Holliday's engine mapping experiment revisited: a design problem with 2-level repeated measures data

Harvey Goldstein

University of Bristol, Bristol, UK
E-mail: h.goldstein@bristol.ac.uk

Toby Lewis*

Centre for Statistics,
University of East Anglia, Norwich, UK
E-mail: t.lewis@uea.ac.uk
*Corresponding author

Abstract: Only a limited amount of work has been done on issues of optimum experimental design in hierarchically structured data, in spite of the importance of design efficiency. We illustrate these issues by a fresh analysis of Holliday's well-known engine mapping experiment. This experiment investigated the relation of automobile engine performance, as measured by engine torque vs. spark advance, to adjustable engine parameters or *factors*, viz. engine speed, load, and air/fuel ratio. We show how an optimum experiment can be designed. Our analysis suggests a general procedure of wide applicability.

Keywords: engine mapping experiments; optimum design; multilevel models; repeated measures.

Reference to this paper should be made as follows: Goldstein, H. and Lewis, T. (2006) 'Holliday's engine mapping experiment revisited: a design problem with 2-level repeated measures data', *Int. J. Vehicle Design*, Vol. 40, No. 4, pp.317–326.

Biographical notes: Harvey Goldstein is Professor of Social Statistics at the University of Bristol and Associate Director of the Centre for Statistics at the University of East Anglia. He has been editor of the Royal Statistical Society's Journal, Series **A**, a member of the Society's Council, and was awarded the Society's Guy Medal in Silver in 1998. He was elected a member of the International Statistical Institute in 1987 and a fellow of the British Academy in 1996. He was awarded an honorary doctorate by the Open University in 2002. He is a leading authority on the methodology of multilevel modelling, and author of the major text *Multilevel Statistical Models* (Goldstein, 2003). He is also well known for his statistical work on issues of educational assesssment and school performance measurement.

Toby Lewis is the Director of the Centre for Statistics, University of East Anglia. Previously he was Professor and head of the Department of Statistics at the University of Hull and then at the Open University. He has wide interests in statistics, both applied and theoretical, and has worked in a variety of areas including applications of statistics to biological, geophysical, industrial and forensic problems. He is a leading authority on the methodology of outlying observations. He was elected a member of the International Statistical Institute in 1972, and was awarded an honorary Sc.D by the University of East Anglia in 2000.

1 Introduction

'At the present time there appears to be little empirical or theoretical work on issues of optimum design for multilevel models' (Goldstein, 1995, p.154). In the same book (Goldstein, 1995, p.162) Goldstein wrote: '… the issue of design efficiency has hardly been explored at all though it is an important topic practically'. In the subsequent nine years, some work on design issues with multilevel data has been published, but not a great deal, and much remains to be done. References include Mok (1995), Liski et al. (1996, 1996–97) and Moerbeek et al. (2000, 2001).

The purpose of this paper is to illustrate a design issue for data with a 2-level repeated measures structure. The data relate to the engine mapping process in the automotive industry, i.e., the modelling of engine behaviour as a function of adjustable engine parameters. A specific aim of engine mapping is the tabulation of chosen measures of engine performance in terms of:

- the speed at which the engine is turning (denoted RPM, symbol R)
- the amount of air entering the combustion chamber on each intake stroke (denoted LOAD, symbol L)

and one or more calibration parameters including

- the air/fuel ratio, i.e. the ratio of the airflow rate to the fuel flow rate (denoted AFR, symbol A); also
- the exhaust gas recycling ratio, i.e., the percentage of the total intake mixture which is recycled exhaust; but this parameter is not used in these data.

2 The engine mapping data set

The particular data set we are concerned with is the well known one published by Holliday (1995) and by Holliday et al. (1998), where it was presented and analysed, but not from our multilevel modelling standpoint and not raising the issue of optimum design which it is our aim to illustrate. The repeated measures structure arises in the following way. Repeated observations of *engine torque* vs. *spark advance* are made within each of a series of *spark sweeps*. A spark sweep, in our data set, is an experiment carried out on an engine at constant speed, load and air/fuel ratio, in which observations are made of engine torque, as a measure of the engine's ability to do work, at successive values of spark advance. Spark advance is the angle of rotation of the crankshaft when the spark is fired, measured in degrees (+ or –) relative to when the piston is at the top of the cylinder. A positive/negative value of spark advance means that the spark plug is fired earlier/later in the cycle, when the piston is travelling up/down the cylinder.

In the data set under discussion, there are ten paired observations of spark advance and engine torque in each spark sweep. The number of settings, 10, was laid down for reasons of automotive practice:

> "The data were collected under the present regime so there are ten observations in each spark sweep … We would prefer more observations in each spark sweep … but the data had to be collected this way." (Holliday, 1995, p.26)

Over the range of observed values of spark advance, denoted s, within each spark sweep, the engine torque rises to a maximum and then falls. We use as measure of the engine's ability to do work, not the torque itself but the transformed value $100 \times \log(\text{torque})$, denoted t. The relationship of t to s within a spark sweep is taken to be parabolic. 27 spark sweeps in all were carried out. Thus the data have a hierarchical (or 'multilevel') structure, with ten repeated observations (t, s) within each spark sweep and 27 spark sweeps, giving 270 observations (t, s) altogether. In the usual language of multilevel statistical modelling, the observations (t, s) are called level 1 units and the spark sweeps within which the set of 10 (t, s) values are nested are called level 2 units. (To give another example of level 1 and level 2 units in multilevel data, this time from the educational field: in a survey of examination performance by pupils in a number of schools, the pupils within the schools would be level 1 units and the schools would be level 2 units).

The 27 spark sweeps were carried out at three levels of each of the factors R, L and A. Individual values varied slightly from the nominal levels; these actual values have been used in the analysis. The nominal levels, equally spaced in the case of R and L, were as follows:

Speed R	1,000	3,000	5,000
Load L	0.2	0.4	0.6
Air/fuel ratio A	11:1	13:1	14.5:1

3 Analysis of the data

The object of the experiment is to model one or more suitably chosen measures of engine performance in terms of the three adjustable factors R, L and A, so that the predicted values of each chosen measure can be tabulated over the ranges of practical interest of R, L and A. The design issue here is to use the results of the experiment to choose the levels of the adjustable factors so as to reduce the amount of data, and hence of future experimentation, which would be required for effective mapping over the ranges of practical interest of the three factors.

There are two main stages to the analysis. For each of the 27 spark sweeps there is a different combination of factor values (R, L, A).

- First, measures of engine performance are calculated for each spark sweep, using a multilevel statistical model.
- Then, these measures are related to R, L and A using a linear regression analysis, in the equivalent form of an analysis of variance.

As regards measures of engine performance we have, for a given spark sweep, ten paired observations of engine torque and spark advance, or equivalently, ten paired observations

(t, s) of $100 \times \log$(engine torque) (t) and spark advance (s). In a graph of t against s, the ten points (t, s) rise to a maximum and then fall, in a relationship which we assume to be parabolic. By fitting a parabola to the ten points we get an estimated value, call it x, of the value of spark advance s giving the maximum torque, and an estimated value, call it y, of the transformed value of this maximum torque:

$$y = 100 \times \log (\text{max torque}) = t_{\text{max}}.$$

We first model the observed values of $100 \times \log$ (torque) (t_{ij}) as a function of spark advance (s_{ij}) as follows:

$$\left.\begin{aligned} & t_{ij} = \beta_0 + \sum_{k=2}^{27} \beta_k \gamma_{kj} + \alpha_{1j} s_{ij} + \alpha_{2j} s_{ij}^2 + e_{ij} \\ & \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix}_j \sim N\begin{pmatrix} \mu_1 & \sigma_1^2 & \\ \mu_2 & \sigma_{12} & \sigma_2^2 \end{pmatrix}, e_{ij} \sim N(0, \sigma_e^2) \end{aligned}\right\}. \tag{1}$$

Here $i(= 1, 2, \ldots, 10)$ indexes the observation within a sweep, and $j(= 1, 2, \ldots, 27)$ indexes the sweep; the γ are dummy variables for the spark sweeps. We are modelling the relationship in sweep j of the ten values of t_{ij} to the ten corresponding values of s_{ij} by a parabolic curve with coefficients of s_{ij} and s_{ij}^2, α_{1j} and α_{2j} respectively. Our model fits a separate intercept for each sweep since the distribution of the mean torque across sweeps does not follow any easily recognisable pattern. The summation term $\sum_{k=2}^{27} \beta_k \gamma_{kj}$ in (1) does not include a term for $k = 1$ because the fitted intercept for sweep $j = 1$ is β_0; $N(\mu, \sigma^2)$ denotes the normal distribution with mean μ and variance σ^2; $N\begin{pmatrix} \mu_1 & \sigma_1^2 & \\ \mu_2 & \sigma_{12} & \sigma_2^2 \end{pmatrix}$ denotes the bivariate normal distribution with means μ_1, μ_2, variances σ_1^2, σ_2^2, and covariance σ_{12}; ~ denotes 'distributed as'. Model (1) is a standard 2-level repeated measures model with a second-order polynomial having random coefficients, except that the intercept terms are represented by a set of dummy variables rather than assuming that they have a normal distribution across sweeps.

Having estimated the parameters in equation (1), we then estimate the random effects (residuals) α_{1j}, α_{2j} for each sweep, so that for each sweep we have a predicted response curve given by substituting the parameter estimates ($\beta_0, \beta_2, \ldots, \beta_{27}$) and estimated residuals $(\hat{\alpha}_{1j}, \hat{\alpha}_{2j})$ in equation (1).

Write the resulting prediction equation as

$$t_{ij} = a_j + b_j s_{ij} + c_j s_{ij}^2 \tag{2}$$

then the value, x_j, of s for which the predicted response in sweep j is a maximum is

$$x_j = -b_j / 2c_j \tag{3}$$

and the maximum predicted response is

$$y_j = a_j - b_j^2 / 4c_j. \tag{4}$$

These estimated values of y_j and of x_j are then used separately as responses, each in a three way (3 × 3 × 3) analysis of variance for the three factors *R*, *L* and *A*, each of which has three levels. It is convenient to give the three levels of factor *R* the values $r = 0, 1, 2$, those of factor *L* the values $l = 0, 1, 2$ and those of factor *A* the values $a = 0, 1, 2$. Each of the 27 values of *j*, indexing the 27 sweeps, corresponds to one of the 27 combinations of *r*, *l* and *a*, viz.

rla = 000, 100, 200, 010, 110, 210, …, 122, 122.

The analysis of variance for the transformed maximum torque values y_j is set out in Table 1. The sum of squares r_{lin} measures how far the variation in *y* is accounted for by a linear dependence on *r*, i.e., by a straight-line relationship between *y* and *r*; the sum of squares r_{quad} measures how far the variation in *y* is further accounted for by a quadratic dependence on *r*, i.e., by a parabolic relationship between *y* and *r*; and similarly for l_{lin}, l_{quad}, a_{lin}, a_{quad}. The sums of squares $r_{\text{lin}} \times l_{\text{lin}}$, $r_{\text{quad}} \times l_{\text{lin}}$, etc., measure the various 2-factor interactions. In Table 1, the 2-factor interactions between *r* and *a* are non-significant at 5% and have been included in the residual; likewise the three 2-factor interactions $l_{\text{lin}} \times a_{\text{lin}}$, $l_{\text{quad}} \times a_{\text{lin}}$, $l_{\text{quad}} \times a_{\text{quad}}$.

The analysis of variance in Table 1 is equivalent to a regression analysis of the 27 values y_j on the 12 predictor (or 'explanatory') variables

$$r, r^2, l, l^2, a, a^2; \quad rl, r^2l, rl^2, r^2l^2; \quad la, la^2.$$

The regression equation is

$$\left.\begin{aligned} y_j &= 222.2 + 19.1r - 24.1r^2 + 154.0l - 33.8l^2 + 14.2a - 6.24a^2 \\ &\quad -14.3rl + 24.0r^2l + 4.24rl^2 - 7.10r^2l^2 - 5.57al + 3.02a^2l \end{aligned}\right\} \tag{5}$$

with standard deviation of fit $\sqrt{(5.85)} = 2.42$, as in Table 1.

In the regression, the respective contributions of the 12 predictor variables to the total variation of the *y* values are the same as the corresponding values in the 'sum of squares' column in Table 1.

Table 1 Analysis of variance for the transformed maximum torque values y_j

Sum of squares		*df*	*Mean square*
r_{lin}	2538.3	1	2538.3
r_{quad}	851.2	1	851.2
l_{lin}	167042.0	1	167042.0
l_{quad}	10288.2	1	10288.2
a_{lin}	82.8	1	82.8
a_{quad}	62.3	1	62.3
$r_{\text{lin}} \times l_{\text{lin}}$	2287.8	1	2287.8
$r_{\text{quad}} \times l_{\text{lin}}$	384.0	1	384.0
$r_{\text{lin}} \times l_{\text{quad}}$	396.6	1	396.6
$r_{\text{quad}} \times l_{\text{quad}}$	67.2	1	67.2
$l_{\text{lin}} \times a_{\text{quad}}$	36.5	1	36.5
Residual	87.8	15	5.85
Total	184124.7	26	–

For the practical purpose of setting up tabulations of y (or similarly of x) with entry values of speed, load and air/fuel ratio, the regression equation (5) is used. For instance, if tabulation was required for (say) $r = 1.5$, $l = 1.5$ and $a = 0.5$, the predicted value of y would be $222.2 + 19.1 \times 1.5 - 24.1 \times 2.25 + \ldots = 381.3$. This illustrates the calculation of a tabulated value of y at levels or r, l and a different from those set in the experiment. But tabulation for values or r, l and a outside the ranges 0 to 2 set in the experiment should be avoided. Turning now to the interpretation of the various component sums of squares in Table 1, the mean responses (mean y values) at each of the three levels of r, l and a and at each of the nine levels of $r \times l$ and of $l \times a$ are shown in Table 2. For example, the first entry 226.0 is the average when $r = 0$ and $l = 0$ of the three values of y as a varies over 0, 1, 2. The row 'all r' shows that the average of the nine values of y as l and a vary over 0, 1, 2 is 323.0 when $r = 0$ and 323.0 again when $r = 1$, but 299.3 when $r = 2$. These three averages cannot be adequately fitted by a straight line relationship with r, and a quadratic term is required in the model as well as a linear term, as shown by the sums of squares for r_{lin} and r_{quad} in Table 1. As another example, consider the averages over the three values of a of the three values of y for fixed r and l. These are

226.0, 221.0, 168.0 respectively for $r = 0, 1, 2$ when $l = 0$

but 397.5, 400.6, 394.8 respectively for $r = 0, 1, 2$, when $l = 2$,

a quite different pattern, with the value for $r = 2$ much lower than those for $r = 0$ and 1 when $l = 0$ but not when $l = 2$. So there is a significant interaction between the factors r and l, as quantified by the four $r \times l$ interaction components in Table 1.

Corresponding to Table 1, the analysis of variance for the values x_j of spark advance giving maximum torque is set out in Table 3, with sums of squares r_{lin}, r_{quad}, l_{lin}, l_{quad}, a_{lin}, a_{quad} ('main effects') and with 2-factor interactions $r_{\text{lin}} \times l_{\text{lin}}$, $r_{\text{quad}} \times l_{\text{lin}}$. All the other 2-factor interactions are non-significant at 5% and have been included in the residual. Corresponding to Table 2, the mean x-values at each of the three levels of r, l and a and at each of the nine levels of $r \times l$ are given in Table 4.

4 Experimental design aspects

We assume that the experiment was a pilot for the setting up of tabulations of y and x for practical use, with entry values of speed, load and air/fuel ratio covering the respective ranges sampled in the experiment; these were:

- speed from 1,000 to 5,000, coded r from 0 to 2
- load from 0.2 to 0.6, coded l from 0 to 2
- air/fuel ratio from 11:1 to 14.5:1, coded a from 0 to 2.

For each of r, l and a, three equispaced levels were used in the experiment, coded 0, 1, 2, making 27 independent spark sweeps in all.

Table 2 Means of transformed maximum torque values at each of the 3 levels or r, l and a and at each of the 9 levels of $r \times l$ and $l \times a$

l/r	*0*	*1*	*2*	*all r*
0	226.0	221.0	168.0	205.0
1	345.6	347.5	335.1	342.7
2	397.5	400.6	394.8	397.6
all *l*	323.0	323.0	299.3	315.1
l/a	0	1	2	all *a*
0	200.9	209.5	204.5	205.0
1	340.0	344.2	343.9	342.7
2	394.7	398.1	400.1	397.6
all *l*	311.9	317.3	316.2	315.1

Table 3 Analysis of variance for the values x_j of spark advance giving maximum torque

Sum of squares		*df*	*Mean square*
r_{lin}	663.7	1	663.7
r_{quad}	8.9	1	8.9
l_{lin}	666.4	1	666.4
l_{quad}	41.9	1	41.9
a_{lin}	61.5	1	61.5
a_{quad}	1.0	1	1.0
$r_{lin} \times l_{lin}$	3.4	1	3.4
$r_{quad} \times l_{lin}$	21.5	1	21.5
Residual	24.6	18	1.37
Total	1493.0	26	–

Table 4 Means of values of spark advance giving maximum torque at each of the three levels of r, l and a and at each of the nine levels of $r \times l$

1/r	*0*	*1*	*2*	*all r*
0	31.2	35.1	42.0	36.1
1	20.4	28.7	33.1	27.4
2	16.4	26.0	29.3	23.9
all *l*	22.7	29.9	34.8	29.1
a	0	1	2	
	27.4	28.8	31.1	29.1

As regards the transformed maximum torque y, Tables 1 and 2 show that the response was much the same for $r = 1$ as for $r = 0$ at all three levels of l, but that for $r = 2$ it was significantly less than for $r = 1$ at $l = 1$ and very significantly less at $l = 0$. Thus the value of y varied substantially over the area $1 < r < 2$, $0 < l < 1$ of the (r, l) plane. But we do not know the nature of the variation, since this region of values of (r, l) was unexplored by the experiment. A revised experiment is required, which has been designed to include an exploration of this region. For example, we might advantageously carry out just 25 spark sweeps (less than the 27 in Holliday's experiment), with levels of r (say) 0, 1, 4/3, 5/3, 2 combined with levels of l (say) 0, 1/3, 2/3, 1, 2.

Note that the low entry 168.0 for $r = 2$, $l = 0$ in Table 2 is not a single value which might be suspected of coming from an aberrant spark sweep; it is the average of three separate determinations 163.0, 174.6, 166.2 at $a = 0$, $a = 1$, $a = 2$ respectively, all three values very similar.

As regards the factor a, there would be no need for a revised experiment to carry out separate spark sweeps for every $r \times l \times a$ combination, since there no significant $r \times l \times a$ interaction. Taking, for example, the revised experiment with 25 spark sweeps suggested above, one would include five different levels of a, say $a = 0$, ½, 1, 1½, 2. The spark sweep at any given (r, l) combination would be carried out at just one of these levels of a, e.g. as in Table 5. The levels c, d, e, f, g of a would be some random permutation of the actual levels 0, ..., 2. Three separate analyses would then be performed on the 5 × 5 2-factor tables $r \times l$, $r \times a$ and $l \times a$.

Table 5 Typical allocations of levels of a to levels of $r \times l$ in improved experimental design

l/r	*0*	*1*	*4/3*	*5/3*	*2*
0	c	d	e	f	g
1/3	g	c	d	e	f
2/3	f	g	c	d	e
1	e	f	g	c	d
2	d	e	f	g	c

Consider now the planning of tabulations of x, the value of spark advance which gives maximum torque in a spark sweep. Tables 3 and 4 show that, as with the y values, there is no significant $r \times l \times a$ interaction, so a similar 5 × 5 design to that recommended above can be used. The previous values suggested for r, l and a would not be ideal; a more informative choice of factor levels would be:

$r = 0$, ½, 1, 1½, 2;

$l = 0$, 0.45, 0.9, 1.35, 2;

$a = 0$, 0.6, 1.2, 1.6, 2.

5 Discussion

We have used Holliday's well known engine mapping data set to illustrate some issues of optimum design in multilevel models. In the engine mapping experiment, the effects on engine performance of three factors, speed R, load L and air/fuel ratio A, were investigated. The levels of each factor were chosen in the experiment to be equally spaced. We have shown that this is an unsatisfactory design, and that a good design would have required a choice of factor values which were not equally spaced.

A further design aspect is the choice of spark advance values (s) in each spark sweep, in particular their central location and range. There were two response values to be analysed as measures of engine performance, which we have denoted as x_j and y_j. These are not actual observations, but functions of the actual observations. The actual observations were ten repeated measurements of torque vs. spark advance within each spark sweep, for 27 spark sweeps. A parabolic relation between $t = 100 \times \log$ (torque) and s = spark advance was assumed for each spark sweep:

$$E(t) = a_j + b_j s + c_j s^2 \text{ for the } j\text{th sweep.}$$

The estimates of a_j, b_j, c_j were therefore functions of the actual observations, and so also were the estimates of x_j and y_j, which were functions of a_j, b_j, c_j given by equations (3) and (4). These values of x_j and y_j were then analysed as responses in separate analyses of variance.

For a well designed experiment, the ten values of s in a spark sweep need to be chosen to produce a good estimate of maximum t, and hence good estimates of the statistics a, b and c. The values of s therefore need to be distributed over a range consistent with a parabolic (s, t) relationship and effectively centred on maximum t. This was achieved in many of the 27 spark sweeps, but not in all. As an example of an unsatisfactory choice of s-values, see Table 6 which gives the values of s, torque, and t, for spark sweep 604 (j = 19).

Table 6 Values of s, torque and t for spark sweep 604 (j = 19, rla = 102)

s	–0.9	4.8	10.9	16.4	21.9	27.4	32.9	38.6	44.1	49.6
torque	0.8	0.9	3.7	5.9	6.9	7.5	7.8	7.8	7.8	7.5
t	–22	–11	131	177	193	201	205	205	205	201

In our analyses of variance (Tables 1 and 3), each value of y or x can be regarded as a response with measurement error. The measurement errors, which are, in fact, calculated from equations (2)–(4), are not the same for each y or x, since they depend on where the maximum torque is situated in the range of spark advance values in the sweep. This would affect such matters as the calculation of confidence intervals and significance tests. However, the estimates of the fixed effects in the model are consistent and unbiased and the design message in Section 4 holds good.

In conclusion, we have presented our analysis of this particular experiment as an example of the gains to be made by optimal choice of factors, levels and settings of predictor variables; in other words, the importance of optimum experimental design. In our example, the information obtainable from the experiment could have been increased for a reduced experimental effort by adopting a more efficient design. Our methods could be used, with benefit, in other similar situations.

Acknowledgements

We wish to thank Dr. Antony Fielding, Dr. Dan Grove and Dr. David Stevens for valuable discussions. We are especially grateful to Professor Anthony Lawrance for his advice and for providing us with the data and other indispensable material.

We also wish to thank the three referees for their very helpful comments which have enabled us to substantially improve the paper.

The support given by the Economic and Social Research Council to our research is gratefully acknowledged.

References

Goldstein, H. (1995) *Multilevel Statistical Models*, 2nd ed., Edward Arnold, London; Wiley, New York.

Goldstein, H. (2003) *Multilevel Statistical Models*, 3rd ed., Edward Arnold, London.

Grove, D.M. and Davis, T.P. (1992) *Engineering Quality and Experimental Design*, Longman, London.

Holliday, T. (1996) *The Design and Analysis of Engine Mapping Experiments: A Two-Stage Approach*, unpublished PhD thesis, University of Birmingham, School of Mathematics and Statistics.

Holliday, T., Lawrance, A.J. and Davis, T.P. (1998) 'Engine-mapping experiments: a two-stage regression approach', *Technometrics*, Vol. 40, pp.120–126.

Liski, E.P., Luoma, A. and Sinha, B.K. (1996) 'Optimal designs in random coefficient linear regression models', *Calcutta Statist. Assoc. Bull.*, Vol. 46, pp.211–219.

Liski, E.P., Luoma, A., Mandal, N.K. and Sinha, B.K. (1996–97) 'Optimal design for an inverse prediction problem under random coefficient regression models', *J. Indian Soc. Agricultural Statist.*, Vol. 49, pp.277–288.

Moerbeek, M., van Breukelen, G.J.P. and Berger, M.P.F. (2000) 'Design issues for experiments in multilevel populations', *J. Educ. Behav. Statist.*, Vol. 25, pp.271–284.

Moerbeek, M., van Breukelen, G.J.P. and Berger, M.P.F. (2001) 'Optimal experimental designs for multilevel logistic models', *The Statistician*, Vol. 50, pp.17–30.

Mok, M. (1995) 'Sample size requirements for 2-level designs in educational research', *Multilevel Modelling Newsletter*, University of Bristol, Vol. 7, No. 2, pp.11–16.