Letters

Phenformin Ban

Gina Bari Kolata (News and Comment, 16 Mar., p. 1094) describes the circumstances leading to removal from the market of phenformin as an "imminent hazard to the public health." What Kolata does not make clear and, indeed, perhaps only those directly involved with the phenformin decision would readily recognize, is that the Food and Drug Administration (FDA) felt that, regardless of the method chosen by Health, Education, and Welfare Secretary Califano to remove phenformin from the general market, the drug should continue to remain available for those few patients in whom the benefits would appear to outweigh the risks. In choosing the option to suspend approval of the applications *for general marketing* of phenformin on the basis that the drug is an "imminent hazard," Secretary Califano was not ruling out permanently all use of the drug. The imminent hazard provision of the law is simply a legal mechanism for suspending marketing while an ultimate solution to an important safety problem is arrived at through the usual processes of a hearing and subsequent court review.

Since the intent was always to permit the drug to be available to a limited population, it is not at all "ironic" that approximately 3000 patients have received phenformin under an ongoing investigational new drug application. It is noteworthy that this represents only 1 percent of the population that received phenformin during any 1 year when the drug was freely available on the market. We attribute this low figure to the fact that most maturity-onset diabetics can be treated successfully with other modes of therapy—diet, insulin, sulfonylureas—and that FDA has developed stringent criteria for patient eligibility to receive phenformin.

Kolata quotes me as stating that calling phenformin an investigatory drug is our way of restricting its distribution. The conditions under which the drug is permitted to be distributed do, of course, confine its use to a small number of individuals. The investigational drug application for phenformin, however, provides another useful purpose.

Because one of the conditions under the application is that physicians are required to report instances of suspected or confirmed lactic acidosis, it permits us to assess the incidence of lactic acidosis when the drug is confined to patients who do not have certain risk factors for lactic acidosis and who receive the drug in dosages associated with a diminished risk for lactic acidosis.

MARION J. FINKEL
Bureau of Drugs,
Food and Drug Administration,
Rockville, Maryland 20857

Statistical Quality

Utah State University offers a course entitled "Quantitative Methods of Natural Resource Management" for juniors in the College of Natural Resources. The take-home exam in this course requires that the student "find an article in your area of professional development in which the hypothesis H_0: $\mu = a$, σ unknown, is tested." These students have had a university statistics course which includes tests of hypotheses and the subject is reviewed in this methods course. Students seek these articles in range, wildlife, forestry, watershed management, outdoor recreation, science, and ecology journals.

The students often find one or more of the following problems:

- The hypothesis being tested is not stated (clearly).
- The test being applied is not identified (clearly). There are numerous tests based on the t-statistic.
- There are not enough data presented to check the application of the test.
- The assumptions that underlie the test are not mentioned, and the design does *not* make it clear that those assumptions are met.
- There are not enough intermediate results (standard deviations, standard errors, numbers of samples, and the like) to check the application of the test.
- The interpretation of the test results is inconsistent with their understanding of what the author did.
- The wrong test is used, the test is incorrectly applied, the calculated values do not follow from the data, and the results are incorrectly interpreted.

The last situation occurs in about 20 percent of the papers where enough information is presented for the student to repeat the test.

This is the first serious look that many students take at the literature in their field. They often tell me that they were afraid of those journals because they thought the material was too esoteric for them. Many are let down by the fact that the published material asks them to accept the conclusions on faith; an objective evaluation is impossible. Some students who have trouble with statistics take heart from the fact that apparently the professionals have not mastered it either; they even challenge my insistence that they learn it.

It is my humble opinion that a smaller number of publications done well would better support good science than this large number of papers done poorly. It also seems that a special class of reviewers (perhaps staff people) need to check quantitative results for assumptions, correct application, correct calculations, correct interpretations, and so forth.

GEORGE S. INNIS
Department of Wildlife Sciences,
College of Natural Resources,
Utah State University,
Logan 84322

Burt's Tables

In "The Cyril Burt question: new findings" (29 Sept. 1978, p. 1177), D. D. Dorfman has analyzed an article by Burt and claims to have found evidence that he "fabricated data," that his frequency distributions were "systematic constructions." The article has in turn led to rather less charitable characterizations in newspaper headlines (often involving the word "fraud"). All of this is unfortunate, in that Dorfman is in error on two major points, and his other points are sufficiently open to reasonable doubt to call his conclusions into serious question.

First, I wish to call attention to a significant misrepresentation of Burt in Dorfman's section entitled "Burt's row totals." Dorfman writes, "The row totals of Burt's tables I to IV and the column totals of his tables III and IV would appear on the basis of Burt's descriptions and discussions to be simply totals per mille." He then goes on to show that the row totals (the proportions in each class) agree perfectly with 1926 data of Spielman and Burt, saying "the coincidence is bizarre indeed." Dorfman's contention that Burt described his row totals as "simply totals per mille" is simply wrong, and his conclusion that the agreement is "bizarre" is uncalled for. What Burt in fact wrote was (*1*, p. 10):

> In constructing the tables the frequencies inserted in the various rows and columns were proportional frequencies and in no way represent the number actually examined: from class I the number actually examined was nearer a hundred and twenty than three. To obtain the figures to be inserted (numbers per mille) we weighted the actual numbers so that the proportions in each class should be equal to the estimated proportions for the total population.

Presumably he got the "estimated proportions" from Spielman and Burt, or some other publication of these data. In other words, Burt is *saying* he has weighted the counts to get precisely the agreement that Dorfman presents as evidence of fabrication.

One of Dorfman's major arguments relates to Burt's column totals, his grouped intelligence assessments for all classes together. Dorfman demonstrates convincingly that Burt's column totals fit a normal distribution exactly, if rounding is allowed for. What does Burt say about that? Immediately after the above-quoted passage, Burt wrote (*1*, p. 10):

> Finally, for purposes of the present analysis we have rescaled our assessments of intelligence so that the mean of the whole group is 100 and the standard deviation 15. This is done because the results of so many intelligence tests nowadays are expressed in terms of conventional I.Q.'s conforming to these requirements.

The question is, what did Burt mean by "rescaled"? For if he meant that he followed the by no means unknown practice of "curving" the scores to fit a normal curve (with mean 100 and standard deviation 15), either by transforming his pooled (over classes) IQ's individually or by reweighting his columns (as he did his rows) to fit "estimated proportions," then Dorfman's case collapses. It is clear that for his purposes Burt would need the father scores and son scores to be comparable. It seems plausible that if the raw data (which were gathered over a half century under widely varying conditions) were in fact "crude" (and thus possibly skew or otherwise markedly nonnormal) and the assessments of adult intelligence "less thorough and reliable" (*1*, p. 9) (and hence possibly more variable), then Burt would rescale the marginal totals to agree with one another so that he could make direct comparisons of within-row variability. Contrary to Dorfman's implications, Burt did believe IQ's were approximately (though not exactly) normal (*2*), and in a "pilot study" he might well choose to rescale using a normal distribution (*3*). In fact, Dorfman's table 7 could be interpreted as showing just how Burt might have proceeded in scaling tables III and IV. It may be significant that Burt used the word "rescale" instead of "standardized," which would be more suggestive of a linear rescaling. (In psychometric terminology, "scaling" has a more general connotation than a simple linear transformation, although linear rescaling is a special case.) It must be admitted that Burt's description of his procedure is extremely vague—one cannot even determine his sample size from his description. Dorfman reports that the consensus is that Burt's sample size was 40,000 pairs. This number must be an error (although not one that originated with Dorfman), based on Burt's statement that for class I the number examined was "nearer a hundred and twenty than three," a 40 to 1 ratio. But he reweighted different classes with different weights and probably chose an extreme ratio to emphasize disparity. Burt made no further statement about the actual number of pairs, but it may have even been less than 1000. One can easily see how many readers could be misled into believing the counts were frequencies, but Dorfman does Burt and the readers of *Science* a great disservice by not even mentioning a reasonable alternative explanation that does not involve either fabrication or fraud.

Another of Dorfman's major errors involves his calculation of regression coefficients. In table 3 he calculates $\bar{X}_c/(\bar{X}_f + 100)$ for each class, where $\bar{X}_c$ is the mean IQ for children and $\bar{X}_f$ is the mean IQ for fathers, and seems surprised at the coincidence that all answers are (to two decimal places) 0.50. He has used the wrong formula. (I thank David L. Wallace for pointing this out to me.) If we really wish to estimate the regression coefficients based on these limited data, we should presumably calculate $(\bar{X}_c - 100)/(\bar{X}_f - 100)$, which gives (to two decimal places) 0.52, 0.48, 0.49, 0.56, 0.50, and 0.49. Dorfman's formula is nonsensical; it adds 100 to the numerator and 200 to the denominator, which biases the results toward 0.50, regardless of the actual data.

As a final point, I note that Dorfman's logic is seriously at fault in his "conclusions." He writes, "These findings show, beyond any reasonable doubt, that Burt fixed the row and column totals of the tables in his highly acclaimed 1961 study. Since the totals are completely determined by the cell entries, Burt determined the cell entries." I have argued that Burt announced that he had weighted the rows to get predetermined row totals and hinted that he had done the same for the columns. But this does not at all imply that he fabricated the individual entries. If a two-way table is reweighted along rows and columns as Burt appears to have done, the individual entries will still be estimates of rates per 1000 for the corresponding cells of the bivariate frequency distribution for the entire population, and it is no abuse of statistical terminology to still refer to the table as "data," as Burt did. The entries will no longer be frequency counts, but, as Burt announced, the cell entries "in no way represent the number actually examined" (*1*, p. 10). In fact, if Burt really did fabricate his data, he chose an extremely difficult way to do it. He would have had to *first* invent a two-way table and *then* rescale his original table to get the predetermined marginal totals. It would be far easier to merely invent a table and skip the rescaling, or at least not bother with a precise rescaling. Contrary to Dorfman's implication, it is not a simple matter to fill in either a 6×10 or a 6×6 table with predetermined marginal totals and get a plausible correlational pattern. A careful inspection of Burt's tables will reassure the reader that they cannot be perfectly fitted by a bivariate normal distribution. For example, in row VI of table I, 45 percent of the scores fall between 80 and 90, suggesting a standard deviation of less than 8.5. Since the mean score in this class is given as 84.9, no more than 0.2 percent of the scores should exceed 110, whereas we are told that 9/261 or 3.4 percent did so.

I do not wish to be interpreted as endorsing either Burt's statistical procedure or his unclear explanation of what he did (and his refusal to present the raw data), but given the standards of Burt's time (*4*) and his repeated disclaimers (it was "merely a pilot inquiry," "data are too crude and limited") (*1*, p. 9), the charges of fabrication or fraud seem, at least in this instance, to be without foundation; the evidence presented is irrelevant to the case.

STEPHEN M. STIGLER
Center for Advanced Study in the Behavioral Sciences, Stanford, California 94305

References and Notes

1. C. Burt, *Br. J. Stat. Psychol.* **14**, 3 (1961).
2. ——, *ibid.* **13**, 98 (1943); *Mental and Scholastic Tests* (King, London, 1922), p. 162.
3. See, for example, J. P. Guilford, *Psychometric Methods* (McGraw-Hill, New York, ed. 2, 1954), pp. 345-346.
4. Burt's paper was published in 1961, but his statistical training was rooted in the 1920's and 1930's, when the normal distribution was considered to be a more "normal" (≡ usual) state than is now the case and was used much more freely as a basis for test scaling. The only statistical reference in Burt's 1961 paper is the 1934 edition of R. A. Fisher's *Statistical Methods for Research Workers*, an excellent work but one that lays considerable stress upon procedures based upon the normal distribution.

I am distressed by the application of statistics in "The Cyril Burt question: new findings" by D. D. Dorfman. Although I have no intention of defending Burt, I think that Dorfman's analyses provide no evidence for his claim that "the eminent Briton is shown, beyond reasonable doubt, to have fabricated data on IQ and social class". . . .

Fixed margins of a table do not "determine" cell entries as Dorfman supposes, even in a 2×2 table. Furthermore, fixing margins is a useful and accepted tool in many statistical problems (*1*). . . .

It is likely that Burt (a) transformed the IQ data to follow a normal distribution with mean 100 and standard deviation 15 for both parents and children, (b) fixed the class margins in his tables I to IV at census data, but (c) did not otherwise alter the data. Burt's descriptions of his tables (*2*, pp. 10, 12, 15) imply (a) and (c) and clearly state (b). Also, (a), (b), and (c) are consistent with (i) the excellent but imperfect fit to normality displayed in the IQ margins of tables I to IV, (ii) the slight differences between parents' IQ margins and children's IQ margins in the tables, and (iii) the slight differences between the IQ margin and the class margin in table III as well as in table IV. The hypothesis that Burt fixed both row and column margins at population proportions is inconsistent with (i), (ii), (iii), and his description of the tables (*2*, pp. 12, 15).

Although Dorfman's statistics do not provide any evidence that Burt fabricated data, there may be such evidence in Burt's tables. Assuming that Burt first normed IQ and then fixed the class margins, his tables I and II present the same data as his tables III and IV but with different boundaries for the IQ categories. By combining the information from tables I and III (parents' data) and from tables II and IV (children's data), it is possible to calculate frequencies in narrower IQ categories. (For example, the frequency 86 for parents having IQ's between 100 and 103 is found by subtracting the sum of frequencies in IQ categories 50-100 in Burt's table I from the sum in IQ categories 50-103 in his table III.) The results are shown here in Table I.

Burt's ambiguous labeling of categories makes it difficult to compare pre-

Table 1. Distribution of IQ's of parents and children from Burt's tables I to IV, and theoretical normal distribution with mean 100 and standard deviation 15.

Subjects	50–60	60–70	70–80	80–90	90–91	91–100	100–103	103–110	110–115	115–120	120–127	127–130	130–140	140–141	141+
Parents	1	23	69	160	8	239	86	162	96	66	56	11	21	0	2
Children	2	22	70	159	8	242	83	164	94	66	56	12	21	0	1
N (100,15)	3	19	68	162	21	226	79	168	94	68	55	13	19	1	3

cisely the observed frequencies in Table 1 with the theoretical normal frequencies (how was a person with IQ 90 or 100 or 110 classified?). Nevertheless, the counts in the 90-91 category look suspiciously small. Furthermore, the parents' and children's margins follow each other more closely than they follow the normal frequencies, with differences between parents' and children's frequencies being regularly "corrected" (that is, the frequencies are identical with the categorization 50–70, 70–90, 90–91, 91–103, 103–115, 115–120, 120–127, 127+). Even assuming bivariate normality with known correlation between parents' and children's IQ's, precise test statistics are difficult to derive because sample sizes in each class are unknown (except the approximate 120 for parents in class I). Yet, if the IQ data are approximately N (100, 15), the patterns in our Table 1 are suspicious, and if the IQ data are not approximately normal, the excellent fits of the IQ margins in Burt's tables I, II, III, and IV to the N (100, 15) model are suspicious.

Moreover, by forming tables analogous to Table 1 within each social class, we uncover a blatant inconsistency between Burt's tables. From his table I we have in class VI 20 with IQ's greater than 100, and from his table III we have in class VI 24 with IQ scores greater than 103. This inconsistency may be the result of recording or computational errors, but it may also be the result of trying to create the entries of tables from specified margins, as Dorfman suspects.

DONALD B. RUBIN
Educational Testing Service,
Princeton, New Jersey 08541

References

1. Y. M. M. Bishop, S. E. Fienberg, P. W. Holland, *Discrete Multivariate Analysis: Theory and Practice* (MIT Press, Cambridge, Mass., 1975), section 3.6, "Classical uses of iterative proportional fitting."
2. C. Burt, *Br. J. Stat. Psychol.* **14**, 3 (1961).

I will begin with my response to Stigler's letter. I will first evaluate his general position and then answer his specific criticisms.

Stigler initially appears to argue that Burt told the reader what he had done and that therefore my charges of fraud are without foundation. Then he acknowledges that "Burt's description of his procedure is extremely vague" and that "one can easily see how many readers could be misled into believing the counts were frequencies," but he rejects the possibility that Burt misled them. By the end of the letter he contradicts his original position that Burt told the reader what he did, and no longer endorses Burt's "unclear explanation of what he did" and "his refusal to present the raw data."

Stigler seems to ascribe the vagueness and omissions primarily to "the standards of Burt's time." But in 1961, the year the paper was published, Burt was simultaneously assistant editor of the *British Journal of Psychology*, assistant editor of the *British Journal of Statistical Psychology*, and aide to the editor of the *British Journal of Educational Psychology*. Hence he must surely have known the requirement for precise descriptions of procedure in scientific communications. Moreover, the extreme vagueness of Burt's descriptions of procedure in his 1961 article is quite inconsistent with his reputation for precision. According to John Cohen, one of his eminent former students (*1*, p. 86), "'sloppy' is the very last word one would use for the precise and punctilious Burt." Cohen goes on to describe him as "an impeccable and meticulous investigator . . . who spared himself no pains to check every figure, every statement and every source" (*1*, p. 87). According to Eysenck (*2*, p. iv): "As the first editor of this *Journal* [*British Journal of Statistical Psychology*], which in a very real sense is his own creation, he set very high standards of critical appraisal; contributors received many pages of detailed discussion and criticism of their work from him." Burt's 1961 highly acclaimed 22-page paper was published in that journal.

Stigler apparently attributes Burt's statistical procedures to R. A. Fisher, "the only statistical reference in Burt's 1961 paper." Fisher was quite clear about manipulation of data. In his classic *Design of Experiments*, he discusses the issue in a section entitled "Manipulation of the data." He states (*3*, p. 45): "If the results of an experiment, as obtained, are in fact irregular, this evidently detracts from their value; and the statistician is not elucidating but falsifying the facts, who rearranges them so as to give an artificial appearance of regularity." Burt clearly did not learn those methods from R. A. Fisher. Finally, Stigler does not endorse Burt's "refusal" to present his observed frequencies, or as Stigler calls them, "the raw data," but gives no plausible explanation for the bizarre "refusal." They could have easily been put in parentheses next to the corresponding computed "data."

To sum up: Stigler has not demonstrated that Burt told the reader what he had done, and he has not provided a plausible explanation for Burt's extreme vagueness and the omission of his observed frequencies. My explanation is simple and straightforward: Burt did not present his observed frequencies because they did not exist, and he was "extremely vague" in his descriptions of statistical procedure in order to mask a scientific fraud.

I will now answer Stigler's specific arguments and criticisms. First, he states that I have significantly misrepresented Burt's row totals, in spite of the fact that my interpretation agrees with that of the experts in the field (*4–11*). Indeed, I invite Stigler to find a single article or book in which those row totals are said to have come "from Spielman and Burt or some other publication of these data." Moreover, I invite Stigler to find a single statement in the 22-page article where the "precise and punctilious" Burt said that he took those row totals "from Spielman and Burt or some other publication of these data." The normal interpretation of the passage that Stigler cites is that Burt weighted his actual frequencies by $1000/N$, where N is the size of the sample. It follows from this customary interpretation that the cell entries are "proportional frequencies" and that the class proportions are "the estimated proportions for the total population" (*12*, p. 10). Furthermore, Stigler's interpretation contradicts Burt's introduction to his tables I and II. Burt said (*12*, p. 9): "In studying the distribution of intelligence among the different occupational classes it is in my view desirable to examine, not only (as is usually done) the class-means, but the entire frequency distributions. Accordingly, in Tables I and II, I give *frequencies* [my italics] both for adults and for children." Thus Burt did not call those numbers "weighted" counts. He called them "frequencies," and he called the distributions of numbers "frequency distributions." In fact, Burt never called those computed numbers "weighted" or adjusted counts anywhere in his paper. Even in the only passage Stigler selected to support his interpretation, Burt referred to those numbers as "proportional frequencies" (*12*, p. 10), not "weighted" or adjusted counts.

Moreover, contrary to Stigler's statement, Burt's use of Spielman and Burt's percentages as the row totals of his tables is quite bizarre. Spielman and Burt characterized their percentages as "nothing more than the roughest approximation" (*13*, p. 15). They were "based mainly upon Charles Booth's survey" (*14*, p. 349) of London in the late 19th century (*15*) combined in an unstated way with some unspecified pre-1926 London census figures for employed male adults aged 14 plus, with or without children and married or unmarried. Burt's purported sample for the 1961 paper was described as a sample of father-son pairs having an average age "difference of 28.4 years" (*12*, p. 16) from an anonymous "London borough selected as typical of the whole county" (*12*, p. 9) for "nearly fifty years, namely, from 1913 onwards" (*12*, p. 4). Thus the Spielman and Burt percentages have a nonsensical relation to Burt's sample and population of interest. Yet Stigler sees nothing bizarre in Burt's using row totals whose source is never given, whose quality is never described, and whose relevance to the problem is never discussed.

Stigler next tries to justify the essentially perfect fit of Burt's column totals to a normal curve. In introducing his tables I and II, Burt says only that "we have rescaled our assessments of intelligence so that the mean of the whole group is 100 and the standard deviation 15" (*12*, p. 10). Stigler argues that Burt "hinted" that he had rescaled to a normal curve with mean 100 and standard deviation 15, but doesn't explain why Burt didn't simply tell the reader that that was what he had done. To support his conjecture, Stigler says that the normal distribution was then "used much more freely as a basis for test scaling." His statement is incorrect for IQ test scores. Indeed, Burt would have strongly disagreed with Stigler's unsupported assertion. In 1957 Burt wrote a paper specifically devoted to the distribution of intelligence, entitled "The distribution of intelligence" (*16*). In that article Burt states (*16*, p. 174), "Except for those intended solely for a single age-group, most [intelligence] tests are scaled by methods independent of any presupposition about the resulting distribution: e.g. (i) by mental age, (ii) by least perceptible or equal-appearing intervals, or (iii) by the number of unit processes performed." Since the scaling did *not* determine the distribution, Burt then argued that the observed frequency distributions of IQ test scores agreed with his predictions from the genetic theory. He concluded his discussion with the following: "With each of these procedures [mental age, equal-appearing intervals, number of unit processes performed] the frequency curves obtained yield, in nearly every case, statistical constants implying a definite though slight degree of leptokurtosis and negative asymmetry. Thus, the results are fully consistent with the twofold hypothesis suggested by genetic arguments and entirely incompatible with the other three hypotheses" (*16*, p. 174). As further evidence that Burt did not view scaling IQ's to a normal curve as a common device, Burt actually raised suspicions about the frequency distributions published by E. L. Thorndike, an eminent leader in educational measurement. Burt expressed his suspicions in 1963 in the *British Journal of Statistical Psychology*, 2 years after he published his almost perfect normal curves in that journal. Burt declared (*17*, p. 176), "On applying the chi-squared test to his [Thorndike's] own measurements he reaches values for *P* ranging from 0.99 to 0.999,999. Now we used to be warned that 'a value of *P* very near to unity should lead the investigator to suspect his hypothesis quite as much as very small values: such very close correspondences are too good to be true' " (*18*). If Burt had considered scaling to a normal curve to be a common device for IQ scores, he surely would not have questioned the close correspondence of Thorndike's data to the normal curve.

Furthermore, Stigler limits his discussion to the question "what did Burt mean by 'rescaled'?" but neglects to mention that Burt did not use the term "rescale" in his introduction to tables III and IV. In introducing table III, Burt said (*12*, p. 12): "If we now *reclassify* [my italics] the actual data for adults according to these new borderlines, we obtain the distribution set out in Table III."

In introducing table IV, Burt said (*12*, p. 15): "Table IV shows the distribution of the children with the scale for intelligence *subdivided* afresh [my italics]." The terms "reclassify" and "subdivide afresh" do not mean "rescale to a normal curve," whereas they are completely consistent with my representation of those tables.

There is another problem with Stigler's explanation of the almost perfect fit of the column totals to the normal curve. He proposes that Burt constructed his tables in one of two ways. He either (i) transformed "his pooled (over [occupational] classes) IQ's individually" to a normal curve and then weighted by the Spielman and Burt proportions or (ii) weighted "his columns (as he did his rows) to fit 'estimated proportions.'" Rubin discusses these two possible explanations, and gives excellent arguments against the second one, which is the one that Stigler prefers. Rubin points out that "the hypothesis that Burt fixed both row and column margins at population proportions is inconsistent with [Rubin's properties] (i), (ii), (iii)." It is also quite inconsistent with the fact that the mean IQ's for the occupational classes given in the last column of Burt's tables I and II disagree in eight of nine possible comparisons with the arithmetic means computed from those tables using the midpoints of the intervals. If alternative two is true, they should be identical in every case. Rubin also points out that alternative two is inconsistent with Burt's "description of the tables." Rubin cites the pages where Burt used "reclassify" and "subdivide afresh," which do not mean "reweight along rows and columns." On the basis of Rubin's excellent arguments, we may conclude that Stigler's alternative two would be a significant misrepresentation of Burt, and Stigler would surely not want to misrepresent Burt. Hence, we only need to evaluate Stigler's alternative one, which is also Rubin's proposal.

I will now show that alternative one won't work. Let us hypothetically construct Burt's four tables using alternative one. We begin with a sample of individual IQ's of fathers and their sons, and then transform each set of scores separately to fit a normal curve (*19*). After this transformation, we must construct a table of sample frequencies for each of Burt's four tables of published numbers. To weight along rows by the Spielman and Burt proportions, we first need to convert the hypothetical tables of sample frequencies to tables of relative frequency distributions by dividing each cell frequency f_{ij} (i^{th} row, j^{th} column) by $N_{i\cdot}$, the sample size for the i^{th} row. To fix the row proportions to the Spielman and Burt proportions ($P^*_{i\cdot}$, $1 \le i \le 6$), we now multiply each $f_{ij}/N_{i\cdot}$ by $P^*_{i\cdot}$. The row totals after this weighting are

$$\sum_{j=1}^{n} (f_{ij}/N_{i\cdot})\, P^*_{i\cdot} = P^*_{i\cdot} \qquad 1 \le i \le 6$$

where n is the number of columns. Burt's published numbers are in per mille, so that $N^*_{i\cdot} = 1000\, P^*_{i\cdot}$, where the $N^*_{i\cdot}$ $(1 \le i \le 6)$ are Burt's published row totals for the table of interest. Alternative one assumes that column totals are not changed by the weighting along rows, so that

$$\sum_{i=1}^{6} \frac{f_{ij}}{N_{i\cdot}} P^*_{i\cdot} = \sum_{i=1}^{6} \frac{f_{ij}}{N_i} \frac{N_{i\cdot}}{N} \qquad 1 \le j \le n \tag{1}$$

or equivalently that

$$\sum_{i=1}^{6} \frac{f_{ij}}{N_{i\cdot}} \left(P^*_{i\cdot} - \frac{N_{i\cdot}}{N} \right) = 0 \qquad i \le j \le n \tag{2}$$

But the matrix $\{(f_{ij}/N_{i\cdot})N^*_{i\cdot},\ 1 \le i \le 6,\ 1 \le j \le n\}$ is Burt's published table by hypothesis. Moreover, each of Burt's published tables can be shown to have rank equal to the number of rows, so that the matrix $\{f_{ij}/N_{i\cdot},\ 1 \le i \le 6,\ 1 \le j \le n\}$ has rank equal to the number of rows. Hence the system of Eqs. 2 is only satisfied by the trivial solution, or equivalently

$$P^*_{i\cdot} = \frac{N_{i\cdot}}{N} \qquad 1 \le i \le 6$$

Thus, weighting the rows by the Spielman and Burt proportions would preserve the column totals if and only if Burt's actual sample proportions in each occupational class equal the nonsensical Spielman and Burt proportions. In brief, if Burt had rescaled the individual IQ's to fit a particular normal curve, he would have lost that normal curve by weighting the rows with proportions different from his actual sample proportions. One can also show that accidental fit to the normal curve would be highly unlikely by trying random vectors of weights in the neighborhood of $P^*_{i\cdot}$ $(1 \le i \le 6)$. In conclusion, Stigler's alternatives do not explain Burt's column totals.

Stigler next expresses concern about the sample size. He apparently believes that he is better able to interpret Burt's descriptions than Dobzhansky, who said "some 40,000" (*5*, p. 19); than Eysenck, a former student and colleague of Burt's, who said "some 40,000" (*7*, p. 62); than Gottesman, who said "some 40,000" (*8*, p. 36); than Herrnstein, who said "40,000" (*9*, p. 436); and Willerman, who said "40,000" (*11*, p. 11). According to Stigler, these authorities "must be" in error, since his second alternative—which Rubin and I have shown is inconsistent with Burt's descriptions and tables—implies otherwise. Stigler evidently sees nothing suspicious in Burt's failure to give the size of his purported sample. My own position is that Burt did not explicitly report the sample size because there was no sample.

Stigler and several others who wrote letters to *Science* are distressed with my calculation of $\bar{X}_{ic}/(\bar{X}_{if} + 100)$ for the occupational classes. They think that I incorrectly derived that formula from a statistical regression equation. In fact, I derived that formula directly from an equation that I believe is part of Burt's fabrication device. I was suggesting that Burt's fabrication gave the following simple relation between the means of the children and the means of the fathers:

$$\bar{X}_{ic} = \frac{1}{2}(\bar{X}_{if} + 100) \tag{3}$$

where $\bar{X}_{ic}$ and $\bar{X}_{if}$ are Burt's published means of the children and the fathers respectively. I inferred this equation "from Conway's discussion" (*20*, p. 1179) in her (Burt's ?) article on social class published in 1959 (*21*). The equation predicts a regression of the children's means toward 100, and therefore I called the coefficient of 1/2 a "regression coefficient." Perhaps I should have called it a "fabrication coefficient." Rearrangement of Eq. 3 gives

$$\frac{\bar{X}_{ic}}{\bar{X}_{if} + 100} = \frac{1}{2}$$

which was verified for every class. It should be emphasized that I never called Eq. 3 a "statistical regression equation" anywhere in the *Science* article and I used the symbol α so that my proposed fabrication constant would not be confused with "β," the standard statistical regression coefficient.

If we subtract 100 from both sides of Eq. 3, we obtain

$$\bar{X}_{ic} - 100 = \frac{1}{2}(\bar{X}_{if} - 100) \tag{4}$$

Rearrangement of Eq. 4 gives

$$\frac{\bar{X}_{ic} - 100}{\bar{X}_{if} - 100} = \frac{1}{2}$$

which is what Stigler tested. Equation 4 is fine as a statistical regression equation, but it would be a nonsensical proposal for a fabrication equation, since

Table 1. The predicted and the published data for Burt's tables I and II (*12*). The predicted percentages and means are in parentheses. They were computed from the theoretical bivariate normal distribution with means 100 and standard deviations 15. The correlation (σ) was fixed at 0.80 for the fathers and 0.40 for the children.

Class	IQ intervals						Total	Mean IQ
	50–70	70–90	90–100	100–110	110–130	130+		
				Fathers				
I	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	140 (137)
II	0 (0)	0 (0)	0 (0)	0 (0)	2 (2)	2 (1)	4 (3)	131 (126)
III	0 (0)	0 (0)	1 (1)	2 (3)	9 (8)	0 (1)	12 (13)	116 (116)
IV	0 (0)	1 (1)	5 (5)	10 (10)	9 (10)	0 (0)	25 (26)	108 (107)
V	1 (0)	5 (7)	14 (13)	12 (10)	2 (3)	0 (0)	34 (33)	98 (98)
VI	2 (2)	17 (15)	5 (7)	1 (2)	1 (0)	0 (0)	26 (26)	85 (85)
Total	3 (2)	23 (23)	25 (26)	25 (25)	23 (23)	2 (2)	101 (101)	
				Children				
I	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	121 (118)
II	0 (0)	0 (0)	0 (0)	1 (1)	2 (2)	0 (0)	3 (3)	115 (113)
III	0 (0)	1 (1)	2 (2)	3 (3)	5 (5)	1 (1)	12 (12)	108 (108)
IV	0 (0)	5 (4)	5 (6)	7 (7)	8 (8)	1 (1)	26 (26)	105 (103)
V	1 (1)	8 (8)	10 (9)	9 (8)	5 (6)	1 (0)	34 (32)	99 (99)
VI	2 (1)	9 (10)	8 (7)	5 (5)	2 (3)	0 (0)	26 (26)	93 (93)
Total	3 (2)	23 (23)	25 (24)	25 (24)	22 (24)	3 (2)	101 (99)	

Burt published means, not deviations of means from 100.

Let us now consider the statistical regression equation

$$\bar{X}_{ic} = \alpha + \beta\bar{X}_{if} \qquad 1 \leq i \leq 6$$

and estimate α and β, which Stigler failed to do. According to my Eq. 3, α should equal 50 and β should equal 0.50. The unweighted least-squares estimates of α and β are in fact 50 and 0.50 respectively, each correct to two significant figures. Figure 1 shows that I have not exaggerated the fit of my proposed fabrication equation to Burt's means. Indeed, this "nonsensical" Eq. 3 predicts the mean IQ of the children with an average absolute error of less than 4/10 of an IQ point.

Also, the product-moment correlation between the $\bar{X}_{if}$ and $\bar{X}_{ic}$ is 0.999, correct to three significant figures, and gives perfect support for the linear relation between those means predicted by this equation. Moreover, there is no a priori reason why that relation should be linear. In fact, Eq. 3 may very well be the best-fitting linear function in the history of psychometric measurement—and the parameters were decided a priori, not estimated from the data. I invite Stigler to find a plausible explanation for $\hat{\alpha} = 50$, $\beta = 0.50$, and the linear relation between $\bar{X}_{if}$ and $\bar{X}_{ic}$ with a correlation of 0.999. Don't forget to assume that Burt's study was "merely a pilot inquiry" (*12*, p. 9), that the "data are too crude and limited" (*12*, p. 9), and that "the assessments of adult intelligence were less thorough and less reliable" (*12*, p. 9) than those of the children. It should also be pointed out that Eq. 3 is genetically bizarre: it follows from the assumption of perfect Mendelian inheritance of IQ (with no dominance or epistasis) and random mating across classes. The genetic theory gives

$$\mu_{ic} = \frac{1}{2}(\mu_{if} + \mu_{im})$$

where μ denotes population mean and μ_{im} is the mean of the mothers for the i^{th} class. Under random mating, $\mu_{im} = 100$, so that

$$\mu_{ic} = \frac{1}{2}(\mu_{if} + 100)$$

which is my Eq. 3 with population means in place of sample means.

Stigler's final point concerns the feasibility of actually fabricating tables of data with predetermined marginal totals. He thinks that "it is not a simple matter" to do that. On the contrary, I will now show that it is quite easy to do. Assume that Burt manufactured each of his tables from a bivariate normal distribution which means 100 and standard deviations 15, with a correlation of 0.80 for each table of fathers' data and 0.40 for each table of children's data. Furthermore, assume that Burt attempted to conceal the fabrication by moving digits in the unit's place from some cells to other cells after conversion from probabilities to per mille. One would then expect that Burt's tables "cannot be perfectly fitted by a bivariate normal distribution." To expect Burt's fraud to be transparent is naive.

Table 1 presents the data of Burt's tables I and II reduced to six intervals of IQ scores and the cell entries rounded to whole percentages. If Burt attempted to conceal the fraud by manipulating the unit's place of his "numbers per thousand," then broader blocking and rounding to whole percentages should help reveal the underlying pattern. The numbers in parentheses are whole percentages from the theoretical bivariate normal distributions. Notice that the maximum deviation between the predicted percentages and Burt's percentages is two percentage points for the fathers and one percentage point for the children. Moreover, the first three rows of percentages for the children—18 numbers—are predicted exactly. The results for Burt's tables III and IV are presented in Table 2. For these tables the maximum deviation between the percentages predicted from my fabrication model and Burt's percentages is also two percentage points for the fathers and one percentage point for the children.

This fabrication model can also be used to predict Burt's published means. Let $p_{i\cdot}$ be Burt's proportion in the i^{th} row (i^{th} class),

$$P_k = \sum_{i=1}^{k} p_{i\cdot}$$

and let Z_k be the inverse value of the standardized normal distribution function at P_k. It follows from the fabrication model that

$$\mu_{if} = 100 + 0.8\,\frac{f_i - f_{i-1}}{p_{i\cdot}}\,15 \qquad (5)$$

and that

$$\mu_{ic} = 100 + 0.4\,\frac{f_i - f_{i-1}}{p_{i\cdot}}\,15 \qquad (6)$$

where f_i is the value of the standardized normal probability density at Z_i, and $f_0 = f_6 = 0$.

Moreover, it follows from Eq. 5 that

$$\frac{f_i - f_{i-1}}{p_{i\cdot}} 15 = \frac{\mu_{if} - 100}{0.8}$$

Hence,

$$\mu_{ic} = 100 + \frac{.4}{.8}(\mu_{if} - 100) = \frac{1}{2}(\mu_{if} + 100) \qquad (7)$$

which is the nonsensical Eq. 3 with population means substituted for sample means. The predicted and published means rounded to whole numbers are given in Table 1. The predicted means were computed from Eqs. 5 and 6. Notice that Burt's means are predicted perfectly for his classes III, V, and VI. The published frequencies for these classes are over 70 percent of the grand total. Furthermore, Eq. 7 predicts the nonsensical relation between the mean of the children and the mean of the fathers seen in Fig. 1. In brief, the fit of the bivariate normal distributions with simple constants of .80 and .40 is excellent.

Perhaps I have discovered the Mendelian laws for the inheritance of social class. Unfortunately, since the model is genetically nonsensical and since Burt's study was "merely a pilot inquiry" (*12*, p. 9), the data "too crude and limited" (*12*, p. 9), and the father's data even "less thorough and less reliable" (*12*, p. 9), that explanation for the excellent fit of the model can be rejected beyond reasonable doubt. Since there are no other plausible explanations, I would conclude that Burt fabricated his tables from bivariate normal distributions with $\rho = .80$ for the fathers and $\rho = .40$ for the children, and contrary to Stigler it was a "simple matter" to do.

I have one comment on Rubin's letter. Rubin failed to notice that the column totals are determined by the row totals through a simple matrix equation. Thus, if Burt had transformed his IQ scores to fit a normal distribution, he would have lost the normal distribution by changing the row totals (fixing the class margins).

DONALD D. DORFMAN

Department of Psychology,
University of Iowa, Iowa City 52242

Table 2. The predicted and the published data for Burt's tables III and IV (*12*). As in Table 1, the predicted percentages are in parentheses and were computed from the bivariate normal distribution with means 100 and standard deviations 15. As in Table 1, the correlation was fixed at 0.80 for the fathers and 0.40 for the children.

Class	IQ intervals				Total
	VI 50-91	V-IV 91-115	III-II 115-141	I 141+	
			Fathers		
I	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
II-III	0 (0)	5 (6)	10 (9)	0 (0)	15 (15)
IV-V	7 (8)	46 (44)	5 (6)	0 (0)	58 (58)
VI	19 (18)	7 (8)	0 (0)	0 (0)	26 (26)
Total	26 (26)	58 (58)	15 (15)	0 (0)	99 (99)
			Children		
I	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
II-III	1 (1)	9 (9)	5 (5)	0 (0)	15 (15)
IV-V	14 (13)	35 (36)	10 (9)	0 (0)	59 (58)
VI	11 (12)	14 (13)	1 (1)	0 (0)	26 (26)
Total	26 (26)	58 (58)	16 (15)	0 (0)	100 (99)

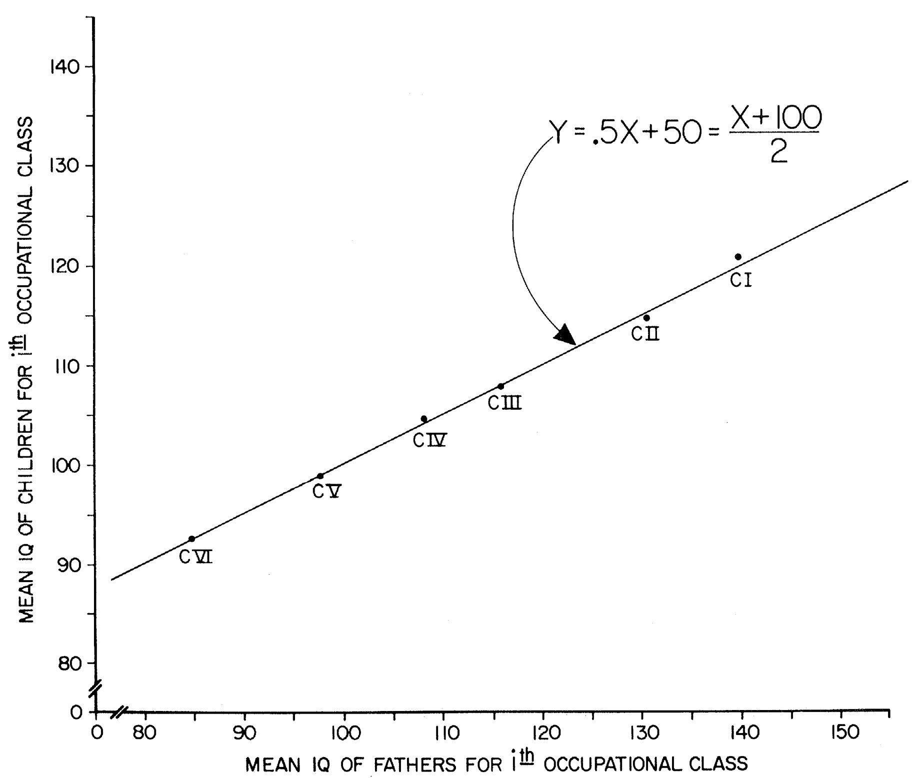

Fig. 1. Burt's published mean IQ of the children plotted against the published mean IQ of the fathers for each occupational class. The proposed fabrication equation is also plotted for purposes of comparison.

References and Notes

1. J. Cohen, *Encounter* **48**, 86 (March 1977).
2. H. J. Eysenck, *Br. J. Stat. Psychol.* **25**, i (1972).
3. R. A. Fisher, *The Design of Experiments* (Oliver & Boyd, London, 1937).
4. L. L. Cavalli-Sforza and W. F. Bodmer, *The Genetics of Human Populations* (Freeman, San Francisco, 1971), p. 795.
5. Th. Dobzhansky, *Genetic Diversity and Human Equality* (Basic Books, New York, 1973).
6. L. Ehrman and A. Parsons, *The Genetics of Behavior* (Sinauer, Sunderland, Mass., 1976), p. 293.
7. H. J. Eysenck, *The I.Q. Argument: Race, Intelligence, and Education* (Library Press, New York, 1971), p. 62.
8. I. I. Gottesman, in *Social Class, Race, and Psychological Development*, M. Deutsch, I. Katz, A. R. Jensen, Eds. (Holt, Rinehart & Winston, New York, 1968), pp. 36-37.
9. R. J. Herrnstein, *Contemp. Psychol.* **20**, 436 (1975).
10. J. Hirsch, *ibid.*, p. 436.
11. L. Willerman, *Individual and Group Differences* (Harper & Row, New York, 1975).
12. C. Burt, *Br. J. Stat. Psychol.* **14**, 3 (1961).
13. W. Spielman and C. Burt, in *A Study in Vocational Guidance, Report No. 33*. F. Gaw, L. Ramsey, M. Smith, W. Spielman (under the general direction of C. Burt) (His Majesty's Stationery Office, London, 1926), pp. 12-17.
14. C. Burt, *Br. J. Psychol.* **14**, 336 (1924).
15. C. Booth, *Life and Labour of the People in London* (Macmillan, London, 1892-1903), 17 volumes. Charles Booth was a Victorian man of business—a Liverpool shipowner and manufacturer who was also an amateur sociologist. It is extraordinary that Burt primarily used Booth's informal survey.
16. C. Burt, *Br. J. Psychol.* **48**, 161 (1957).
17. ______, *Br. J. Stat. Psychol.* **16**, 175 (1963).
18. The reference Burt cites for the quotation is G. U. Yule and M. G. Kendall, *An Introduction to the Theory of Statistics* (Griffin, London, 1937).
19. The success of the transformation depends on the number of distinct obtained scores. Successful transformation to normality is by no means certain.
20. D. D. Dorfman, *Science* **201**, 1177 (1978).
21. J. Conway, *Br. J. Stat. Psychol.* **12**, 5 (1959). The Conway of 1959 appears to have been invented by Burt.