Biol. Rev. (1986), **61**, *pp.* 295–312
Printed in Great Britain

ARE MENDEL'S RESULTS REALLY TOO CLOSE?

BY A. W. F. EDWARDS
Gonville and Caius College, Cambridge

(*Received* 26 *March* 1985, *accepted* 29 *May* 1986)

CONTENTS

I. Introduction 295
II. A survey of 'explanations' 296
III. Previous statistical analyses 299
IV. Goodness-of-fit χ^2 302
V. A fresh analysis 304
VI. Conclusions 309
VII. Summary. 310
VIII. References 311

I. INTRODUCTION

Gregor Mendel read his famous paper 'Experiments in plant hybridization' to the Brünn Natural History Society in 1865 (Mendel, 1866). It attracted wide attention only from 1900 onwards, and in 1911 R. A. Fisher, then starting his third year as an undergraduate in Gonville and Caius College, Cambridge, noted in a talk to the Cambridge University Eugenics Society:

> It is interesting that Mendel's original results all fall within the limits of probable error; if his experiments were repeated the odds against getting such good results is about 16 to one. It may have been just luck; or it may be that the worthy German abbot, in his ignorance of probable error, unconsciously placed doubtful plants on the side which favoured his hypothesis (Norton & Pearson, 1976; Bennett, 1983).

Twenty-five years later Fisher published his analysis of Mendel's data (Fisher, 1936), admitting, in a covering letter to the editor of the journal, 'I had not expected to find the strong evidence which has appeared that the data had been cooked' (Bennett, 1965). Nearly 20 years later still he prepared an introduction and commentary to Mendel's paper which was published posthumously in 1965 (Bennett, 1965), and his final verdict was, 'The data have evidently been sophisticated systematically, and after examining various possibilities, I have no doubt that Mendel was deceived by a gardening assistant, who knew too well what his principal expected from each trial made.' A letter from Fisher to E. B. Ford dated 2 January 1936 (Bennett, 1983) amplifies what Fisher called his 'abominable discovery': 'I cannot conceive that Mendel himself had any hand in it'. From his comments in 1936 it seems that Fisher had by then forgotten what he had written in 1911.

Fisher's remark about the odds of 16 to 1 was probably taken directly from Weldon (1902), but that is not to suggest that Weldon himself harboured any suspicions about Mendel's results. He subjected them to a statistical analysis using probable errors but

concluded only that Mendel's statements were 'admirably in accord with his experiment'. Fisher had encountered Mendel's paper itself in the English translation in Bateson's newly published *Mendel's Principles of Heredity* (Bateson, 1909) which he had bought as a freshman (Fisher, 1952). In 1936 he referred to a footnote of Bateson's as raising 'a point of great interest' because it hinted that Mendel's paper should not be taken as a literal account of his experiments, and this may be said to have been the first such suggestion.

Fisher's 1936 conclusion slowly became the received wisdom, but his painstaking analysis and his defence of Mendel's integrity have sometimes been incorrectly reported as having exposed a scientific fraud of major proportions, and the name of Mendel is in danger of acquiring the connotations of Piltdown or Burt. Recently there has been a spate of papers seeking to defend Mendel against the imaginary charges. Thus Pilgrim (1984) writes 'The purpose of this [i.e. his own] paper is to demonstrate that Fisher's reasoning was faulty and to clear the name of an honest man', and 'There is no evidence that Mendel did anything but report his data with impeccable fidelity. It is to the discredit of science that it did not recognize him during his lifetime. It is a disgrace to slander him now'. An answer to Pilgrim is Edwards (1986). A similar example of missionary zeal is afforded by Monaghan & Corcos (1985), who write that 'suggestions that Mendel's data were biased in support of an hypothesis about inheritance in peas are not supported'. Their paper is, however, completely untrustworthy in its statistical analysis (see Section III below). A recent biography of Mendel (Orel, 1984) concludes, on the basis of Weiling's analysis to which I refer below, 'This would seem to have removed the last shadow of doubt as to the credibility of Mendel's experiments and his published results'.

As I will show, such views are not supported by the evidence. In the present paper, after reviewing the many 'explanations' of Mendel's data, I shall let the facts speak for themselves as plainly as possible: there is no point in addressing the issue with statistical techniques of such complexity that they are more difficult to interpret than the original data. I shall assume that readers are familiar with Mendel's paper and that they have an English translation to hand, preferably that printed in Bennett (1965), to which my page references will apply (thus: M, p. 20). I cannot myself comment on the quality of this translation from the German, but it certainly reproduces Mendel's numerical material without error. Others have felt the need for an improved translation (see Stern in the foreword to Stern & Sherwood, 1966) and one has been made (Sherwood, *loc. cit.*).

II. A SURVEY OF 'EXPLANATIONS'

The critical re-evaluation of experiments and observations is an important part of scientific endeavour. ('There is no substitute for a careful, or even meticulous, examination of all original papers purporting to establish new facts', Fisher, 1936.) Though in certain notorious cases this has led to the exposure of fraudulent work, it generally results in an increased appreciation of the original material or even a fresh discovery from it. Occasionally some puzzling aspects of the data surface on a close examination, and may even lead to the suggestion that the experiments did not take place exactly as the author reported them. Thus Galileo's unrepeatable results in his experiments on acceleration down an inclined plane mentioned in the *Dialogo* led

Mersenne (1637) to doubt whether he had actually performed them. In that example it was the incomprehensibly bad data which raised doubts, but long before the era of modern statistics it was understood that suprisingly good data could equally well create justifiable suspicion. Babbage (1830), writing before the time of Mendel, classified some of the methods used to produce data too close to expectation:

> *Trimming* consists in clipping off little bits here and there from those observations which differ most in excess from the mean, and in sticking them on to those which are too small.
>
> *Cooking*... One of its numerous processes is to make multitudes of observations, and out of these to select those only which agree, or very nearly agree.

Many commentators on Mendel's results share Fisher's belief that some kind of 'cooking' has taken place (for example: de Beer, 1964; Sturtevant, 1965; Crew, 1966), but most have sought alternative explanations for the surprisingly good fit of the data to the hypothesis.

A popular one is that the data are the result of an imperfectly understood sampling process. An early example of such a misunderstanding occurs in the work of Francis Galton, who, by the standards of his time, had a remarkably clear head for statistical enquiry. In *Hereditary Genius* (Galton, 1892), however, he concluded that since judges were all men and came from families of average size five, each must have (on average) $2\frac{1}{2}$ sisters and $1\frac{1}{2}$ brothers. Galton is here wrongly assuming binomial sampling, with probability one half, *including* the judge; might not we be assuming the wrong sampling process for Mendel's data? (Galton, 1904, later realized his error.)

The natural suggestion is that Mendel stopped counting the peas in each experiment when he was satisfied with the result, and this has been put forward numerous times (Dunn, 1965*a*, *b*; Olby, 1966; Beadle, 1967; van der Waerden, 1968; C. D. Darlington, reported to me by J. H. Edwards; J. Kříženecký, reported by Orel, 1968). The difficulty is that there is no evidence whatsoever that Mendel did this, and in many of the experiments his description clearly excludes the possibility, so that to accept it as a means of exonerating Mendel from having reported data which had been adjusted in some way is to saddle him with the charge of not having reported his experimental method accurately. 'The style throughout suggests that he expects to be taken entirely literally; if his facts have suffered much manipulation the style of his report must be judged disingenuous' (Fisher, 1936). Van der Waerden (1968) is the only one of the authors quoted above who pursued the suggestion in detail, and although he found it possible to account for some of the earlier instances of surprisingly good fits in the experiments if he assumed sequential sampling, he concluded 'All in all my impression is that some data from later years were probably biased.'

Another favourite suggestion has been that the data exhibit non-binomial variability not because of the sampling process, but because of the reproductive physiology of the pea (Sturtevant, 1965; Weiling, 1966; Thoday, 1966; Beadle, 1967). Sturtevant and Beadle 'explored this possibility to see if it was sufficient to account for the apparent bias. It works in the right direction but is not sufficient' (Beadle, 1967). Weiling (1966, 1971, 1985) made a detailed comparison of Mendel's data with data from other authors and concluded that the results were all consistent with the hypothesis that the variance

of the segregation was smaller than the binomial variance by a factor of between 0·6 and 1·0. In the next section I give my reasons for rejecting this suggestion: the evidence is that the pea is in fact an excellent randomizer!

To help explain the infra-binomial variance in the experiments in which Mendel progeny-tested F_2 plants which displayed one of the dominant plant-characteristics to see if they were homozygous or heterozygous (M, pp. 20, 26), Weiling (1966) suggested that on average only 8 of the 10 seeds sown to test each parent for heterozygosity will have produced a scorable plant, thus making the probability $(3/4)^8$ that no recessive was raised, rather than the $(3/4)^{10}$ assumed by Fisher (1936). Taking into account the resulting misclassification of some heterozygotes as homozygotes, the expectations are then even further from Mendel's data, and the fit consequently *worsened*. By contrast, Sturtevant (1965) thought that 'if the experiment included at least 10 seeds but often more than 10, then the correction to the 2:1 expectation will be less, and Fisher's most telling point will be weakened', and Wright (1966; also reported by Dunn, 1965*b*) advanced the same idea: 'if the average was 12, the probability of misclassifying [the heterozygote] falls from ·056 $[(3/4)^{10}]$ to ·031 $[(3/4)^{12}]$'. This would *improve* the fit of Mendel's results. As a matter of fact this suggestion had already been made by Fisher (1936) in connection with the 473 plants with coloured flowers from the trifactorial experiment which needed progeny-testing to determine whether they were homozygous or heterozygous, for which Mendel did not restate his procedure. 'It was presumably the same as before', wrote Fisher, adding:

> if we could suppose that larger progenies, say fifteen plants, were grown on this occasion, the greater part of the discrepancy would be removed. However, even using families of 10 plants the number required is more than Mendel had assigned to any previous experiment... Such explanations, moreover, could not explain the discrepancy observed in the first group of experiments, in which the procedure is specified,....

Nothing could illustrate better than these two opposing theories the ingenuity that has been expended on accounting for Mendel's surprisingly good data. Here are two explanations, one postulating that *fewer* than ten plants were scored, and the other that *more* than ten were scored, both with a view to accounting for results judged too close to a 2:1 ratio. In the next section I shall comment further on Fisher's handling of this question, and show that Weiling's suggestion is of little importance statistically, but once again we should note that there is no evidence that Mendel did either of the things suggested. He stated explicitly (M, p. 20) that 'ten seeds of each were cultivated', implying that there were exactly ten which grew. In Sherwood's translation (Stern & Sherwood, 1966) '*angebaut*' – 'cultivated' – is rendered as 'sown', but I am advised (K. Martin, personal communication) that this is an unwarranted change.

One puzzling aspect of the problem of Mendel having apparently overlooked the fact that progeny-testing with only ten plants will have led to some heterozygotes being misclassified as homozygotes is that towards the end of his paper (M, p. 49) he was completely clear that a form with low probability (one-eighth in the case he was treating) might by chance not appear in a particular experiment.

Wright (1966) also suggested that Mendel might often have been able to distinguish a 'segregating group from a nonsegregating one even in the absence of any recessives'.

In other words, for some characters Mendel might have been able to spot the heterozygote, and Wright mentioned seed-coat colour especially. This was indeed often true for seed-coat colour (M, p. 16), though Mendel never stated he used the fact. Indeed, he explicitly employed progeny-testing (M, p. 20) or 'further investigation' (M, p. 26), which all commentators have assumed to mean progeny-testing, in the case of seed-coat colour as with the other plant characters. Mendel (M, p. 15) also mentioned that F_1 hybrids from short × tall parents were usually taller than the tall parent, but he thought this possibly due to the 'greater luxuriance which appears in all parts of plants when stems of very different length are crossed' and did not mention its occurrence in the F_2. It is one thing to notice that tall offspring are generally taller than their tall parent, but another thing altogether to notice that a tall population of plants actually consists of two sub-populations. Even if there were some truth in the suggestion that in respect of seed-coat colour Mendel could distinguish between a mixed group of ten plants and a homozygous group, the explanation would only affect one character and the statistical consequences would be small. But the most telling objection to this hypothesis is, again, the fact that Mendel explicitly described a quite different procedure (M, p. 20).

In addition to the papers cited above, there have been many relevant reviews, though mostly of a somewhat inconclusive nature, and all lacking any extensive statistical analysis of the scope of Fisher (1936) or even Weiling (1966). Mention may, however, be made of the wide-ranging commentaries by Piegorsch (1983), who studied Weiling's analysis with care, and Root-Bernstein (1983), who concluded his review of earlier work by saying 'The issue remains unresolved,' and then, accepting the existence of bias, suggested that some plants were difficult to classify and on that account were tallied last in a direction to fit the ratios (cf. Fisher, 1911, quoted above in Section I). In respect of the others, I believe that I have read all the published material on this question, and that no useful purpose would be served by referencing papers which, though they may have fulfilled a valuable educational role at the time, are of no lasting value. Similarly, I make no reference to the many excellent commentaries on Mendel's paper that do not touch on the goodness-of-fit issue.

III. PREVIOUS STATISTICAL ANALYSES

Fisher (1936) gave the results of the most extensive statistical analysis of the data so far. On recomputing his χ^2 values (1936, table V) which he calculated to four places of decimals, I find them all correct, though Piegorsch (1983) noted that the famous 'P-value' for $\chi^2 = 41{\cdot}6056$ on 84 d.f. should be ·99997 and not ·99993!

There are, however, two minor criticisms of Fisher's analysis. First, the 20 d.f. *Illustrations of plant variation* are not independent of the 2 d.f. *3:1 ratios – seed characters* since the data for the former are included in the latter. Subtracting them out increases χ^2_2 from 0·2779 to 0·6232. Secondly, in those experiments in which, as Fisher was the first to point out, 2:1 ratios are not to be expected (see Section II above), he did in fact use them in his calculations of χ^2, presumably because they were what he thought Mendel expected. Root-Bernstein (1983) regarded this as a 'ploy' of Fisher's (a choice of word more revealing of Root-Bernstein's thought than Fisher's), but there are arguments both ways.

Fisher was testing the divergence between Mendel's counts and their expectations,

and the problem is *which* expectations, Mendel's or nature's? The difficulty is that any goodness-of-fit test confounds two sources of departure from expectation, the systematic (due to using the incorrect expectation) and the random (due to binomial variability). Usually one is assuming binomial variability and testing an expectation, in which case the null hypothesis under test clearly involves that expectation, but if one is assuming an expectation and testing the variability against the binomial model, as when the implied alternative is that the data have been adjusted towards the expectation, it is arguable that one should then assume *that* expectation. However, we need not explore this statistical by-way further (for which see Edwards, 1972, chapter 9) since even assuming the alternative expectations on Mendel's data, Fisher's overall conclusion is barely affected: Weiling (1966) computed that the χ^2_{84} is increased from 41·6056 to 48·910 and remarked correctly, 'This value is also highly significant.' Finally, on Fisher's statistical analysis, we should note that van der Waerden's (1968) suggestion of a numerical error seems to be a mistake.

After Fisher, Weiling (1966) has given the next most extensive statistical treatment. He squarely addressed the question of whether the variance might not be binomial by postulating that it is infra-binomial by a factor c, for the reasons given in Section II above, and estimating c. Exactly the same thought occurred to me at a later date (Edwards, 1972, chapter 9) in a general discussion of χ^2. My k^2 is his c, my figure 25 highly reminiscent of Weiling's figures, and my final comment indicative of an intention to do what, unbeknown to me, Weiling had already done:

> Since the alternative hypothesis [in considering the goodness-of-fit of Mendel's data] is that the variance of the results is less than expected, rather than that the means are in doubt, the use of χ^2 seems entirely appropriate. It would be interesting to rework Fisher's analysis using the justification of χ^2 offered by the Method of Support.

Of course, once one has estimated c (for which Weiling found the broad limits 0·6–1·0) there is nothing left to test, and Weiling persuaded himself that Mendel's results were compatible with those of other workers who repeated his experiments around the turn of the century. However, the facts indicate otherwise. The overwhelming majority of the post-Mendelian data on segregation in the pea comes from two sources, Bateson & Killby (1905) and Darbishire (1908, 1909), for which Weiling calculated $\chi^2_{408} = 411{\cdot}101$ and $\chi^2_{654} = 597{\cdot}689$ respectively, or $\chi^2_{1062} = 1008{\cdot}79$ in aggregate. The estimate of c is therefore $1008{\cdot}79/1062 = 0{\cdot}9500$. In other words these massive data exhibit a standard deviation of about 97·5 % of that expected on a binomial model, and far from this lending support to the biological hypothesis that the pea does not segregated randomly, it fills one with admiration for the perfection of the randomizing mechanism! The compatibility which impressed Weiling is but a reflection of the wide interval estimate he obtained for c on Mendel's data.

Moreover, the conclusion that the pea is a good randomiser is actually supported by an alternative analysis Weiling gave which he misinterpreted. If the data really do depart from the binomial model in the way he suggested, then not only should the observed χ^2 values be less than their expectations (= their degrees of freedom), but more than half of the component χ^2_1's should be less than their theoretical median value of 0·4549. He found (his table 4) that for Bateson & Killby's data 116 were less and 120 more,

and for Darbishire's 347 were less and 307 more (I assume that only 236 d.f. in Bateson & Killby's data refer to single-factor segregations). These figures, 463 *vs.* 427, lend no real support to the infra-binomial-variance hypothesis, but Weiling did not appreciate this because he wrote:

> In spite of the consequent results of our investigations into all the analyzed pea crosses (yielding an appreciable deviation of the value c from 1), the χ^2-test may, without hesitation, at least be applied to the evaluation of the individual segregations. Then the test of frequency distribution of all other inquiries into the underlying individual probabilities shows no noticeable deviation in the direction of P = 1·0, as [his] Table 4 illustrates.

Weiling here conceded 'no noticeable deviation' but thought this consistent with an appreciable overall departure of c from 1. He thus assumed that a collective of individual χ_1^2's can exhibit no systematic departure from their theoretical distribution even when their sum does, but this is a misunderstanding of the nature of χ^2. I consider the logic of χ^2 further in the next section.

Weiling also analysed the much less extensive data of Correns and Tschermak, and the fact that I have concentrated my remarks on the data of Bateson & Killby and Darbishire is explained by the circumstance that I had already chosen these data-sets for a complete re-analysis because of their extent and their ready availability to me. When I started the re-analysis I was not aware of the existence of the other smaller data-sets in the German literature, and since I do not read German I must leave their reconsideration to others. I only labour this point in order to deflect any suggestion that I have deliberately ignored data which supports this or that hypothesis. I hope to publish a full analysis of the very large English data sets in due course.

Weiling is to be congratulated for his determined attempt to rescue the Mendelian experiments from the Fisherian conclusion. He left no stone unturned in his search for explanations, increasing the value of χ^2 by using the true instead of the 2:1 ratio and by postulating that not all the 10 seeds sown in some of the experiments yielded mature plants, and then accounting for the too-small χ^2 still remaining by the hypothesis of a diminished variance. But in the end the attempt fails: there is too much to explain. The diminished-variance hypothesis is not supported by other more extensive data, and, as I show in my own analysis below, simply reducing the variance is in any case not enough to explain the peculiarities of Mendel's data.

No such congratulations can be offered to Monaghan & Corcos (1985). All their four tables contain substantial errors. At the trivial level, many of the χ^2 values they give are simply wrong to the accuracy claimed. Even the simple addition of these values is wrong in table I, and this departure alone from Fisher's value should have altered the authors to the fact that either they or Fisher were in error. Table III contains the astonishing assertion that $0{\cdot}95 < P < 0{\cdot}99$ for $\chi^2_{67} = 32{\cdot}57$, on which obvious error the authors hang their conclusions. Table IV *Results of other workers compared to those of Mendel* leads to the observation 'that the results of Bateson for cotyledon shape [*sic*] and Tschermak for cotyledon color deviate very widely from the other workers', which is not surprising given that Monaghan & Corcos have copied them both down wrongly from the literature. The list of references is a bibliographic disaster area. I spare the reader further criticism.

My overall impression from reading all the commentaries since Fisher (1936) is that a good deal of special pleading, not to mention downright advocacy, has failed to make any substantial impact on Fisher's conclusion. Moreover, as Fisher was the first to point out, much of the special pleading requires that Mendel's account itself would have to be judged 'disingenuous'. Certainly, there is every reason for each student of Mendel to keep an open mind and to make his own analysis and form his own conclusion. I give my own attempt in Section V, preceded in Section IV by a discussion of the precise nature of goodness-of-fit χ^2, the statistical procedure on which the whole issue hinges and which has often been misunderstood.

IV. GOODNESS-OF-FIT χ^2

Karl Pearson published his famous χ^2 goodness-of-fit test in the year that Mendel's paper was rediscovered (Pearson, 1900), and it is of interest that he warned against its use with large numbers of degrees of freedom:

> Thus, if we take a very great number of groups our test becomes illusory. We must confine our attention in calculating P to a finite [*sic*] number of groups, and this is undoubtedly what happens in actual statistics. n will rarely exceed 30, often not be greater than 12.

The basis of Pearson's concern seems to have been that in a large number of dimensions practically the whole of the probability in a multivariate normal distribution is to be found in the 'tails', or regions remote from the central region of high probability *density*. He seems to have sensed the dilemma facing anyone who ponders the question of what a 'typical' sample-point might be from a standardized multivariate normal distribution in, say, 84 dimensions (the number of degrees of freedom in Mendel's data). If, for example, the choice falls somewhere in the region of high probability density within one unit (= standard deviation in one dimension) of the centre, or mean, then the corresponding χ^2_{84} is less than one and, by this criterion, the point is surprisingly close to the centre. Yet if the choice falls about nine units from the centre, to make χ^2_{84} about 81 and thus near its expectation, it is in a region where the probability density has fallen to about $2{\cdot}6 \times 10^{-18}$ of its maximum value.

It is interesting to find this implicit concern for the logic of 'tail-area' significance tests so early in their history. More recent explicit criticisms are well-known, often in the writings of Bayesians (for example, Jeffreys, 1939) but not always (for example, Edwards, 1972). What is the 'rejection region' to be? Taking the 5 % significance level as an example, is it to be the region delimited by the line (or surface, etc.) of constant probability density which divides the sample space into probabilities of 0·95 and 0·05 such that the probability density everywhere in the latter part is less than the probability density everywhere in the former part? Or is it to be the line which delimits that part of the sample space beyond a certain distance from the mean, the distance chosen so that just 5 % of the probability is in it? In either case, what transformations of the sample space are allowed (since they manifestly affect both criteria)?

The χ^2 distribution is a transformation of the multivariate normal sample space, many dimensions being mapped into one by combining all points of equal probability density. On either of the above criteria the rejection region for the standardized

multivariate normal will be the outer skirts delimited by a hypersphere, but in the case of χ^2, which is not a symmetrical univariate distribution, it will be the two tails, consisting either of equal probability content 0·025 or cut off at equal probability *densities* and totalling 0·05 in probability. The corresponding region in the multivariate normal is not that described above, but rather two disjoint regions delimited by hyperspheres, one surrounding the centre (the image of the left-hand tail of χ^2) and one far removed from it (the image of the right-hand tail). A χ^2 value much below expectation (such as 41·6056 on 84 d.f.) is represented in the multivariate normal space by a point within the hyperspherical region surrounding the centre, the region of highest probability density, yet in the χ^2 space it is in a region of low probability density. The criticism of Mendel's results is that they are too close to expectation in the normal space as judged by being too far from expectation in the χ^2 space. *Quis custodiet ipsos custodes?*

We may emphasize this *Paradox of the Probable* in another way. Suppose the χ^2 test had antedated Mendel, and that in his paper he had reported a value of 84·0000 on 84 d.f. The reaction of a latter-day Fisher might well have been to conclude that Mendel's assistant had known that what Mendel really needed for his paper was not good mendelian ratios but a good value of χ^2.

I have discussed these questions at length in my book *Likelihood: ' The quantification of surprise in terms of probability is likely to tell only half the story '* (Edwards, 1972). The other half depends on the fact that the more improbable a result on a particular hypothesis as judged on *some* criterion, the greater the possibility of an alternative hypothesis leading to the result with a substantially higher probability. In other words there is a greater possibility of a hypothesis of higher likelihood, and hence one which, other things being equal, is more satisfactory. In particular, any *pattern* which we recognize in some data and which is unexplained on the current hypothesis is a signal that we should seek an alternative hypothesis, because an alternative which accounts for it is almost bound to have a higher likelihood. This may be the case even where the actual data are more probable than any other that might have occurred given the hypothesis, for we can then often more readily think of an alternative explanation than would be the case for some more arbitrary outcome. Thus numerical equality of the sexes at a children's party (see *Likelihood*, p. 190) is readily explained although, being the most probable outcome of a random selection of children, it might be thought to be the least in need of explanation. It is surely for this reason that attempts to base a theory of statistical estimation on maximising probabilities (Barndorff-Nielsen, 1976) have failed.

To sum up, we must not allow our judgment to be dominated by tests of significance and other calculations of probability which are at best pointers for further thought and at worst misleading. This is especially important when, as in the case of Mendel's experiments, the suggestion is that the data are ' too good ' in the sense of being *too probable*, too close to the target. If it were just a question of having hit the bull's eye with a single shot we might conclude, as some commentators have (e.g. Pilgrim, 1984), that Mendel was simply lucky, but when a whole succession of shots comes close to the bull's eye we are entitled to invoke skill or some other factor. A χ^2 of 41·6056 on 84 d.f. may look like a single lucky shot, but in reality it is a succession of 84 shots which must be manifesting some kind of pattern calling out for investigation. This will be the basis of my own analysis of Mendel's data in the next section.

Finally, there is one technical point about χ^2 which should be noted. It has been implied (Weiling, 1966) that one of the reasons for doubting the χ^2 analysis is that it is based on the assumption of normally distributed data, whereas genetic segregations are (at least on the simplest hypothesis) binomial. But it is a fact that the *expectation* of χ^2 on a binomial model is *exactly* equal to its degrees of freedom (Pearson, 1900). This is true *by definition*, for Pearson obtained his goodness-of-fit criterion by assuming that the normal distribution in question had a mean and variance given by the binomial distribution. In modern notation:

Let x be normally distributed with mean μ and variance σ^2. Then by definition $\chi_1^2 = (x-\mu)^2/\sigma^2$. Now put $\mu = np$ and $\sigma^2 = npq$, the mean and variance of a binomial distribution, and let the observed number of successes x be a and of failures be $b = n\text{-}a$. Then, substituting for μ and σ^2,

$$\begin{aligned}\chi^2_1 &= (a\text{-}np)^2/npq,\\ &= (a\text{-}np)^2/np + (a\text{-}np)^2/nq,\\ &= (a\text{-}np)^2/np + (b\text{-}nq)^2/nq,\\ &= \Sigma[\,(\text{obs}-\text{exp})^2/\text{exp}] \text{ exactly.}\end{aligned}$$

It is obvious that the expectation is exactly 1 because $E(a\text{-}np)^2 = npq$ by the definition of variance, and if all the χ_1^2's are independent then the expectation of their sum will exactly equal the degrees of freedom.

It follows that although the fine detail of the χ^2 distribution may be wrong its mean and variance are right, and there can be no doubt over its applicability save when sample sizes are very small, much smaller than Mendel's, and even then the expectation is still correct.

V. A FRESH ANALYSIS

The 84 d.f. in Mendel's data may be individually identified in a number of different ways depending on the order in which the simultaneous segregations are considered. I shall arbitrarily order the loci according to Mendel's own classification. Thus in analysing the results of selfing AaBb where A, B is the order of the loci in Mendel's list, I shall take the segregation at the A locus first and then the two segregations at the B locus for each of the two A-locus phenotypes, making three independent comparisons in all. With this arrangement we will then have 84 independent variates to analyse.

For our purposes it will be better to work with χ rather than χ^2, partly because it preserves the information about the direction in which the segregation departs from expectation, but also because χ is expected to have a normal distribution to an excellent approximation, and methods for analysing a sequence of normal variables are commonplace. By the argument of the previous section the expectation of each χ on the binomial hypothesis is exactly 0 and its variance exactly 1, so that its distribution is $N(0, 1)$ for all practical purposes. It may also be noted that since in the binomial case χ is given exactly by dividing the observed deviation from the hypothetical binomial mean by the hypothetical standard deviation, whether we derive it as the square-root of Pearson's χ^2 criterion or directly as a standardized deviation is immaterial.

There is, of course, also an arbitrariness as to which departures from expectation we regard as positive and which as negative, but we will consistently regard a deviation in favour of the first of the two segregating classes as positive, and the first class will always

Table 1. *Notation for Mendel's seven characters (M, p 17).*

	Seed characters
A, a	Form of seed
B, b	Colour of albumen
	Plant characters
C, c	Colour of the seed-coats
D, d	Form of pods
E, e	Colour of the unripe pods
F, f	Position of flowers
G, g	Length of stem

be taken to be the one with the larger expectation. In the case of the experiments on gametic ratios, where 1:1 is expected, the first class will always be the one with the larger number of dominant genes, so the order will be either AA:Aa or Aa:aa as the case may be.

The notation for the characters is given in Table 1, and agrees with Mendel's for the major part of his paper. Table 2 gives the 84 values of χ and their origins. The page references are to Mendel's paper in Bennett (1965), and 'A' stands for AA+Aa and 'a' for aa in accordance with the usual convention. As mentioned at the beginning of our Section III, the data for the illustration of plant variation (M, p. 17) should be subtracted out from the data in the first two experiments (M, pp. 16–17), and this has been done. Similarly, modified expectations based on the ratio 0·629124:0·370876 have been used instead of 2:1 where appropriate. All calculations have been rounded to four significant figures after completion.

It will be seen that of the fifteen χ values for the segregation ratio 0·6291:0·3709 (rows 30–35 and 61–69), 13 are positive and only 2 negative, suggesting something of a bias towards the larger class. The total segregation is 720:353, or 0·6710:0·3290, with an associated χ value of +2·8408, indicating a very poor fit indeed. Mendel (M, pp. 20, 21) explicitly accepted this type of experiment as establishing the 2:1 ratio, and as an aid to judging the extent to which the data conform to this rather than the true ratio we can compute the likelihood ratio for the 2:1 hypothesis *vs.* the 0·6291:0·3709 hypothesis on the assumption of binomial variation (for likelihood techniques, see Edwards, 1972). The result is 57·90, or a support value (taking logs) of just over 4, which is large enough to substantiate what Sturtevant (1965) called 'Fisher's most telling point' in his argument that the data have been biased in the direction of agreement with what Mendel expected.

Having concluded from an examination of the *means* for these fifteen segregations that possibly all is not well, we now turn to an examination of the *variances* of the other 69 segregations. For we know from the analyses considered in Section III that it is the infra-binomial variance that raises doubts about the data. As far as the means are concerned, of their 69 χ values, 38 are positive, 30 negative, and 1 zero. Their mean is 4·0880 which, dividing by $\sqrt{69}$ to make it a standard normal deviate, is equal to 0·4291, as unexceptionable a value as one could wish for. By contrast, the variance of the χ values is certainly exceptional. χ^2 is now the appropriate statistic, and the sum of the 69 χ^2 values from Table 2 is 30·8138, highly remarkable on any interpretation of tests of significance.

Table 2. *Mendel's segregations and their χ^2 values.*

	Character	Segregation: Expected	Observed		Total	χ
		F_2, Seed characters, pp. 16–17				
1	A	3:1	5138	1749	6887	−0·7583
2	B	3:1	5667	1878	7545	+0·2193
		F_2, Illustrations of plant variability, p. 17				
3	A	3:1	45	12	57	+0·6882
4	A	3:1	27	8	35	+0·2928
5	A	3:1	24	7	31	+0·3111
6	A	3:1	19	10	29	−1·1793
7	A	3:1	32	11	43	−0·0880
8	A	3:1	26	6	32	+0·8165
9	A	3:1	88	24	112	+0·8729
10	A	3:1	22	10	32	−0·8165
11	A	3:1	28	6	34	+0·9901
12	A	3:1	25	7	32	+0·4082
13	B	3:1	25	11	36	−0·7698
14	B	3:1	32	7	39	+1·0170
15	B	3:1	14	5	19	−0·1325
16	B	3:1	70	27	97	−0·6448
17	B	3:1	24	13	37	−1·4237
18	B	3:1	20	6	26	+0·2265
19	B	3:1	32	13	45	−0·6025
20	B	3:1	44	9	53	+1·3482
21	B	3:1	50	14	64	+0·5774
22	B	3:1	44	18	62	−0·7332
		F_2, Plant characters, p. 18				
23	C	3:1	705	224	929	+0·6251
24	D	3:1	882	299	1181	−0·2520
25	E	3:1	428	152	580	−0·6712
26	F	3:1	651	207	858	+0·5913
27	G	3:1	787	277	1064	−0·7788
		F_3, Seed characters, pp. 19, 20				
28	A	2:1	372	193	565	−0·4165
29	B	2:1	353	166	519	+0·6518
		F_3, Plant characters, p. 20				
30	C	0·63:0·37	64	36	100	+0·2252
31	D	0·63:0·37	71	29	100	+1·6743
32	E	0·63:0·37	60	40	100	−0·6029
33	F	0·63:0·37	67	33	100	+0·8462
34	G	0·63:0·37	72	28	100	+1·8813
35	E	0·63:0·37	65	35	100	+0·4322
		F_2, Bifactorial experiment, p. 23				
36	A	3:1	423	133	556	+0·5876
37	B, among 'A'	3:1	315	108	423	−0·2526
38	B, among 'a'	3:1	101	32	133	+0·2503

Table 2. (*cont.*)

	Character	Segregation Expected	Observed	Total	χ
		F_3, Bifactorial experiment, p. 24			
39	A, among 'AB'	2:1	198	103	301	−0·3261
40	A, among 'Ab'	2:1	67	35	102	−0·2100
41	B, among 'aB'	2:1	68	28	96	+0·8660
42	B, among Aa 'B'	2:1	138	60	198	+0·9045
43	B, among AA 'B'	2:1	65	38	103	−0·7664
		F_2, Trifactorial experiment, p. 26, seed characters			
44	A	3:1	480	159	639	+0·0685
45	B, among 'A'	3:1	367	113	480	+0·7379
46	B, among 'a'	3:1	122	37	159	+0·5037
		F_3, Trifactorial experiment, p. 26, seed characters			
47	A, among 'AB'	2:1	245	122	367	+0·0369
48	A, among 'Ab'	2:1	76	37	113	+0·1330
49	B, among 'aB'	2:1	79	43	122	−0·4481
50	B, among Aa 'B'	2:1	175	70	245	+1·5811
51	B, among AA 'B'	2:1	78	44	122	−0·6402
		F_2, Trifactorial experiment, p. 26, plant character			
52	C, among AaBb	3:1	127	48	175	−0·7419
53	C, among AaBB	3:1	52	18	70	−0·1380
54	C, among AABb	3:1	60	18	78	+0·3922
55	C, among AABB	3:1	30	14	44	−1·0445
56	C, among Aabb	3:1	60	16	76	+0·7947
57	C, among AAbb	3:1	26	11	37	−0·6644
58	C, among aaBb	3:1	55	24	79	−1·1043
59	C, among aaBB	3:1	33	10	43	+0·2641
60	C, among aabb	3:1	30	7	37	+0·8542
		F_3, Trifactorial experiment, p. 26, plant character			
61	C, among AaBb	0·63:0·37	78	49	127	−0·3488
62	C, among AaBB	0·63:0·37	38	14	52	+1·5174
63	C, among AABb	0·63:0·37	45	15	60	+1·9384
64	C, among AABB	0·63:0·37	22	8	30	+1·1816
65	C, among Aabb	0·63:0·37	40	20	60	+0·6020
66	C, among AAbb	0·63:0·37	17	9	26	+0·2610
67	C, among aaBb	0·63:0·37	36	19	55	+0·3903
68	C, among aaBB	0·63:0·37	25	8	33	+1·5276
69	C, among aabb	0·63:0·37	20	10	30	+0·4257
		Gametic ratios, p. 32, seed characters, first experiment			
70	A	1:1	43	47	90	−0·4216
71	B, among AA	1:1	20	23	43	−0·4575
72	B, among Aa	1:1	25	22	47	+0·4376
		Gametic ratios, p. 32, seed characters, second experiment			
73	A	1:1	57	53	110	+0·3814
74	B, among Aa	1:1	31	26	57	+0·6623
75	B, among aa	1:1	27	26	53	+0·1374

Table 2. (*cont.*)

	Character	Segregation: Expected	Observed		Total	χ
	Gametic ratios, p. 32, seed characters, third experiment					
76	A	1:1	44	43	87	+0·1072
77	B, among AA	1:1	25	19	44	+0·9045
78	B, among Aa	1:1	22	21	43	+0·1525
	Gametic ratios, p. 32, seed characters, fourth experiment					
79	A	1:1	49	49	98	0·0000
80	B, among Aa	1:1	24	25	49	−0·1429
81	B, among aa	1:1	22	27	49	−0·7143
	Gametic ratios, p. 32, plant characters					
82	G	1:1	87	79	166	+0·6209
83	C, among Gg	1:1	47	40	87	+0·7505
84	C, among gg	1:1	38	41	79	−0·3375

We now go further than any previous analysis of Mendel's data by looking at the distribution of the 69 χ values themselves to see whether the unexpectedly low variance is a reflection of any particular anomaly. Fig. 1 displays the values on normal probability paper. It is immediately obvious that the reduced variance is not characteristic of the whole data, as Weiling's theory would require, but is confined to the tails of the distribution, where the extreme variates are not extreme enough to conform to expectation. Between the upper and lower quartiles (the 25-percentiles) of the expected distribution the slope of the plot is acceptable, but outside these points it is not. It must be remembered that a normal probability plot is not a linear regression, so that intuitive judgments of the departure from expectation may be misleading; for comparison, a simulated plot from a standard normal distribution is given in Fig. 2, using the first 69 random normal deviates from Lindley & Scott (1984), reading columnwise.

Another way to judge the apparent lack of extreme deviates is by examining the range, which is from −1·4237 (row 17) to +1·5811 (row 50), a total of 3·0048, whereas the expected range for a sample of 69 random normal deviates is 4·7440. Indeed, the expected absolute values of the first three most extreme deviates out of 69 are 2·6216, 2·2755, and 2·0828, and nothing on Mendel's data approaches these. The inescapable conclusion is that some segregations beyond the outer 5-percentiles (approximately) have been systematically biased towards their expectations so as to fall between the 5-percentiles and the 25-percentiles. Further analysis shows that the effect is not confined to particular sample sizes or segregation ratios, but is quite general.

It is relevant to recall here that Mendel thought it worthwhile to repeat an experiment in which a segregation of 60:40 had appeared where he had expected 2:1 (M, p. 20) a segregation with a corresponding χ-value of only −1·4142. As Fisher (1936) observed, this is not a significant value; yet only twice out of 69 times does Mendel record more extreme segregations (those quoted above). We should not forget, however, that when Mendel actually singled out the most extreme segregations in the first two experiments for comment (M, p. 17) they were no different from what one might expect (Fisher, 1936).

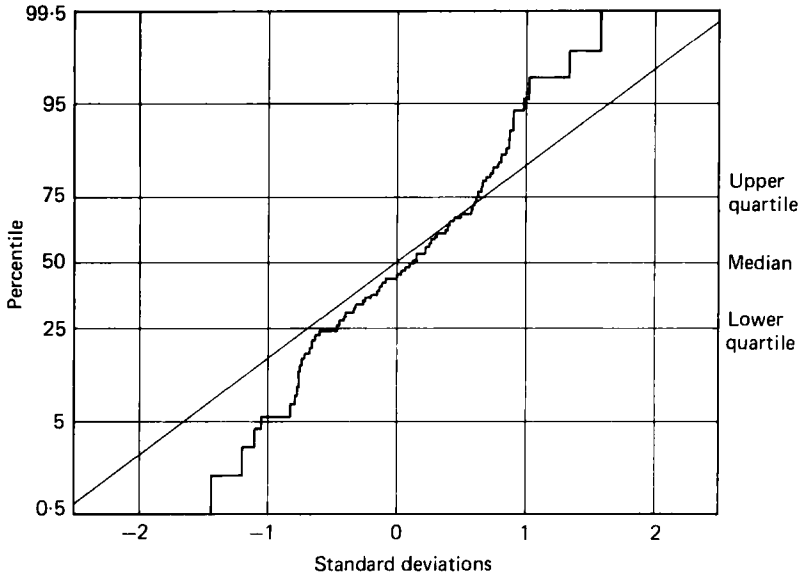

Fig. 1. The values of χ for the 69 segregations with undisputed expectations plotted on normal probability paper.

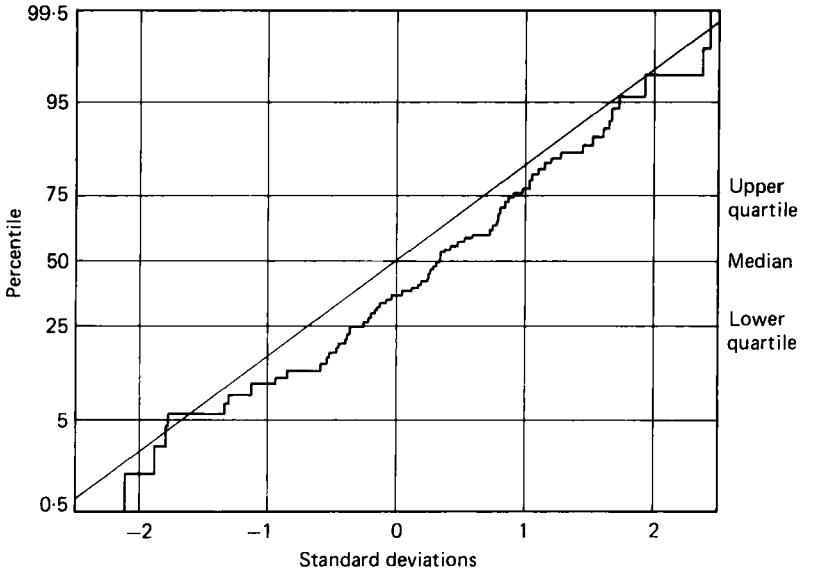

Fig. 2. The values of 69 simulated standard normal deviates plotted for comparison with Fig. 1.

VI. CONCLUSIONS

We thus see that Mendel's data exhibit two independent peculiarities both of which point to the same conclusion, namely, that in general the segregations agree more closely with what Mendel expected than chance would dictate. Where Mendel expected a 2:1

ratio even though the true expectation must have been different it was the former ratio which was achieved, and throughout the rest of his results there is a persistent lack of extreme segregations.

In Sections II and III above I reviewed earlier analyses and explanations and concluded that only Fisher (1936) could withstand detailed criticism. It may be helpful if I admit at this point that for many years I supposed that Fisher's analysis was going to be able to be faulted because of its total reliance on the 'repeated sampling' logic of the χ^2 goodness-of-fit test which I had come to mistrust, but a complete review of the whole problem has now persuaded me that his 'abominable discovery' must stand. I agree with Sturtevant (1965): 'In summary, then, Fisher's analysis of Mendel's data must stand essentially as he stated it.' As to the precise method by which the data came to be adjusted I would rather not speculate, though it does seem to me that any criticism of Mendel himself is quite unwarranted. Even if he were personally responsible for the biased counts his actions should not be judged by today's standards of data recording. In 1967 Dobzhansky reviewed six separate volumes celebrating the centenary of Mendel's paper and had this to say:

> I believe that a far simpler explanation [than Fisher's suggestion of an over-zealous assistant] is at least as plausible. Few experimenters are lucky enough to have no mistakes or accidents happen in any of their experiments, and it is only common sense to have such failures discarded. The evident danger is ascribing to mistakes and expunging from the record perfectly authentic experimental results which do not fit one's expectations. Not having been familiar with chi-squares and other statistical tests, Mendel may have, in perfect conscience, thrown out some crosses which he suspected to involve contamination with foreign pollen or other accident.

With the reservation that in order to explain some of the results misclassification rather than simple throwing away of plants will have to be invoked, I agree with Dobzhansky. Perhaps after all the undergraduate Fisher came close to the correct solution 75 years ago: 'It may be that the worthy German abbot, in his ignorance of probable error, unconsciously placed doubtful plants on the side which favoured his hypothesis.'

VII. SUMMARY

1. All the analyses and discussions of Mendel's data in his 1866 paper 'Experiments in plant hybridization' are reviewed, with special reference to the suggestion by Fisher (1936) that the segregations are in general closer to Mendel's expectations than chance would dictate. It is concluded that in spite of many attempts to find an explanation, Fisher's suggestion that the data have been subjected to some kind of adjustment must stand.

2. A fresh analysis based on the individual consideration of the 84 independent segregations reported by Mendel confirms this conclusion in two separate ways. In the words of my title, Mendel's results really are too close.

VIII. REFERENCES

BABBAGE, C. (1830). *Reflections on the Decline of Science in England and on Some of its Causes*. Fellowes, London.

BATESON, W. (1909). *Mendel's Principles of Heredity*. Cambridge University Press.

BATESON, W. & KILLBY, H. (1905). Experimental studies in the physiology of heredity: Peas (*Pisum sativum*). In *Royal Society Reports to the Evolution Committee* **2**, 55–80.

BARNDORFF-NIELSEN, O. (1976). Plausibility inference. *Journal of the Royal Statistical Society* B**38**, 103–131.

BEADLE, G. W. (1967). 'Mendelism, 1965'. In *Heritage from Mendel* (ed. R. A. Brink), pp. 335–350. University of Wisconsin Press, Madison.

BENNETT, J. H. (ed.). (1965). *Experiments in Plant Hybridisation*, by Gregor Mendel, with a Commentary and Assessment by R. A. Fisher. Oliver & Boyd, Edinburgh.

BENNETT, J. H. (ed.). (1983). *Natural Selection, Heredity, and Eugenics*. Clarendon Press, Oxford.

CREW, F. A. E. (1966). *The Foundations of Genetics*. Pergamon, Oxford.

DARBISHIRE, A. D. (1908). On the result of crossing round with wrinkled peas, with especial reference to their starch-grains. *Proceedings of the Royal Society* B**80**, 122–135.

DARBISHIRE, A. D. (1909). An experimental estimation of the theory of ancestral contributions in heredity. *Proceedings of the Royal Society* B **81**, 61–79.

DE BEER, G. (1964). Mendel, Darwin, and Fisher (1865–1965). *Note and Records of the Royal Society* **19**, 192–226.

DOBZHANSKY, T. (1967). Looking back at Mendel's discovery. *Science* **156**, 1588–1589.

DUNN, L. C. (1965*a*). *A Short History of Genetics*. McGraw-Hill, New York.

DUNN, L. C. (1965*b*). Mendel, his work and place in history. *Proceedings of the American Philosophical Society* **109**, 189–198.

EDWARDS, A. W. F. (1972). *Likelihood*. Cambridge University Press. (Reprinted 1984.)

EDWARDS, A. W. F. (1986). More on the too-good-to-be-true paradox and Gregor Mendel. *Journal of Heredity* **77**, 138.

FISHER, R. A. (1936). Has Mendel's work been rediscovered? *Annals of Science* **1**, 115–137. (Reprinted in Bennett, 1965, pp. 59–87, and Stern & Sherwood, 1966, pp. 139–172.)

FISHER, R. A. (1952). Statistical methods in genetics. *Heredity* **6**, 1–12.

GALTON, F. (1892). *Hereditary Genius*, 2nd ed., Macmillan, London.

GALTON, F. (1904). Average number of kinsfolk in each degree. *Nature* **70**, 529 and 626.

JEFFREYS, H. (1939). *Theory of Probability*. Clarendon Press, Oxford. (Reprinted 1983.)

LINDLEY, D. V. & SCOTT, W. F. (1984). *New Cambridge Elementary Statistical Tables*. Cambridge University Press.

MENDEL, G. (1866). Versuche über Pflanzen-Hybriden. *Verhandlungen des naturforschenden Vereines in Brünn* **4**, 3–47. (English translations in Bateson, 1909 [reprinted in Bennett, 1965, pp. 7–51] and Stern & Sherwood, 1966, pp. 1–48.)

MERSENNE, M. (1637). *L'Harmonie Universelle, Seconde Partie*. Ballard, Paris.

MONAGHAN, F. & CORCOS, A. (1985). Chi-square and Mendel's experiments: where's the bias? *Journal of Heredity* **76**, 307–309.

NORTON, B. & PEARSON, E. S. (1976). A note on the background to, and refereeing of, R. A. Fisher's paper '[On] The correlation between relatives on the supposition of Mendelian inheritance'. *Notes and Records of the Royal Society* **31**, 151–162.

OLBY, R. C. (1966). *Origins of Mendelism*, Schocken, New York, (2nd ed. University of Chicago Press, 1985).

OREL, V. (1968). Will the story on 'too good' results of Mendel's data continue? *BioScience* **18**, 776–778.

OREL, V. (1984). *Mendel*. Oxford University Press.

PEARSON, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *London, Edinburgh, and Dublin Philosophical Magazine, Fifth series* **50**, 157–175.

PIEGORSCH, W. W. (1983). The questions of fit in the Gregor Mendel controversy. *Communications in Statistics: Theory and Methods* **12**, 2289–2304.

PILGRIM, I. (1984). The too-good-to-be-true paradox and Gregor Mendel. *Journal of Heredity* **75**, 501–502.

ROOT-BERNSTEIN, R. S. (1983). Mendel and methodology. *History of Science* **21**, 275–295.

STERN, C. & SHERWOOD, E. R. (1966). *The Origin of Genetics: A Mendel Source Book*. Freeman, San Francisco.

STURTEVANT, A. H. (1965). *A History of Genetics*. Harper & Row, New York.

THODAY, J. M. (1966). Mendel's work as an introduction to genetics. *Advancement of Science* **23**, 120–124.

VAN DER WAERDEN, B. L. (1968). Mendel's experiments. *Centaurus* **12**, 275–288.

WEILING, F. (1966). Hat J. G. Mendel bei seinen Versuchen 'zu genau' gearbeitet? – Der χ^2-Test und seine Bedeutung für die Beurteilung genetischer Spaltungsverhältnisse. *Der Züchter* **36**, 359–365. (An English translation by W. W. Piegorsch is available as Paper BU-718-M of the Biometrics Unit, Cornell University, Ithaca, New York.)

WEILING, F. (1971). Mendel's 'too good' data in *Pisum* experiments. *Folia Mendeliana* **6**, 75–77.

WEILING, F. (1985). What about R. A. FISHER's statement of the 'too good' data of J. G. MENDEL's *Pisum*-paper? *45th Session of the International Statistical Institute*. Amsterdam.

WELDON, W. F. R. (1902). Mendel's laws of alternative inheritance in peas. *Biometrika* **1**, 228–254.

WRIGHT, S. (1966). Mendel's ratios. In Stern & Sherwood (1966), 173–175.