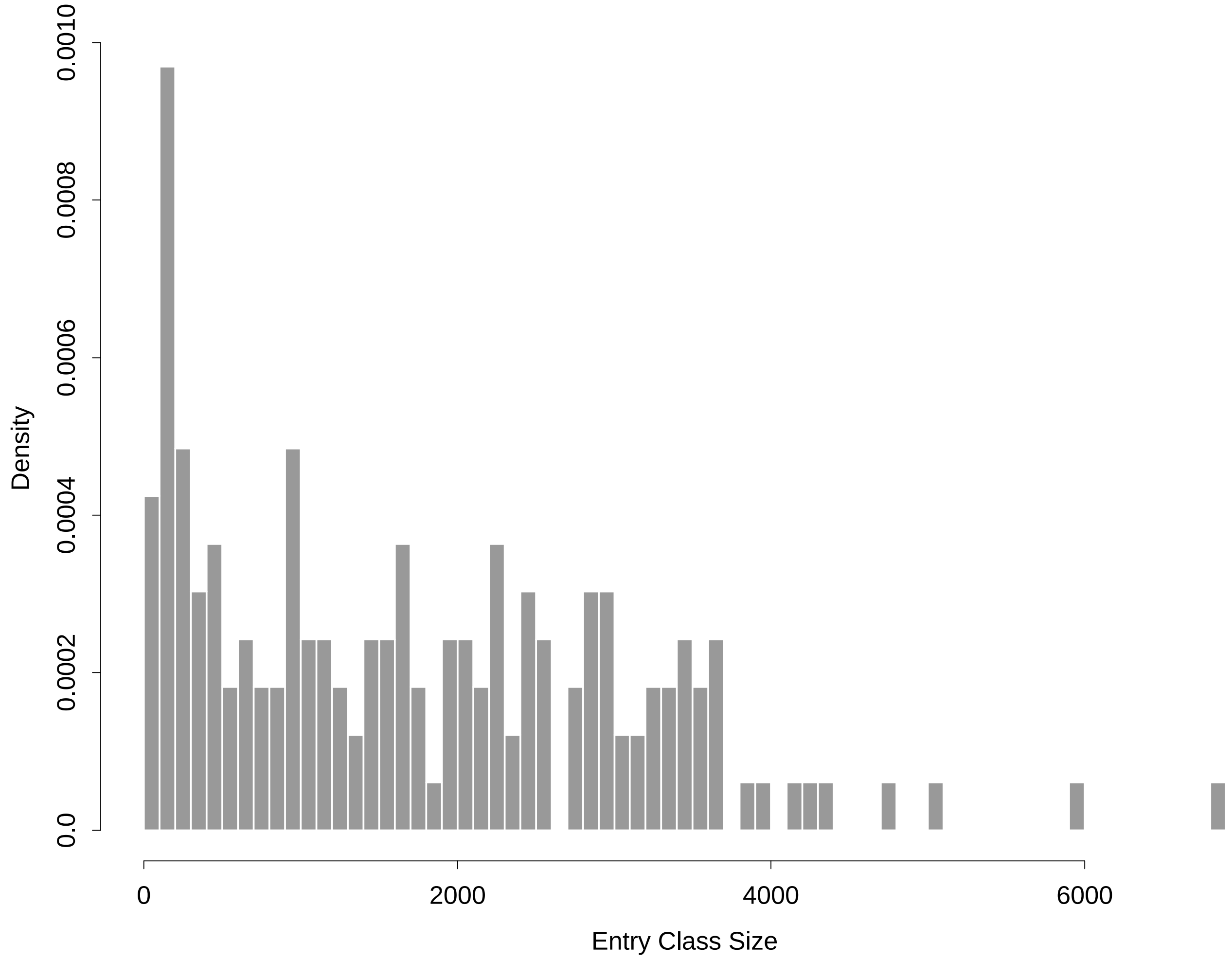
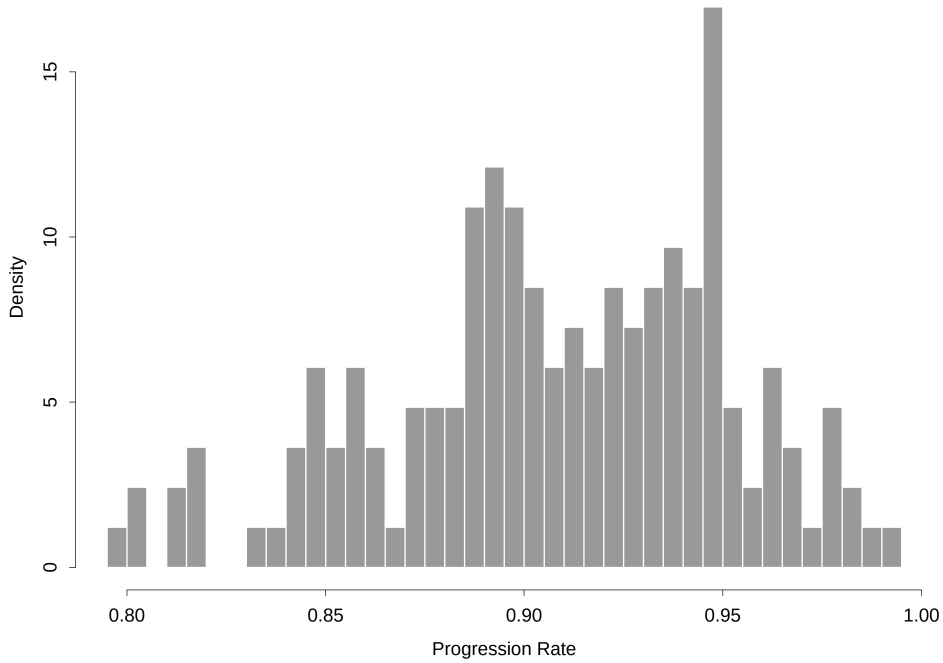
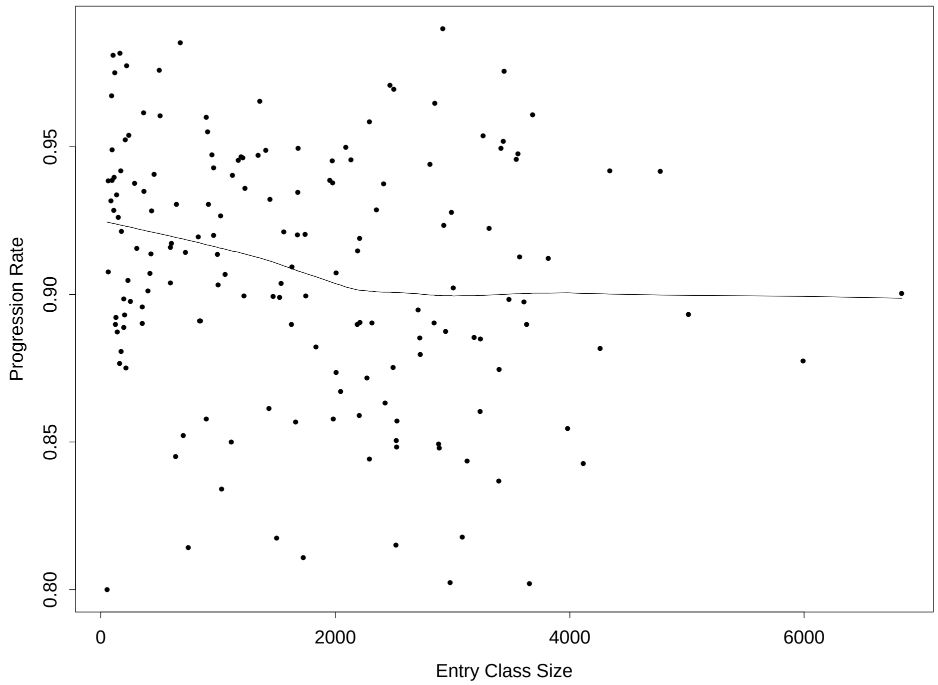


Figure 1. *Histogram of numbers of entering students at the 165 UK universities in 1996–97.*

Progression Rates

The dichotomous progression outcome had a mean of **0.901** in 1996–97 (Bath **0.93**; Bristol **0.95**; Oxford **0.97**).





PCFs for Student Dropout

By linking with the *Universities and Colleges Admissions Service* (**UCAS**) and *Higher Education Statistics Agency* (**HESA**) data bases, HEFCE have available the following eight **PCFs** at the student level for possible adjustment:

- **age**, as a dichotomy for young ($\leq$ 21 on entry to this university) or not;

```
                                        Progression
   Age |       Freq.      Percent          Rate
-------+----------------------------------------
Mature |       81905        28.8           .847
 Young |      202494        71.2           .924
-------+----------------------------------------
 Total |      284399       100.0           .901
```

- **gender** (1 = male, 0 = female);

```
                                        Progression
Gender |       Freq.      Percent          Rate
-------+----------------------------------------
  Male |      138740        48.8           .885
Female |      145659        51.2           .917
-------+----------------------------------------
 Total |      284399       100.0           .901
```

PCFs (continued)

- **entry qualification** (categorical at 21 levels: type of qualification if not A-level, otherwise 2- or 4-point A-level categories) (note the treatment of **missing data** here and below);

```
       Entry                          Progression
        Qual. |   Freq.     Percent      Rate
    ----------+-----------------------------------
         None |    5234        1.84      .787
       Others |    8211        2.89      .793
      Unknown |    8031        2.82      .832
     BTEC/ONC |   15308        5.38      .849
       GNVQ3+ |    8015        2.82      .856
           HE |   26493        9.32      .852
      ACC/FND |   21906        7.70      .864
A pts Unknown |   13611        4.79      .871
     A pts  4 |    8353        2.94      .869
     A pts  8 |   17678        6.22      .889
     A pts 10 |   12138        4.27      .903
     A pts 12 |   13434        4.72      .905
     A pts 14 |   14539        5.11      .915
     A pts 16 |   15112        5.31      .926
     A pts 18 |   15375        5.41      .942
     A pts 20 |   15371        5.40      .945
     A pts 22 |   13763        4.84      .952
     A pts 24 |   13275        4.67      .961
     A pts 26 |   12289        4.32      .967
     A pts 28 |   10891        3.83      .972
     A pts 30 |   15372        5.41      .984
    ----------+-----------------------------------
        Total |  284399       100.0      .901
```

PCFs (continued)

- **subject** of study (categorical at 13 levels);

```
                                   Progression
     Subject |    Freq.     Percent    Rate
-------------+--------------------------------
 Engineering |    22638        7.96    .864
Maths & Comp |    20569        7.23    .877
Architecture |     7151        2.51    .877
    Combined |    34748       12.2     .889
    Business |    36610       12.9     .891
 Agriculture |     2511        0.880   .903
 SocSt + Law |    33567       11.8     .903
 Allied to M |    15240        5.36    .904
Art + Design |    23317        8.20    .906
   Education |    14543        5.11    .913
 Biol + Phys |    36498       12.8     .916
  Lang + Hum |    30199       10.6     .930
    Medicine |     6808        2.39    .980
-------------+--------------------------------
       Total |   284399      100.0     .901
```

PCFs (continued)

- Was the student's secondary education at a **state school** or not?

```
  State                               Progression
 School? |     Freq.   Percent      Rate
---------+-------------------------------------
 Unknown |    103945     36.6       .856
     Yes |    146295     51.4       .925
      No |     34159     12.0       .938
---------+-------------------------------------
   Total |    284399    100.0       .901
```

- **Parental occupation**: Is the occupation of the principal wage-earner in the student's family **skilled manual**, **semi-skilled**, or **unskilled**?

```
 Parental                               Progression
Low Skills? |     Freq.   Percent      Rate
------------+-------------------------------------
    Unknown |     89669     31.5       .854
        Yes |     50480     17.8       .905
         No |    144250     50.7       .929
------------+-------------------------------------
      Total |    284399    100.0       .901
```

PCFs (continued)

- **Low HE participation**: Does the **address** from which the student applied to university have a postal code with a **low rate of university participation**? (Highly correlated with **income**.)

```
                                        Progression
Low Income? |     Freq.     Percent       Rate
------------+-------------------------------------
    Unknown |     11566       4.07        .846
        Yes |     37955      13.4         .875
         No |    234878      82.6         .908
------------+-------------------------------------
      Total |    284399     100.0         .901
```

- **Year of (program of) HE study**: (most people for whom this variable is 1 are just starting university).

```
                                             Progression
   Study Year |        Freq.     Percent       Rate
--------------+-----------------------------------------
2+ or Unknown |        32612        11.5        .878
            1 |       251787        88.5        .904
--------------+-----------------------------------------
        Total |       284399       100.0        .901
```

(HEFCE has also worked with several PCFs based on **geography**.)

What HEFCE Does

HEFCE's method of adjusting for the PCFs is a version of **league-table** (Goldstein and Spiegelhalter 1996) or **input-output** (IO; Draper 1995) quality assessment (in health policy this is also called **provider profiling**).

NB **HEFCE themselves actively discourage forming league tables across the entire HE sector.**

In this approach the quality of the institution (university) is inferred **indirectly**, by measuring its **outputs** and adjusting for its **inputs**, with no attempt to directly measure the quality of the **processes** going on inside the institution.

Imagine cross-tabulating all of the PCFs against each other, and let $\boxed{M}$ be the number of nonempty cells (**PCF categories**) in this table.

HEFCE's adjustment method is based on a **further cross-tabulation** of the $\boxed{N}$ universities by these M PCF categories.

The Basic HEFCE Grid

	PCF Categories				Weighted
University	1	2	$\dots$	M	Row Mean
1	$\widehat{p}_{11}$	$\widehat{p}_{12}$	$\dots$	$\widehat{p}_{1M}$	$\widehat{p}_{1\cdot}$
2	$\widehat{p}_{21}$	$\widehat{p}_{22}$	$\dots$	$\widehat{p}_{2M}$	$\widehat{p}_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
N	$\widehat{p}_{N1}$	$\widehat{p}_{N2}$	$\dots$	$\widehat{p}_{NM}$	$\widehat{p}_{N\cdot}$
Weighted Column Mean	$\widehat{p}_{\cdot 1}$	$\widehat{p}_{\cdot 2}$	$\dots$	$\widehat{p}_{\cdot M}$	$\widehat{p}_{\cdot\cdot}$

	PCF Categories				Row
University	1	2	$\dots$	M	Sum
1	n_{11}	n_{12}	$\dots$	n_{1M}	n_{1+}
2	n_{21}	n_{22}	$\dots$	n_{2M}	n_{2+}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
N	n_{N1}	n_{N2}	$\dots$	n_{NM}	n_{N+}
Column Sum	n_{+1}	n_{+2}	$\dots$	n_{+M}	n_{++}

Here $\widehat{p}_{ij}$ and n_{ij} are the **observed progression rate** and **number of students**, respectively, in PCF category j at university i, $n_{i+} = \sum_{j=1}^{M} n_{ij}$ is the **entry class size** at university i (and similarly for n_{+j}),

$$\widehat{p}_{.j} = \frac{\sum_{k=1}^{N} n_{kj}\widehat{p}_{kj}}{\sum_{k=1}^{N} n_{kj}} = n_{+j}^{-1} \sum_{k=1}^{N} n_{kj}\widehat{p}_{kj} \qquad (1)$$

is the observed **national progression rate** for students in PCF category j, and $\widehat{p}_{i.}$ is the **observed progression rate** for university i.

The HEFCE Grid: An Example

For illustration consider a **small world** with only $N = 5$ universities and $M = 4$ PCF categories, defined by cross-tabulating **age** against **gender**:

$\boxed{\widehat{p}_{ij}}$	PCF Categories				
	Young		Mature		Weighted
University	Male	Female	Male	Female	Mean
RCN	—	—	.800	.800	.800
Newport	.838	.889	.819	.878	.858
UWE	.897	.923	.868	.905	.900
Bath	.941	.972	.884	.887	.947
Cambridge	.993	.992	.958	.969	.990
Weighted Mean	.933	.947	.870	.903	.924

$\boxed{n_{ij}}$	PCF Categories				
	Young		Mature		Row
University	Male	Female	Male	Female	Sum
RCN	0	0	5	50	55
Newport	198	271	227	205	901
UWE	2133	2099	1443	1156	6831
Bath	712	501	69	62	1344
Cambridge	1467	1176	144	130	2917
Column Sum	4510	4047	1888	1603	12048

Standardization

All adjustment methods like HEFCE's have a **counterfactual** character (e.g., Rubin 1974).

What actually *happened* at a given university with regard to progression rate, given that university's students, is **factual** data: this is the **observed** progression rate $\boxed{\hat{O}_i = \hat{p}_{i\cdot}}$ at university i.

All IO methods require the estimation of **counterfactual** data of the form {what *would have* happened at this university as far as progression rate is concerned *if* ...}: this defines the **expected** progression rate $\boxed{\hat{E}_i}$ at university i.

There are **two main counterfactuals** of interest in IO analysis, each corresponding to a different kind of adjustment.

Both adjustments are based on the method of **standardization** (e.g., Anderson et al. 1980), as follows.

Indirect vs. Direct Standardization

- **Direct** standardization (to the *national* cohort). We can ask the question "What would the observed overall progression rate have been at this university if its progression rates (in the PCF categories) had been what they were, but its **distribution** of students across the PCF categories had instead matched the national distribution?"

This is like imagining **sending the whole country to this university** instead of its own students, and involves computing $\widehat{E}_i$ by holding the $\widehat{p}_{ij}$ constant and changing the n_{ij} to n_{+j}.

- **Indirect** standardization (to the *university* cohort). Or we can ask "What would the observed overall progression rate have been at this university if its distribution of students across the PCF categories had been what it was, but its **progression rates** in the PCF categories were replaced by the national rates?"

This is like imagining **how well the whole country would do** with this university's students, and involves computing $\widehat{E}_i$ by holding the n_{ij} constant and changing the $\widehat{p}_{ij}$ to $\widehat{p}_{\cdot j}$.

This (**indirect** standardization) is what HEFCE actually does.

The HEFCE Procedure

In the above notation the observed progression rate $\widehat{O}_i$ at university i is a **weighted average** of the form

$$\widehat{O}_i = \widehat{p}_{i.} = \frac{\sum_{j=1}^{M} n_{ij} \widehat{p}_{ij}}{\sum_{j=1}^{M} n_{ij}} = n_{i+}^{-1} \sum_{j=1}^{M} n_{ij} \boxed{\widehat{p}_{ij}}, \qquad (2)$$

and HEFCE's expected rate based on indirect standardization, which they call the **benchmark**, is also a weighted average:

$$\widehat{E}_i = n_{i+}^{-1} \sum_{j=1}^{M} n_{ij} \boxed{\widehat{p}_{.j}}. \qquad (3)$$

HEFCE compute (but do not emphasize) the **difference** $\widehat{D}_i = \widehat{O}_i - \widehat{E}_i$ and refer it to its standard error $\widehat{SE}\left(\widehat{D}_i\right) = \sqrt{\widehat{V}\left(\widehat{D}_i\right)}$, via the ratio

$$\widehat{z}_i = \frac{\widehat{D}_i}{\widehat{SE}\left(\widehat{D}_i\right)}, \qquad (4)$$

as a basis for identifying unusually **"good"** and unusually **"bad"** universities.

In HEFCE's reports on PIs, all universities with $|\widehat{z}_i| > 3$ and $|\widehat{D}_i| > .03$ are flagged with an **asterisk** $(*)$ (this addresses both **statistical** and **practical significance**).

Remarks on the HEFCE Procedure

- We have **all the UK data** for any given year, so in a sense the standard errors are **predictive**, not inferential: we are thinking of the students at a given university in a given PCF category as **like a random sample of such students in the future** (assuming that **underlying educational quality does not change**).
  - Epidemiologists (e.g., Greenland 1998) concentrate on the **ratio** $\frac{\widehat{O}_i}{\widehat{E}_i}$ rather than the difference between observed and expected (results with $(\widehat{O}_i - \widehat{E}_i)$ and $\frac{\widehat{O}_i - \widehat{E}_i}{\widehat{E}_i} = \left(\frac{\widehat{O}_i}{\widehat{E}_i} - 1 \right)$ tend to be **similar**).

- In HEFCE's first report on PIs, the only PCFs in the adjustment process were **entry qualification** and **subject** of study ($M = 272$ PCF categories), but separate tables were given for **young** and **mature** students.
  - HEFCE's method as stated is **"not model-based"**, in the sense that there is no **regression** or **multilevel** model underlying their derivation (but see below).

- **Direct** standardization (without model-based smoothing) fails in this problem whenever M is even modestly large, because this leads to **empty cells** in the grid and local estimates of the $\widehat{p}_{ij}$ are not available.

Linear Regression Formulation

- An **alternative** regression-based way to think about and calculate the $\widehat{D}_i$:

(1) Fit a **generalized linear model** to the entire data set (ignoring the multilevel structure) in which Y is the binary outcome and the model is **fully saturated**, in the sense that all possible interactions are included (for example, if the set of PCFs includes the variables (X_1, X_2, X_3) then the model would include an intercept, the 3 main effects for the Xs, all 3 sets of 2–way interactions, and the single set of 3–way interactions).

(2) Obtain **predicted** values $\hat{Y}$ from this model.

(3) Then $\widehat{O}_i$ is just the **mean of the Y values** at university i and $\widehat{E}_i$ is the **mean of the $\hat{Y}$ values** at that university, from which $\widehat{D}_i$ may then be calculated as usual.

Because the model is fully saturated, the predicted values are just the **cell means** in the cross-tabulation of universities by PCF categories, so you can use **any link function** you want (e.g., *least-squares* regression is faster than *logistic* regression).

It should be possible to get a **standard error** for $\widehat{D}_i$ from this regression approach, but it's not easy.

The Right Variance

We have shown by another route that (a) HEFCE's standard errors in the early years of PI generation were in some cases **incorrect** and (b) the right theoretical variance for $\widehat{D}_i$ can be obtained by writing it as a **weighted sum** of all NM cells in the grid,

$$\widehat{D}_i = \sum_{k=1}^{N} \sum_{j=1}^{M} \lambda_{ikj} \, \widehat{p}_{kj}, \tag{5}$$

where

$$\lambda_{ikj} = \left\{ \begin{array}{cc} \frac{n_{ij}}{n_{i+}} \left(1 - \frac{n_{ij}}{n_{+j}} \right) & \text{for } i = k \\ -\frac{n_{ij} n_{kj}}{n_{i+} n_{+j}} & i \neq k \end{array} \right\} \tag{6}$$

(some algebra reveals that $\sum_{k=1}^{N} \sum_{j=1}^{M} \lambda_{ikj} = 0$ for all i). Then in **repeated sampling**

$$V\left(\widehat{D}_i\right) = \sum_{k=1}^{N} \sum_{j=1}^{M} \lambda_{ikj}^2 \, V\left(\widehat{p}_{kj}\right). \tag{7}$$

The problem becomes how to **estimate** $V\left(\widehat{p}_{kj}\right)$.

If the grid is not too **sparse** and the $\widehat{p}_{kj}$ are not too close to 0 or 1, simple **local** variance estimation should work well:

$$\widehat{V}_l\left(\widehat{p}_{kj}\right) = \frac{\widehat{p}_{kj}\left(1 - \widehat{p}_{kj}\right)}{n_{kj}}. \tag{8}$$

Results in the Small-World Example

This local variance estimation method produces the following results in our **small world** in 1996–97 with $N = 5$ universities and $M = 4$ PCF categories (**NB** by construction $\sum_{i=1}^{N} n_{i+} \widehat{D}_i = 0$):

University	n_{i+}	$\widehat{O}_i$	$\widehat{E}_i$	$\widehat{D}_i$	$\widehat{SE}\left(\widehat{D}_i\right)$	$\widehat{z}_i$
RCN	55	.800	.900	$-.100$	.053	-1.89
Newport	901	.858	.914	$-.056$	.011	-5.18
UWE	6831	.900	.919	$-.018$	.002	-9.25
Bath	1344	.947	.933	$+.014$	.006	$+2.39$
Cambridge	2917	.990	.934	$+.056$	.003	$+20.6$

The standard errors have the right **monotone** relationship with n_{i+}, but the $|\widehat{z}_i|$ values do seem rather large: **4** of the 5 universities would be identified as unusual with a rule of the form $|\widehat{z}_i| > 2$, and even with HEFCE's more stringent rule the proportion of extreme universities would be $\frac{2}{5} =$ **40%**.

Are these standard errors **too small**?

Answering this question requires a **null simulation** in which there are no "good" and "bad" universities, and we can see if the $\widehat{z}_i$–scores are **well calibrated**.

Doing this requires a **model** in which the quality of a university appears directly.

Multilevel Modeling: Fixed Effects

In the input-output approach to quality assessment the supposedly causal factor S at the university level is **unobserved**.

This suggests directly fitting models which include terms (either fixed or random effects) that **stand for** S.

For example, a linear version of such a **fixed-effects** model would look like

$$\begin{array}{c} y_{ij} = \beta_0 + \sum_{k=1}^{p} \beta_k \left(x_{ij k} - \bar{x}_k \right) + \boxed{\alpha_i} + e_{ij}, \\ e_{ij} \overset{\text{IID}}{\sim} N\left(0, \sigma_e^2\right), \quad \sum_{i=1}^{N} n_{i+} \alpha_i = 0, \end{array} \qquad (9)$$

where y_{ij} and x_{ijk} are the outcome and PCF "carrier" k for student j in university i and $\bar{x}_k$ is the grand mean of predictor k.

This model may be fit iteratively by maximum likelihood, e.g., using the **EM algorithm**.

Fixed-Effects Model Fitting

From (e.g.) least-squares starting values for the $\widehat{\beta}$, the two **iterative EM equations** are specified by

$$\begin{array}{c} \boxed{1}\ \widehat{\alpha}_i = y_{i\cdot} - \left[\widehat{\beta}_0 + \sum_{j=1}^{n_{i+}} \sum_{k=1}^{p} \widehat{\beta}_k \left(x_{ijk} - \bar{x}_k \right) \right], \\ \boxed{2}\ \text{Regress } (y_{ij} - \widehat{\alpha}_i) \text{ on the PCF} \\ \text{carriers } x_1, \ldots, x_p \text{ to get new } \widehat{\beta} \text{ values,} \end{array} \qquad (10)$$

where $y_{i\cdot} = \widehat{O}_i$ is the **observed progression rate** at university i.

Equation $\boxed{1}$ in (10) has the form

$$\widehat{\alpha}_i = (\text{observed rate}) - (\text{predicted rate}) \qquad (11)$$

and, since we have shown earlier that $\widehat{E}_i$ may be obtained from **regression predictions**, it should come as no surprise to find that

$$\boxed{\widehat{\alpha}_i \doteq \widehat{D}_i} \qquad (12)$$

to a good approximation (the only difference is that the $\widehat{\beta}$ values have y_{ij} as their outcome variable in the **regression** formulation and $(y_{ij} - \widehat{\alpha}_i)$ in the **multilevel modeling** formulation; in practice the two sets of $\widehat{\beta}$s are similar).

So the HEFCE method is effectively based on a **fixed-effects multilevel model** where the binary outcome is predicted with **linear** (not logistic) regression.

By this argument, and using a particular **variance estimation method** to be described below, we have demonstrated that

Indirect standardization to the institutional cohort is functionally equivalent to the following fixed-effects hierarchical model:

$$\begin{array}{c} y_{ij} = \beta_0 + \sum_{k=1}^{p} \beta_k \left(x_{ijk} - \bar{x}_k \right) + \boxed{\alpha_i} + e_{ij}, \\ e_{ij} \overset{\text{IID}}{\sim} N\!\left(0, \sigma_e^2\right), \quad \sum_{i=1}^{N} n_{i+}\, \alpha_i = 0. \end{array}$$

HEFCE Method vs. Multilevel Modeling

As an example of the correspondence between the HEFCE method and the fixed-effects multilevel model (9), we fit both methods to a **medium world** in 1996–97 consisting of 10 universities and 4 PCFs—gender, age, state school, and low participation—giving rise to $M = 2 \times 2 \times 3 \times 3 = 36$ PCF categories.

University (i)	n_i	$\widehat{D}_i$	$\widehat{\alpha}_i$
UWE	6831	0.0304	0.0325
Glasgow	3314	0.0262	0.0253
City	1113	–0.0306	–0.0314
Northampton	2205	–0.0276	–0.0293
Falmouth	289	0.0681	0.0712
Salford	3238	–0.0126	–0.0131
Central England	2889	–0.0284	–0.0290
Greenwich	2292	–0.0235	–0.0250
Abertay Dundee	1031	–0.0458	–0.0471
Royal Free	116	0.0441	0.0475

University (i)	n_i	$\widehat{SE}\left(\widehat{D}_i\right)$	$\widehat{SE}\left(\widehat{\alpha}_i\right)$
UWE	6831	0.00320	0.00344
Glasgow	3314	0.00507	0.00543
City	1113	0.00947	0.00960
Northampton	2205	0.00659	0.00663
Falmouth	289	0.01909	0.01904
Salford	3238	0.00532	0.00531
Central England	2889	0.00569	0.00567
Greenwich	2292	0.00645	0.00649
Abertay Dundee	1031	0.00996	0.00990
Royal Free	116	0.03014	0.03026

Here the $\widehat{\alpha}_i$ are estimated by maximum likelihood, with large-sample (Fisher-information-based) standard errors, and $\widehat{SE}\left(\widehat{D}_i\right)$ is based on a **global** variance estimate to be explained later.

EM Algorithm in the Big World

Recall from above that, from (e.g.) least-squares starting values for the $\widehat{\beta}$, the two **iterative EM equations** are specified by

$$\boxed{1}\ \widehat{\alpha}_i = y_{i\cdot} - \left[\widehat{\beta}_0 + \textstyle\sum_{j=1}^{n_{i+}} \sum_{k=1}^{p} \widehat{\beta}_k \left(x_{ijk} - \bar{x}_k \right) \right],$$

$\boxed{2}$ Regress $\left(y_{ij} - \widehat{\alpha}_i \right)$ on the PCF carriers $x_1, \ldots, x_p$ to get new $\widehat{\beta}$ values,

where $y_{i\cdot} = \widehat{O}_i$ is the **observed progression rate** at university i.

The EM updating step for the $\widehat{\alpha}_i$ is trivial, but the regression step is anything but trivial in the case of the **big world**: the full HEFCE grid with all $N = 165$ universities, 284,399 students, and $M = 17,799$ nonempty PCF categories formed by fully crossing all 8 available PCFs.

Brute-force regression $Y = X \beta + e$, leading to $\widehat{\beta} = \left(X^T X \right)^{-1} X^T Y$, fails when X has 284,399 rows and 17,799 columns.

Regression With 17,799 Predictors

However, in this case all of the columns of X are **indicator variables** for membership in the PCF categories, so $X^T X$ and $X^T Y$ have special forms:

$$X^T X = \left(\begin{array}{ccccc} d_0 & d_1 & d_2 & \dots & d_{M-1} \\ d_1 & d_1 & 0 & \dots & 0 \\ d_2 & 0 & d_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ d_{M-1} & 0 & 0 & \dots & d_{M-1} \end{array} \right), \qquad (13)$$

where $d_0 = n_{++}$ is the total number of students and (for $j > 0$) $d_j = n_{+j}$ is the national number of students in PCF category j.

Moreover the jth entry in the $M \times 1$ vector $X^T Y$ is s_j, where $s_0 = \sum_{i=1}^N \sum_{j=1}^M \sum_{k=1}^{n_{ij}} y_{ijk}$ and (for $j > 0$) $s_j = \sum_{i=1}^N \sum_{k=1}^{n_{ij}} y_{ijk}$ is the **national sum of the** y **values** in PCF category j.

We used `Maple` to **symbolically** invert $X^T X$, making regression with $M = 17,799$ predictors easy.

From this we have fit the fixed-effects multilevel model (9) to the entire HEFCE data set and confirmed that the $\hat{\alpha}_i$ values **closely match** HEFCE's $\hat{D}_i$, and so do the standard errors from both approaches (`GLIM4`, with the `eliminate` command, may be used to compute $\widehat{SE}\,(\hat{\alpha}_i)$, with some difficulty, in the big world).

Null Simulation

HEFCE's method is like fitting a **linear** multilevel model to the binary outcomes, but for simulation purposes it is more natural to fit the closely related **logistic multilevel model**

$$\begin{aligned} \left(y_{ij} \,|\, p_{ij} \right) & \stackrel{\text{indep}}{\sim} \text{Bernoulli} \left(p_{ij} \right) , \qquad (14) \\ \log \left(\frac{p_{ij}}{1 - p_{ij}} \right) & = \beta_0 + \textstyle\sum_{k=1}^{p} \beta_k \left(x_{ijk} - \bar{x}_k \right) + \alpha_i , \\ \textstyle\sum_{i=1}^{N} n_{i+} \, \alpha_i & = 0 , \end{aligned}$$

to avoid simulated y_{ij} values below 0 or above 1.

Thus we can create a null simulation world by using the fixed-effects logistic model (14), holding the PCFs constant at their values in the real data set, fixing the target overall progression rate at the actual $\hat{p}_{\cdot\cdot}$, and **setting all the α_i to 0**.

We did this for the **small world** of $N = 5$ universities and $M = 4$ PCF categories using local variance estimation, repeating the simulation **2,000** times.

$\boxed{\textbf{Results:}}$ z scores had mean **0.064** (0.005), SD **1.06** (0.010), Gaussian shape, $\hat{z}_i < -1.96$ **1.7%** (0.1%) of the time, $\hat{z}_i > 1.96$ **3.8%** (0.2%) of the time, $|\hat{z}_i| > 1.96$ **5.5%** (0.2%) of the time (Monte Carlo SEs in parenthesis).

Asymmetry and Sparseness

The **asymmetry** in this result is because the overall progression rate is close to 1.

When we repeated this simulation with $\widehat{p}_{\cdot\cdot} = \mathbf{0.5}$ instead of 0.9, results were as follows:

z scores had mean **0.00** (0.005), SD **1.03** (0.009), Gaussian shape, $\widehat{z}_i < -1.96$ **2.6%** (0.2%) of the time, $\widehat{z}_i > 1.96$ **2.5%** (0.2%) of the time, $|\widehat{z}_i| > 1.96$ **5.1%** (0.2%) of the time.

We conclude that local variance estimation is **reasonably well calibrated** when the university–PCF grid is not too **sparse** and the $\widehat{p}_{kj}$ are not too close to 0 or 1.

However, when the grid becomes too sparse—as the number M of PCF categories grows—the **local variance estimation method starts to break down**.

For example, fully crossing all 8 available PCFs produces $2 \cdot 2 \cdot 21 \cdot 13 \cdot 3 \cdot 3 \cdot 3 \cdot 2 =$ **58,968** potential PCF categories, of which $M =$ **17,799** are nonempty.

And while **284,399** sounds like a lot of students, $\frac{284{,}399}{165 \cdot 17{,}799}$ is only an average of **0.1** students per cell in the $N \times M$ grid.

Local and Global Variance Estimation

When we repeated our simulation 500 times, using $(*)$ **local** variance estimation on the grid with $N =$ **165** simulated universities and all $M =$ **17,799** PCF categories (holding the PCFs and the overall progression rate $\widehat{p}_{\cdot\cdot}$ **constant** at their values in the real data set), results were as follows:

z scores had mean **0.097** (0.003), SD **1.60** (0.008), $\widehat{z}_i < -1.96$ **8.1%** (0.1%) of the time, $\widehat{z}_i > 1.96$ **10.7%** (0.1%) of the time, $|\widehat{z}_i| > 1.96$ **18.8%** (0.1%) of the time (i.e., local $\widehat{SE}\left(\hat{D}_i\right)$s much too small).

One way to overcome the sparseness would be to use a **global** variance estimate, e.g.,

$$\hat{V}_g\Big(\widehat{p}_{kj}\Big) = \frac{\widehat{p}_{\cdot\cdot}\,(1 - \widehat{p}_{\cdot\cdot})}{n_{kj}}. \tag{15}$$

Under simulation conditions $(*)$ above with 200 replications, the z scores had mean **−0.002** (0.002), SD **0.93** (0.004), Gaussian shape, $\widehat{z}_i < -1.96$ **2.1%** (0.1%) of the time, $\widehat{z}_i > 1.96$ **1.7%** (0.1%) of the time, $|\widehat{z}_i| > 1.96$ **3.8%** (0.1%) of the time.

Thus global variance estimation is **a bit too conservative** but not far off.

Shrinkage Variance Estimation

A natural alternative to both local and global variance estimation would involve **shrinking** the local estimate $\widehat{p}_{kj}$ some distance toward the global estimate $\widehat{p}_{\cdot\cdot}$:

$$\widehat{V}_\gamma\Big(\widehat{p}_{kj}\Big) = \frac{\widehat{p}^*_{kj}\left(1 - \widehat{p}^*_{kj}\right)}{n_{kj}}, \tag{16}$$

where (for some $0 \leq \gamma \leq 1$)

$$\widehat{p}^*_{kj} = \gamma\,\widehat{p}_{\cdot\cdot} + (1-\gamma)\,\widehat{p}_{kj}\ . \tag{17}$$

Through experimentation we have found that $\gamma = 0.5$ **performs very well in calibrating the HEFCE method**.

When we repeated our simulation 200 times, using **shrinkage** variance estimation on the grid with $N =$ **165** simulated universities and all $M =$ **17,799** PCF categories (holding the PCFs and the overall progression rate $\widehat{p}_{\cdot\cdot}$ **constant** at their values in the real data set), results were as follows:

z scores had mean **0.015** (0.002), SD **1.00** (0.004), $\widehat{z}_i < -1.96$ **2.5%** (0.1%) of the time, $\widehat{z}_i > 1.96$ **2.7%** (0.1%) of the time, $|\widehat{z}_i| > 1.96$ **5.1%** (0.1%) of the time.

Results on the Real Data

- When **shrinkage** variance estimation is used on the **actual** 1996–97 progression data set (all $N = 165$ universities and 284,399 students) with adjustment for all 8 available PCFs ($M = 17,799$ PCF categories), the results are as follows:

z scores had mean **0.24**, SD **2.90**.

If the **Bonferroni multiple-comparisons** method is used to account for the fact that **165** significance tests are being made, the appropriate $|\hat{z}_i|$ cutoff is **3.61**; with this cutoff **16** (**9.7%**) and **12** (**7.3%**) of the universities are classified as "bad" and "good", respectively.

HEFCE's more **stringent** rule points the finger at only **12** (7.3%) "bad" and **8** (4.8%) "good" universities.

A **detailed table** of results follows.

Top and Bottom 15 Universities

Inst	n	$\hat{O}$	$\hat{E}$	$\hat{D}$	$\widehat{SE}$	$\hat{z}$	Sig
9	1031	0.83	0.89	-0.06	0.007	-8.11	** H
110	3658	0.80	0.84	-0.04	0.005	-8.00	** H
78	1728	0.81	0.87	-0.06	0.007	-7.97	** H
151	2981	0.80	0.84	-0.04	0.006	-7.31	** H
145	4115	0.84	0.87	-0.03	0.004	-6.79	**
154	3126	0.84	0.88	-0.03	0.005	-6.31	** H
7	2889	0.85	0.88	-0.03	0.005	-6.00	** H
64	1501	0.82	0.86	-0.04	0.007	-5.80	** H
101	639	0.85	0.91	-0.06	0.011	-5.63	** H
122	2519	0.82	0.85	-0.03	0.006	-5.49	** H
86	1836	0.88	0.91	-0.03	0.006	-5.12	** H
4	2205	0.86	0.89	-0.03	0.006	-4.82	**
25	748	0.81	0.86	-0.05	0.011	-4.30	** H
3	1113	0.85	0.88	-0.03	0.009	-3.99	** H
137	2192	0.91	0.94	-0.02	0.005	-3.92	**
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
20	3482	0.90	0.88	0.02	0.005	3.57	
30	3819	0.91	0.90	0.01	0.004	3.57	
63	1023	0.93	0.90	0.03	0.008	3.59	
35	1975	0.95	0.93	0.02	0.005	3.66	*
129	3559	0.95	0.93	0.01	0.004	3.70	*
123	1445	0.93	0.91	0.03	0.006	3.99	*
159	3186	0.89	0.87	0.02	0.005	4.10	*
164	1685	0.95	0.93	0.02	0.006	4.16	*
57	1751	0.90	0.87	0.03	0.006	4.53	*
17	2294	0.96	0.94	0.02	0.005	4.57	*
121	949	0.95	0.91	0.04	0.008	4.88	* H
144	1173	0.95	0.91	0.04	0.007	5.37	* H
71	913	0.96	0.91	0.04	0.008	5.65	* H
44	3262	0.95	0.93	0.03	0.004	7.10	*
46	3573	0.91	0.88	0.04	0.004	8.63	* H

Results (continued)

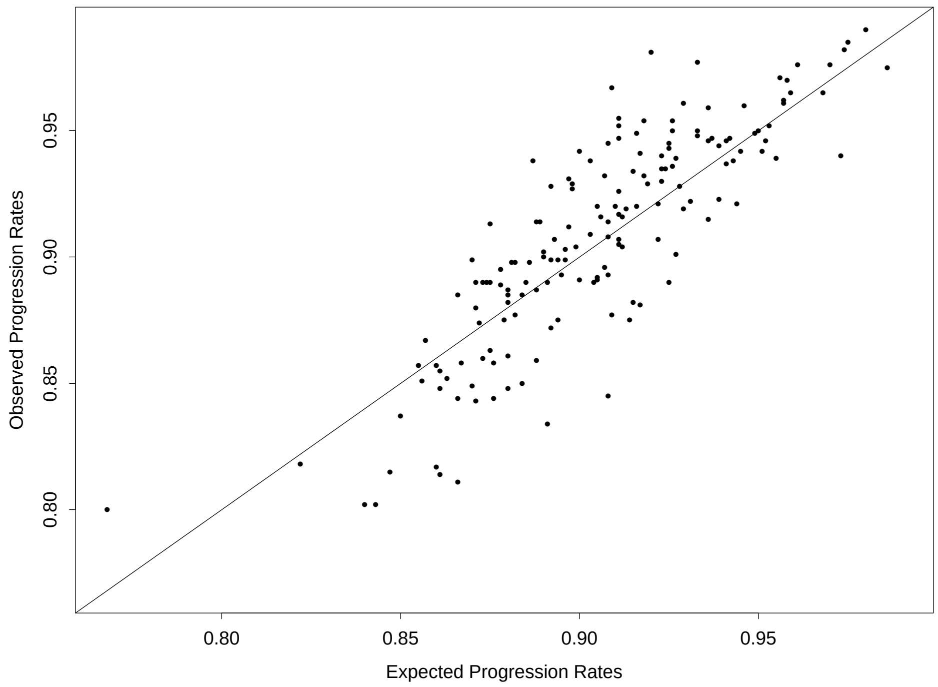
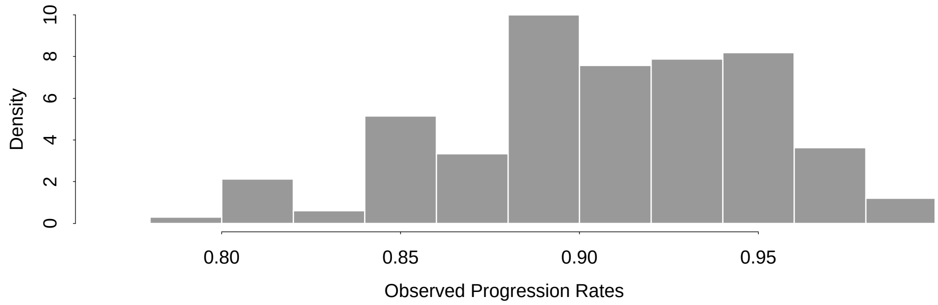


Figure 4. *O and E have correlation +0.85; 99 of the 165 universities (60%) have $O \geq E$.*



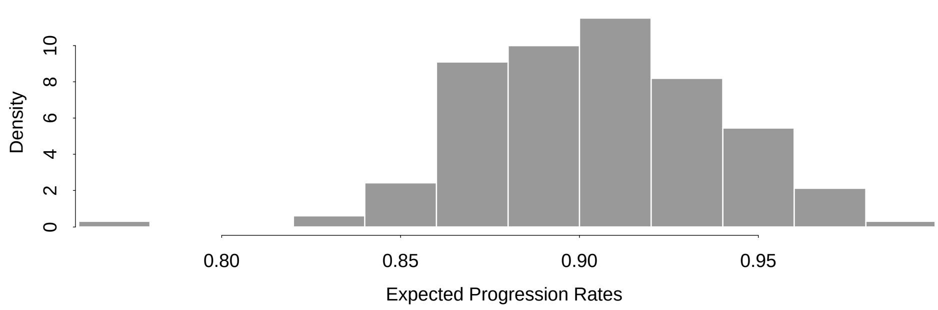


Figure 5. *63% of the variance in O is "explained" by E at the university level.*

Results (continued)

- It is interesting to examine how the results change **as PCFs are added** (the table below is based on **global** variance estimation, and row k averages over all $\left(\begin{array}{c} 8 \\ k \end{array} \right)$ possible subsets of PCFs):

Number of		Mean	z–scores		%	%	%
PCFs	Models	M	Mean	SD	"Bad"	"Good"	Unusual
0	1	1.0	0.860	6.20	0.212	0.280	0.509
1	8	6.1	0.762	5.29	0.185	0.258	0.442
2	28	31.1	0.678	4.64	0.165	0.227	0.392
3	56	132.4	0.602	4.17	0.150	0.200	0.351
4	70	483.9	0.531	3.81	0.139	0.176	0.316
5	56	1,498.9	0.461	3.53	0.127	0.154	0.282
6	28	3,933.3	0.389	3.30	0.116	0.135	0.252
7	8	8,906.4	0.314	3.10	0.111	0.114	0.225
8	1	17,799.0	0.236	2.90	0.097	0.073	0.170

- One **problem** with looking at z–scores is that they are based on **wildly different sample sizes**.

For example, in our small world with only 5 universities and 4 PCF categories, the RCN had $\hat{z}_i = -1.89$ and Bath had $\hat{z}_i = +2.39$, and there is no way to tell just by looking at the z–scores that we are **much less certain** about underlying quality at the RCN ($n_{i+} = 55$ students) than at Bath ($n_{i+} = 1,344$).

Results (continued)

A **simple graphical solution** plots $\widehat{D}_i \pm 1.96\,\widehat{SE}\left(\widehat{D}_i\right)$ for each university after sorting the $\widehat{D}_i$ from smallest to largest.

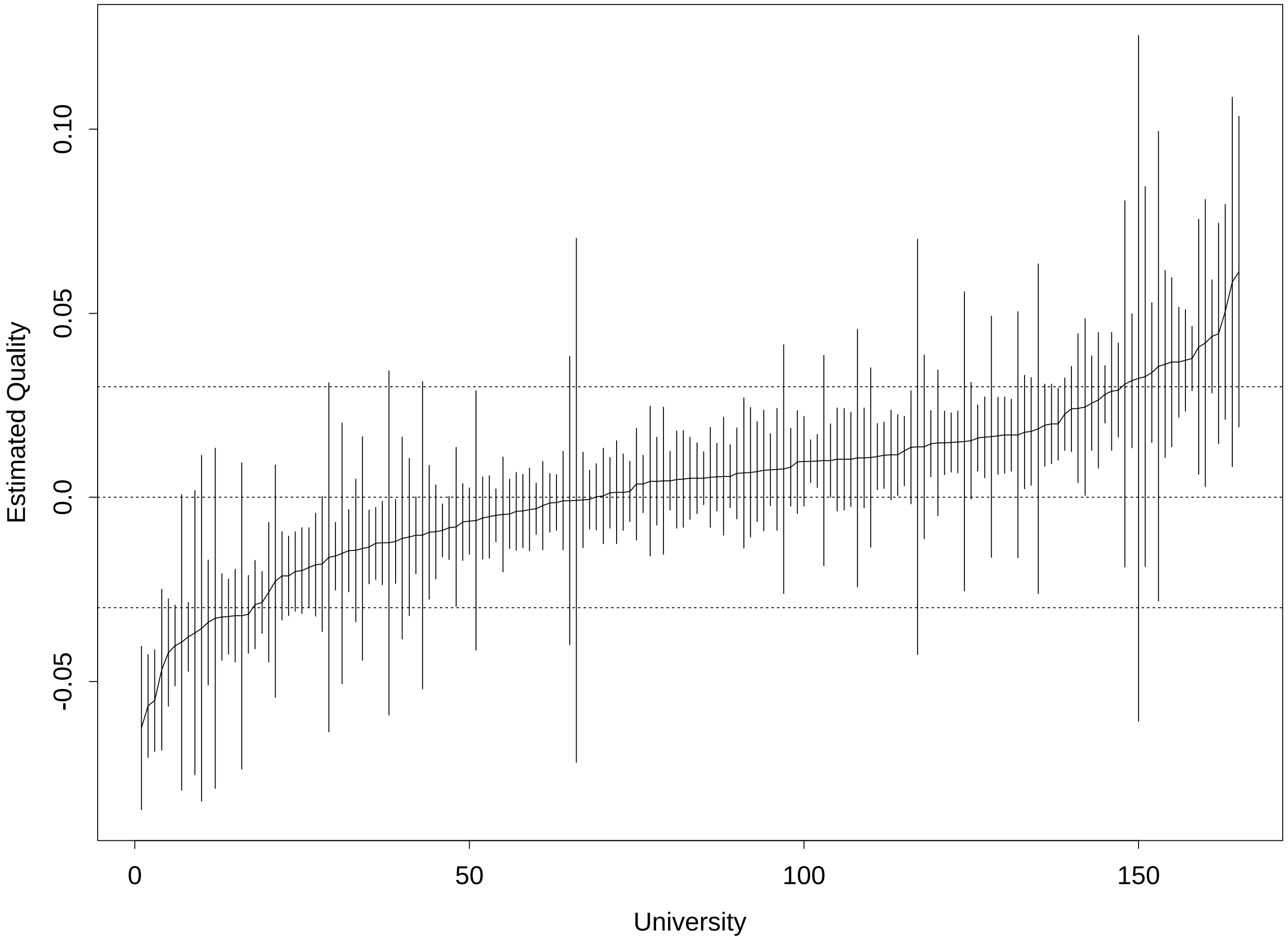


Figure 6. *Estimated quality $\widehat{D}_i \pm 1.96\,\widehat{SE}\left(\widehat{D}_i\right)$ with the universities renumbered from smallest to largest in $\widehat{D}_i$.*

30 universities have their **95% "quality interval"** entirely below 0 and **44** entirely above 0, but only **3** and **0** universities have their entire interval below -0.03 and above 0.03, respectively.

Non-Null Simulations

How **certain** are we that we are **right** when we identify a university as "bad" or "good"?

One answer to this question is based on **non-null simulations**, in which data sets are generated with a model like

$$\left(y_{ij} \,|\, p_{ij} \right) \stackrel{\text{indep}}{\sim} \text{Bernoulli} \left(p_{ij} \right), \qquad (18)$$

$$\log \left(\frac{p_{ij}}{1 - p_{ij}} \right) = \beta_0 + \sum_{k=1}^{p} \beta_k \left(x_{ijk} - \bar{x}_k \right) + \alpha_i,$$

$$\sum_{i=1}^{N} n_{i+} \, \alpha_i = 0,$$

for various choices of $\alpha_i \neq 0$.

For illustration we fit this model to the entire data set, via maximum likelihood, but using only the $M = 272$ PCF categories based on **entry qualifications** and **subject** employed in the HEFCE December 1999 publication of PIs, obtaining $\widehat{\alpha}_i$ values.

We then set $\alpha_i = \widehat{\alpha}_i$ and generated **5,000 random replications** of all 165 universities (with the same n_{i+} values as in the actual data), keeping track of the mean and SD of the $\widehat{z}_i$ scores and the percentage of time each university was classified as "bad", "OK", and "good" (with the HEFCE z–scores on the real data as **"truth"**).

Non-Null Simulations (continued)

Bonferroni Cutoff (3.61)

Univ.	True		%	%	%		Mean	SD
No.	z	Truth	Good	OK	Bad	n	of z	of z
110	-10.60	-1	0.000	0.000	100.0	3660	-10.2	1.01
78	-8.45	-1	0.000	0.000	100.0	1730	-8.40	1.03
151	-6.81	-1	0.000	0.267	99.7	2980	-6.72	1.03
7	-6.64	-1	0.000	0.133	99.9	2890	-6.55	0.975
154	-6.41	-1	0.000	0.133	99.9	3130	-6.37	0.98
122	-6.26	-1	0.000	1.07	98.9	2520	-6.23	1.10
145	-6.02	-1	0.000	2.00	98.0	4120	-5.95	1.06
64	-5.93	-1	0.000	1.47	98.5	1500	-5.92	1.04
101	-5.15	-1	0.000	6.67	93.3	639	-5.13	0.999
81	-4.51	-1	0.000	21.7	78.3	3400	-4.49	1.07
86	-4.50	-1	0.000	17.7	82.3	1840	-4.49	0.949
4	-4.48	-1	0.000	21.2	78.8	2200	-4.42	1.02
75	-4.37	-1	0.000	23.2	76.8	3080	-4.38	1.06
25	-4.37	-1	0.000	21.5	78.5	748	-4.40	1.04
141	-4.26	-1	0.000	24.3	75.7	2500	-4.28	0.975
6	-4.20	-1	0.000	28.7	71.3	3240	-4.16	1.02
3	-4.02	-1	0.000	32.9	67.1	1110	-4.05	1.02
21	-4.01	-1	0.000	34.4	65.6	1440	-3.96	1.03
9	-3.94	-1	0.000	39.2	60.8	1030	-3.89	1.02
8	-3.65	-1	0.000	48.1	51.9	2290	-3.68	1.04
147	-3.41	0	0.000	57.5	42.5	2880	-3.39	1.06
137	-3.32	0	0.000	60.7	39.3	2190	-3.35	0.909
62	-3.30	0	0.000	65.6	34.4	2270	-3.23	0.983
139	-3.23	0	0.000	67.1	32.9	2930	-3.18	0.902
84	-2.52	0	0.000	87.3	12.7	2430	-2.45	1.02
74	-2.40	0	0.000	87.9	12.1	3980	-2.35	1.08
42	-2.38	0	0.000	88.9	11.1	901	-2.32	1.04
133	-2.23	0	0.000	95.5	4.53	4340	-2.18	0.838
55	-2.13	0	0.000	95.3	4.67	176	-2.00	0.944
49	-2.04	0	0.000	94.1	5.87	2520	-1.98	1.07
70	-1.90	0	0.000	94.9	5.07	5990	-1.84	1.03

Non-Null Simulations (continued)

Bonferroni Cutoff (3.61)

Univ.	True		%	%	%		Mean	SD
No.	z	Truth	Good	OK	Bad	n	of z	of z
56	2.17	0	5.73	94.3	0.000	1120	2.18	0.915
16	2.21	0	6.80	93.2	0.000	1680	2.27	0.913
158	2.22	0	9.07	90.9	0.000	1680	2.28	0.960
105	2.29	0	10.0	90.0	0.000	172	2.35	0.930
87	2.30	0	8.00	92.0	0.000	1950	2.32	0.913
125	2.35	0	12.0	88.0	0.000	2210	2.41	1.080
95	2.42	0	8.8	91.2	0.000	92	2.46	0.837
117	2.50	0	15.3	84.7	0.000	433	2.56	1.030
36	2.60	0	16.5	83.5	0.000	210	2.66	0.967
37	2.63	0	17.1	82.9	0.000	456	2.66	0.979
120	2.68	0	17.5	82.5	0.000	239	2.72	0.918
104	2.83	0	16.4	83.6	0.000	2010	2.65	0.998
159	2.84	0	22.4	77.6	0.000	3190	2.83	1.040
66	3.01	0	28.5	71.5	0.000	723	3.05	1.070
93	3.02	0	15.9	84.1	0.000	106	2.98	0.693
11	3.15	0	28.4	71.6	0.000	222	3.14	0.751
5	3.18	0	33.5	66.5	0.000	289	3.21	1.040
83	3.22	0	34.3	65.7	0.000	2090	3.24	0.909
14	3.23	0	20.7	79.3	0.000	2920	3.22	0.494
59	3.32	0	41.6	58.4	0.000	964	3.40	0.981
65	3.34	0	38.5	61.5	0.000	3010	3.36	1.020
79	3.37	0	37.1	62.9	0.000	3440	3.38	0.666
99	3.40	0	40.5	59.5	0.000	2320	3.38	1.070
63	3.52	0	45.7	54.3	0.000	1020	3.52	1.010
57	3.53	0	50.9	49.1	0.000	1750	3.58	1.100
35	3.82	1	58.5	41.5	0.000	1980	3.79	0.912
100	3.93	1	66.1	33.9	0.000	2470	3.94	0.737
20	4.00	1	64.7	35.3	0.000	3480	4.02	1.040
119	4.05	1	67.1	32.9	0.000	648	4.02	1.070
123	4.16	1	74.7	25.3	0.000	1440	4.23	0.964
129	4.19	1	74.4	25.6	0.000	3560	4.18	0.903

Non-Null Simulations (continued)

Bonferroni Cutoff (3.61)

Univ.	True		%	%	%		Mean	SD
No.	z	Truth	Good	OK	Bad	n	of z	of z
30	4.49	1	79.9	20.1	0.000	3820	4.45	1.040
144	4.95	1	91.5	8.53	0.000	1170	4.97	0.977
164	5.02	1	94.1	5.87	0.000	1680	4.99	0.919
149	5.04	1	94.3	5.73	0.000	507	5.03	0.869
17	5.30	1	98.1	1.87	0.000	2290	5.37	0.839
71	5.90	1	99.5	0.533	0.000	913	5.91	0.887
121	6.25	1	99.6	0.400	0.000	949	6.26	0.998
44	7.39	1	100.0	0.000	0.000	3260	7.42	0.895
46	8.83	1	100.0	0.000	0.000	3570	8.84	1.040
1	9.1	1	100.0	0.000	0.000	6830	9.25	1.080

The means of the z–scores tracked the “true” HEFCE values **almost perfectly**.

The SDs of the z–scores ranged from **0.49** to **1.16**, with a mean of **0.96**.

If we define a **success** as a university where the percentage of time its classification in the simulations matches its “true” status is at least (say) 75%, then the success rate across the 165 universities was **86%**.

Sensitivity to Omitted PCFs

We have adjusted for 8 PCFs—what about a **lurking PCF 9** that we forgot to adjust for?

In other words, how **sensitive** are the findings here to omitted PCFs?

One empirical answer takes the world based on 8 PCFs as **truth** and asks how close working with only $7, 6, \ldots$ PCFs comes to **reproducing** that truth.

(Results in the next three tables are averaged across all possible removals in each row.)

Number of PCFs Removed	**Overall Misclassification** (%) Using Cutoff	
	Bonferroni	HEFCE
0	0.00	0.00
1	7.65	4.92
2	11.52	7.64
3	15.24	11.21
4	19.07	14.71
5	19.66	18.97
6	21.90	24.39
7	23.79	31.21
8	38.79	39.39

When the Bonferroni or HEFCE cutoffs are used, omitting **1** of the 8 PCFs leads to an average overall misclassification rate of only **5–8%** (8–12 universities out of 165), and even with **2** missing PCFs the HEFCE cutoff has an average error rate of only **7.6%**.

Omitted PCFs (continued)

Number of PCFs Removed	**Bad But Not Called Bad** (%) Using Cutoff	
	Bonferroni	HEFCE
0	0.00	0.00
1	4.69	1.04
2	6.47	0.60
3	7.37	0.74
4	8.21	0.71
5	8.04	0.89
6	7.37	1.79
7	7.81	4.17
8	12.50	8.33

Number of PCFs Removed	**Good But Not Called Good** (%) Using Cutoff	
	Bonferroni	HEFCE
0	0.00	0.00
1	8.33	9.38
2	14.29	14.73
3	17.86	20.54
4	19.88	25.00
5	20.39	29.02
6	20.54	33.93
7	21.88	40.63
8	25.00	50.00

When classification errors occur with the Bonferroni or HEFCE cutoffs, they almost all involve **incorrectly labeling a university as “good” when actually it’s “OK”**; the average rate of failing to identify “bad” universities is only **0–2%** with the HEFCE cutoff up to and including **6** omitted PCFs.

Omitted PCFs (continued)

Omitted: **Low HE Participation**
(pseudo–R^2 with progression .019)

	True Status Using Cutoff					
	3.61			HEFCE		
	Bad	OK	Good	Bad	OK	Good
Classified	16	137	12	12	145	8
Bad	14	2	0	12	1	0
OK	2	130	0	0	142	0
Good	0	5	12	0	2	8
(False Neg. %) Overall Error %	(12.5)		5.5	(0.0)		1.8

Omitted: **Parental Occupation** (pseudo–R^2 .004)

	True Status Using Cutoff					
	3.61			HEFCE		
	Bad	OK	Good	Bad	OK	Good
Classified	16	137	12	12	145	8
Bad	15	2	0	12	1	0
OK	1	131	0	0	138	1
Good	0	4	12	0	6	7
(False Neg. %) Overall Error %	(6.3)		4.2	(0.0)		4.8

Omitted: **Entry Qualification** (pseudo–R^2 .049)

	True Status Using Cutoff					
	3.61			HEFCE		
	Bad	OK	Good	Bad	OK	Good
Classified	16	137	12	12	145	8
Bad	15	9	0	12	4	0
OK	1	111	2	0	127	3
Good	0	17	10	0	14	5
(False Neg. %) Overall Error %	(6.3)		17.6	(0.0)		12.7

Omitted PCFs (continued)

Omitted: **Subject of Study** (pseudo–R^2 .009)

	True Status Using Cutoff					
	3.61			HEFCE		
	Bad	OK	Good	Bad	OK	Good
Classified	16	137	12	12	145	8
Bad	15	4	0	12	1	0
OK	1	116	2	0	129	0
Good	0	17	10	0	15	8
(False Neg. %) Overall Error %	(6.3)		14.5	(0.0)		9.7

Omitted: **State School** (pseudo–R^2 .021)

	True Status Using Cutoff					
	3.61			HEFCE		
	Bad	OK	Good	Bad	OK	Good
Classified	16	137	12	12	145	8
Bad	16	1	0	12	1	0
OK	0	133	0	0	140	0
Good	0	3	12	0	4	8
(False Neg. %) Overall Error %	(0.0)		2.4	(0.0)		3.0

Omitted: **Year of Program** (pseudo–R^2 .001)

	True Status Using Cutoff					
	3.61			HEFCE		
	Bad	OK	Good	Bad	OK	Good
Classified	16	137	12	12	145	8
Bad	15	2	0	11	1	0
OK	1	129	4	1	143	0
Good	0	6	8	0	1	8
(False Neg. %) Overall Error %	(6.3)		7.9	(8.3)		1.8

Omitted PCFs (continued)

Omitted: **Age** (pseudo–R^2 .020)

	True Status Using Cutoff					
	3.61			HEFCE		
	Bad	OK	Good	Bad	OK	Good
Classified	16	137	12	12	145	8
Bad	16	2	0	12	0	0
OK	0	128	0	0	142	1
Good	0	7	12	0	3	7
(False Neg. %) Overall Error %	(0.0)		5.5	(0.0)		2.4

Omitted: **Gender** (pseudo–R^2 .004)

	True Status Using Cutoff					
	3.61			HEFCE		
	Bad	OK	Good	Bad	OK	Good
Classified	16	137	12	12	145	8
Bad	16	2	0	12	0	0
OK	0	131	0	0	141	1
Good	0	4	12	0	4	7
(False Neg. %) Overall Error %	(0.0)		3.6	(0.0)		3.0

Overall (averaging across all 8 omitted PCFs)
(entries are % of all 165 universities)

	True Status Using Cutoff					
	3.61			HEFCE		
Classified	Bad	OK	Good	Bad	OK	Good
Bad	9.2	1.8	0.0	7.2	0.7	0.0
OK	0.5	76.4	0.6	0.1	83.5	0.5
Good	0.0	4.8	6.7	0.0	3.7	4.4
Overall Error		**7.7**			**4.9**	

University Summary Reports

The following is a draft version of the sort of report that may be generated by HEFCE in future to help universities **interpret** the annual PI results.

University of Poppleton: Student Progression 1996–97

This summary gives the findings on the **first-year progression rate** for **University of Poppleton** students who commenced a program of study during academic year 1996/1997.

Students are classed as having successfully progressing into their second year if they are **still present at a higher-education establishment** at the start of academic year 1997/1998.

- Poppleton successfully progressed **80.2%** of its starting students for 1996/1997.
- The progression rate for all the universities was **90.1%**.
- After taking into account the **qualifications of entry** and **subject of study** for Poppleton's students, the university's expected progression rate is **85.9%**.

Poppleton's progression benchmark is 85.9% and its actual progression performance is 80.2%. This means that Poppleton is **underperforming** against its benchmark percentage by **5.7 percentage points**.

This difference has been identified as both **statistically** and **practically significant** based on Poppleton's university profile.

University Summary Reports (continued)

This **significant difference** could be due to either or both of two effects:

- Poppleton is **not performing as well as it should do** with its student population, compared to how the rest of the country's universities perform with students similar to Poppleton's; and/or
- Poppleton's student cohort and/or the university itself is **unusual** in ways not taken into account by the analysis, i.e., variables not relating to student qualifications or subject of study.

The following information identifies student types where Poppleton's progression rate differs from that of the rest of the UK HE sector by an amount that is **large in practical and statistical terms**.

These student types should be examined by the university to discover why Poppleton is performing **differently** from the rest of the UK HE sector.

NB (1) This analysis focuses on student cohorts with more than 20 people at Poppleton who are progressing either **above or below expectation**.

A **full table** of how Poppleton is progressing its student types, with regard to subject and qualification, is given in the Appendix.

(2) Cohorts with $|z| > \mathbf{3}$ are singled out.

University Summary Reports (continued)

Problem Student Cohorts

Student Profile			Progression Rate			
Qual.	Subject	n	Poppleton	Sector	Difference	SE
HE	Soc. Sci. & Law	267	11.2%	77.3%	-66.1%	1.9%
Unkn.	Business & Libr.	31	35.5%	81.5%	-46.0%	8.7%
Unkn.	Engin. & Tech.	34	41.2%	84.1%	-42.9%	8.6%
Unkn.	Soc. Sci. & Law	22	45.5%	81.8%	-36.3%	10.9%
Unkn.	Math. & Comput.	25	48.0%	80.0%	-32.0%	10.2%
Unkn.	Comb. & Subj.	39	46.2%	73.7%	-27.6%	8.1%

The largest area of concern relates to the **267** students, with some type of Higher Education qualification on entry, studying Social Sciences and Law. These students are **not progressing as well as expected**. There may be a **misclassification** problem, e.g., these students might be on a single year course and not be tagged as such.

Excellent Student Cohorts

Student Profile			Progression Rate			
Qual.	Subject	n	Poppleton	Sector	Difference	SE
A-L Pts. 8-9	Math. & Comput.	48	95.8%	86.4%	+9.5%	2.9%

Poppleton had **some excellent progression performances** for this year. The **Mathematical and Computer Science** subject areas are progressing certain types of students at a higher rate than expected. These areas should be examined in order to learn more about **why excellent progression rates are being obtained**.

University Summary Reports (continued)

Subject Area Analysis

			Prog. Rates, Adjusted For				
Subject			Qualif., Subject		All 8 Vars.		
Area	n	Obs.	Exp.	Diff.	Exp.	Diff.	z
Medicine	437	89.7%	86.7%	3.0%	86.6%	3.1%	2.2
Biol. & Phys. Sci.	141	85.1%	86.8%	-1.7%	84.2%	0.9%	0.3
Agricult.	72	80.6%	87.7%	-7.2%	82.2%	-1.7%	-0.5
Math. & Comput.	401	85.0%	83.9%	1.2%	83.6%	1.4%	0.9
Engin. & Tech.	353	77.1%	81.9%	-4.8%	80.5%	-3.5%	-1.7
Archit.	44	79.5%	85.3%	-5.8%	82.9%	-3.3%	-0.6
Soc. Stud. & Law	488	41.2%	81.7%	-40.5%	71.3%	-30.1%	-16.4
Busin. & Librar.	294	81.6%	87.2%	-5.6%	85.1%	-3.4%	-1.6
Arts & Design	449	89.5%	89.5%	0.0%	89.5%	0.0%	0.0
Educat.	256	96.1%	92.0%	4.1%	92.5%	3.6%	2.9
Comb. Subj.	723	86.7%	86.1%	0.6%	86.2%	0.5%	0.4

Social Studies and Law are doing unusually badly, but this difficulty was noted earlier as a possible **student misclassification** problem.

Education and **Medicine** are doing unusually well.

Standardization vs. Modeling

How does the HEFCE **standardization** method compare to **model-based** approaches as a function of **interactions included**? Example: the **medium world** ($M = 36$ PCF categories).

	Linear			Linear		
	Main Effects Only			Fully Saturated		
University	α	SE	z	α	SE	z
UWE	0.032	0.003	9.36	0.033	0.003	9.44
Glasgow	0.026	0.005	4.88	0.025	0.005	4.67
City	-0.032	0.010	-3.36	-0.031	0.010	-3.28
Northampton	-0.029	0.007	-4.34	-0.029	0.007	-4.41
Falmouth	0.071	0.019	3.71	0.071	0.019	3.73
Salford	-0.013	0.005	-2.49	-0.013	0.005	-2.47
Central England	-0.029	0.006	-5.08	-0.029	0.006	-5.09
Greenwich	-0.025	0.007	-3.83	-0.025	0.006	-3.87
Abertay Dundee	-0.046	0.010	-4.68	-0.047	0.010	-4.75
Royal Free	0.048	0.030	1.59	0.047	0.030	1.57

	Standardization		
University	$\hat{D}$	SE	z
UWE	0.030	0.003	9.50
Glasgow	0.026	0.005	5.17
City	-0.031	0.009	-3.23
Northampton	-0.028	0.007	-4.19
Falmouth	0.068	0.019	3.57
Salford	-0.013	0.005	-2.37
Central England	-0.028	0.006	-4.99
Greenwich	-0.024	0.006	-3.64
Abertay Dundee	-0.046	0.010	-4.60
Royal Free	0.044	0.030	1.46

	Logistic			Logistic		
	Main Effects Only			Fully Saturated		
University	α	SE	z	α	SE	z
UWE	0.294	0.035	8.47	0.293	0.035	8.46
Glasgow	0.288	0.061	4.74	0.288	0.061	4.74
City	-0.307	0.085	-3.62	-0.307	0.085	-3.63
Northampton	-0.301	0.060	-5.01	-0.299	0.060	-4.99
Falmouth	0.818	0.243	3.37	0.818	0.243	3.37
Salford	-0.146	0.048	-3.01	-0.145	0.048	-3.00
Central England	-0.280	0.050	-5.59	-0.280	0.050	-5.58
Greenwich	-0.239	0.057	-4.21	-0.239	0.057	-4.21
Abertay Dundee	-0.423	0.084	-5.05	-0.424	0.084	-5.06
Royal Free	0.606	0.392	1.54	0.640	0.392	1.54

Standardization vs. Modeling

		Linear				Logistic			
Univ.	**Truth**	F	3	2	1	F	3	2	1
UWE	Good	✓	✓	✓	✓	✓	✓	✓	✓
Glasgow	Good	✓	✓	✓	✓	✓	✓	✓	✓
City	Bad	✓	✓	✓	✓	✓	✓	✓	✓
Northampton	Bad	✓	✓	✓	✓	✓	✓	✓	✓
Falmouth	Good	✓	✓	✓	✓	✓	✓	✓	✓
Salford	OK	✓	✓	✓	✓	×	×	×	×
C. Engl.	Bad	✓	✓	✓	✓	✓	✓	✓	✓
Greenwich	Bad	✓	✓	✓	✓	✓	✓	✓	✓
Abertay Dundee	Bad	✓	✓	✓	✓	✓	✓	✓	✓
Royal Free	OK	✓	✓	✓	✓	✓	✓	✓	✓
“Error” (%)	0	0	0	0	0	10	10	10	10

(This table uses the **Bonferroni** cut-off (2.81) and the **non-model based university status** as truth; F = full model, $k = 1$ means main effects only, and $k = 2, 3$ means main effects plus all k–way interactions.)

Model Description	**“Incorrect” University Status (Out of 165)**	**“Error Rate”**
Linear Full	6	4%
Linear Main	14	8%
Logistic Full	15	9%
Logistic Main	17	10%

In the big world, adjusting only for entry qualification and subject, **model-based** classifications differ from the HEFCE standardization method for **4–10%** of the universities, depending on modeling method and interactions included.

When all 8 adjustors are considered, both **linear** and **logistic** modeling with **main effects only** classify universities into the categories {good, OK, bad} in a way that differs from the HEFCE standardization method for **13%** of the universities.

Extensions

• There are situations in which fixed-effects models **cannot be fit**, e.g., when it is desirable to adjust for **institution-level PCFs** (these will be **confounded** with the institutional **dummy variables** in the fixed-effects models).

A leading alternative to the fixed-effects logistic model (18) in such situations is a **random-effects** formulation:

$$\begin{aligned} \left(y_{ij} \,|\, p_{ij} \right) & \stackrel{\text{indep}}{\sim} \text{Bernoulli} \left(p_{ij} \right) , \qquad (19) \\ \log \left(\frac{p_{ij}}{1 - p_{ij}} \right) & = \beta_0 + \textstyle\sum_{k=1}^{p} \beta_k \left(x_{ijk} - \bar{x}_k \right) + \boxed{q_i} , \\ & \boxed{q_i \stackrel{\text{IID}}{\sim} N\left(0, \sigma_q^2 \right)} . \end{aligned}$$

This model may be fit either by **quasi-likelihood** (QL) methods or in a **Bayesian manner** (e.g., with diffuse priors) via MCMC (and **full likelihood** methods that **integrate** over the random effects are starting to become available too).

Browne and Draper (2004) show that **Bayesian fitting** of models like (19) **can perform substantially better than QL** in terms of bias of estimates and actual coverage of nominal 95% intervals for parameters.

Extensions (continued)

The random effects q_i (the measures of university "quality") will exhibit **shrinkage behavior** relative to their fixed-effects counterparts α_i: extreme q_i at universities with little data will be drawn back toward 0.

This can be regarded as a Bayesian analogue to frequentist **multiple-comparisons adjustment** (using, e.g., Bonferroni) to account for the fact that, in effect, the $\widehat{z}_i$ method is based on $N = 165$ significance tests.

- One advantage of the Bayesian MCMC random-effects formulation is that the **ranks** of the universities can be easily monitored along with the q_i, and this will show clearly **how little can be said** about which universities are "better" than which others.

Extensions (continued)

• It is straightforward to look for **changes over time** in quality by modeling student dropout data **longitudinally**, e.g., through models like

$$y_{ijk} = \mu + \alpha_i^t + \alpha_j^u + \alpha_{ij}^{tu} + \text{covariates}_{ijk} + e_{ijk}, \quad (20)$$

where α_i^t is the **time** effect, α_j^u is the baseline **university** effect, α_{ij}^{tu} is the **interaction** effect time × university, and covariates_{ijk} are the **adjustment variables**.

• Model (19) can be expanded to explicitly acknowledge the possibility of **unexplained student-level variation** (i.e., unmeasured PCFs) by adding **student-level random effects**:

$$\begin{aligned} \left(y_{ij} \,|\, p_{ij} \right) & \stackrel{\text{indep}}{\sim} \text{Bernoulli}\left(p_{ij} \right), \qquad (21) \\ \log\left(\frac{p_{ij}}{1 - p_{ij}} \right) & = \beta_0 + \textstyle\sum_{k=1}^{p} \beta_k \left(x_{ijk} - \bar{x}_k \right) + q_i + \boxed{u_{ij}}, \\ q_i \stackrel{\text{IID}}{\sim} N\left(0, \sigma_q^2 \right), & \ \boxed{u_{ij} \stackrel{\text{IID}}{\sim} N\left(0, \sigma_u^2 \right)}. \end{aligned}$$

Conclusions of Main Case Study

• This work generalizes in obvious ways to **quality assessment in health** (based, e.g., on patient mortality) and other fields.

• $\boxed{\text{Interpreting the } \widehat{z}_i \text{ scores:}}$ a big negative $\widehat{z}_i$ means either (a) that university i is not doing as well with its students as the national average (**causal conclusion**) or (b) that its student cohort and/or the university itself are unusual in ways not measured and correctly adjusted for (**confounding explanation**).

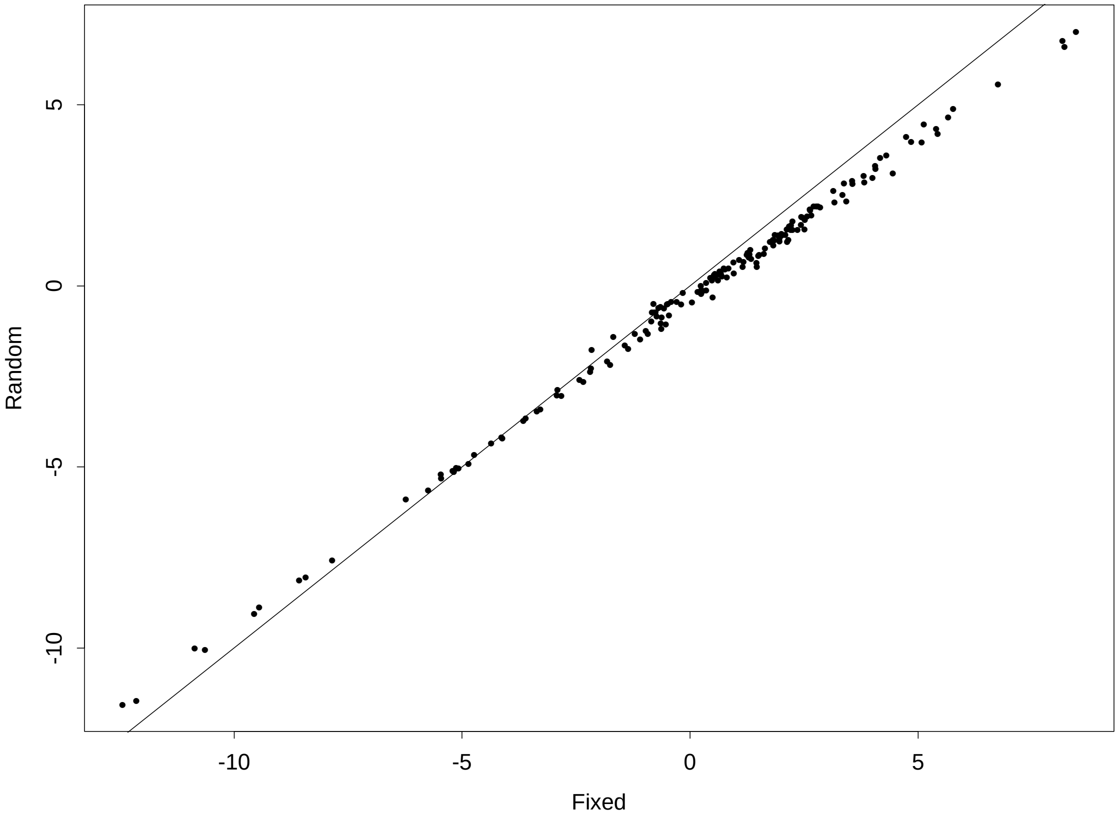
The better the set of PCFs available, the **less plausible** the second explanation becomes.

In our view the current full set of PCFs is rich enough to **cast doubt** on the performance of a number of UK universities on student drop-out, and to single out a number of other universities for **good results**.

Methodological Summary

- With a **binary** outcome variable, indirect **standardization** to the institutional cohort is essentially equivalent to **linear fixed-effects hierarchical modeling** as in model (9) (p. 22), with estimation either by maximum likelihood or Bayesian methods with diffuse priors.

- When both fixed- and random-effects models (e.g., a linear version of model (19), p. 51) can be fit, the rank-ordering of the institutions tends to be **similar**, but the random-effects formulation will typically produce **more shrinkage** (more thought needs to go into where to draw the line between “good,” “OK,” and “bad” institutions using random-effects models; **null simulations** provide a good way forward).



- **Bayesian** fitting of HMs (for profiling or other purposes) can produce **better-calibrated uncertainty assessments** than those from likelihood-based alternatives (e.g., Browne and Draper 2004: random-effects logistic regression models).

Pros and Cons of Institutional Profiling

Potential **advantages**:

- It's (a lot) **cheaper** than **explicit** process measurement—if quality assessment is too expensive it **won't be done**.

- The very **act** of estimating quality (e.g., through profiling) can **raise** quality, by fostering an environment of **healthy criticism** and **improvement**.

Potential **disadvantages**:

- Profiling works by **subtraction**: quality is what's left after we adjust for **the PCFs we remember to measure**—obviously if other major PCFs are unadjusted for the results are **suspect** (but the case study presented here offers a simple method for measuring the sensitivity of the findings to unmeasured PCFs).

- The very **act** of estimating quality can **lower** quality, by creating undesirable **distortions** in the behavior of institutions and individuals.