

# Statistics is not enough: revisiting Ronald A. Fisher's critique (1936) of Mendel's experimental results (1866)

Avital Pilpel

*Philosophy Department, University of Haifa, 1901 Eshkol Tower, Mount Carmel, Haifa 31905, Israel*

Received 28 April 2006; received in revised form 1 November 2006

## Abstract

This paper is concerned with the role of rational belief change theory in the philosophical understanding of experimental error. Today, philosophers seek insight about error in the investigation of specific experiments, rather than in general theories. Nevertheless, rational belief change theory adds to our understanding of just such cases: R. A. Fisher's criticism of Mendel's experiments being a case in point. After an historical introduction, the main part of this paper investigates Fisher's paper from the point of view of rational belief change theory: what changes of belief about Mendel's experiment does Fisher go through and with what justification. It leads to surprising insights about what Fisher had done right and wrong, and, more generally, about the limits of statistical methods in detecting error. © 2007 Elsevier Ltd. All rights reserved.

**Keywords:** R. A. Fisher; Gregor; Mendel;  $\chi^2$  test; Error; Belief change

When citing this paper, please use the full journal title *Studies in History and Philosophy of Biological and Biomedical Sciences*

## 1. Introduction

Ever since Kuhn and Quine, philosophy of science had been shifting its focus from the general and normative to the specific and descriptive. It is not that philosophers of science had ceased to ask the big questions, but new insights into such questions were being sought by investigating the actual behavior of scientists performing particular experiments. The focus on actual, specific historical experiments also legitimized new areas of philosophical research. Traditionally, philosophy of science saw experimental error as an incidental, annoying feature of no intrinsic philosophical interest. But investigation of experiments had shown that experimental error of various sorts is ubiquitous and intrinsic to scientific work, and its effects

cannot be dismissed, “subtracted” away easily, or ignored if one wished to gain insight into the nature of science. Indeed, investigation of error is crucial for such insight (see Mayo, 1996). Nevertheless, general normative theories—in particular, rational belief change theory—can sometimes add important insights to our analysis of experimental error.

A scientific paper usually claims there is a good reason to change our beliefs in some way. Typically this is due to presenting new data, but some scientific papers show that results of a previous experiment justify coming to believe error of some sort exists in it. Such papers are of course an exception: typically, error is suspected, discovered, and dealt with *in medias res*, and without those suspecting error bothering with formal analysis as to why

*E-mail address:* [avital.pilpel@gmail.com](mailto:avital.pilpel@gmail.com)

such a suspicion is justified. Sometimes, however, such an error-detecting paper does show, implicitly or explicitly, a commitment to a more robust view of rational belief change. In such cases one can look critically at the rational belief change theory involved and consider both its recommendations and whether its holder applies it consistently. One such case is R. A. Fisher's criticism of Mendel in 'Has Mendel's work been rediscovered?' (Fisher, 1936). Using statistical analysis, in particular  $\chi^2$  tests, Fisher concluded that Mendel's results show a systematic "fudging" of the results to make them fit with his expectations. I start with an historical introduction and then analyze Fisher's paper from the point of view of rational belief change theory. My conclusions are threefold. First (as others noted before), Fisher did not accuse Mendel of cheating. Second, his conclusions do not follow automatically from the statistics, and cannot be reduced to essentially inter-statistical "hypothesis testing"; Fisher was, rather, following (implicit) rules for rational belief change. Third, Fisher should have, according to his own lights, considered the possibility of other sources of error than data "cooking". He missed this possibility because he used a "rule of thumb" for rational belief change (for example, when high  $p$  values justify coming to believe deliberate cheating occurred), which usually works, but in this case was inapplicable. This lends credence to the views of Hon (1989) and others that a typology of experimental error has value over and above statistical analysis: often, statistical analysis can show the existence of error, and even point towards its likely source, but it is not enough.

## 2. Historical background

On 8 February and 8 March 1865, Gregor Mendel reported on his experiments with pea plants, conducted between 1856 and 1863, in meetings of the Natural Science Society of Brunn. He published them in the society's journal (Mendel, 1866). Mendel is generally accepted today as the "father of genetics" for his discovery of the combinatorial rules of inherited traits, and for his hypothesis that such traits are caused by the combination of discrete 'factors' (i.e., genes), though he was unclear about the ontological status of his 'factors'.

In his paper, Mendel discusses the inheritance laws concerning the two distinct 'differentiating characters' (*differierende Merkmale*, i.e., phenotypes) of seven different traits: form of seed, color of seed albumen, color of seed coat, form of ripe pods, color of unripe pods, position of flowers, and length of stem. For example, a seed can be either round or 'irregularly angular', but not both, nor anything else or in between (ibid., §3).

Mendel's goal was 'to observe these variations in the case of each pair of differentiating characters, and to deduce the law according to which they appear in successive generations' (ibid.). For each trait, Mendel started by creating a generation of hybrids (the 'F1' generation). Mendel's first set of experiments concerned the inheritance

laws for the phenotypes of a single trait (ibid., §5). The F1 hybrids—all exhibiting the dominant phenotype (round pea, green albumen, etc.)—were self-fertilized, demonstrating the 3:1 segregation ratio between the dominant and recessive phenotype in the second generation (F2). His second series of experiments attempted in various ways to determine their genotype—in Mendel's words, their 'internal composition', *inneren Beschaffenheit* (ibid., §§6, 7)—discovering the famous 1:2:1 law. Finally, Mendel considered whether 'the law of development discovered in these applied to each pair of differentiating characters when several diverse characters are united in the hybrid by crossing' (ibid., §8). In fact, it did.

For various reasons whose exact nature is debated among historians of biology, Mendel's work was ignored. Its importance was not generally recognized until it was rediscovered (independently) in 1900 by de Vries, Correns, and von Tschermak. Soon afterwards, Mendel's work was criticized by Weldon (1902). His criticism is based on statistical analysis—probable error and 'the method of Pearson', that is, the newly developed  $\chi^2$  tests (Pearson, 1900). He first considers Mendel's results for 'individuals with dominant characters in the second hybrid generation' (i.e., the results of Mendel's experiments with the 3:1 law) and concludes that:

Only one determination has a deviation from the hypothetical frequency greater than the probable error, and one has a deviation sensibly equal to the probable error; so that a discrepancy between the hypothesis and observations which is equal to or greater than the probable error occurs twice in seven times, and deviations much greater than probable error do not occur at all. These results then accord so remarkably with Mendel's summary that if they were repeated a second time ... the chance that the agreement between observation and hypothesis would be worse than that actually obtained is about 16 to 1. (Weldon, 1902)

Mendel's luck holds throughout: when considering the 'proportion of plants with dominant characters among hybrids of the second generation, which transmitted only the dominant characters to their offspring' (i.e., the experiments establishing the 1:2:1 law) Weldon finds that 'Mendel's statement is admirably in accord with his experiment'. Finally, looking at the frequency of 'various possible combinations of characters in hybrids of the second generation from races which differed in three characters' (see Mendel, 1866, §8), he finds that 'applying the method of Pearson':

the chance that a system will exhibit deviations as great as or greater than these from the result indicated by Mendel's hypothesis is about 0.95 ... or if the experiment were repeated a hundred times, we should expect to get a worse result about 95 times, or the odds against a result as good as this or better are 20 to 1. (Weldon, 1902)

Weldon's paper needs to be seen in the context of the time: the conflict between the Darwinist biometricians

(such as Weldon and Karl Pearson) and the Mendelians (such as Bateson), the key issue being whether evolution is continuous or discontinuous. Indeed, the bulk of Weldon's paper, after the criticism of Mendel, is an attempt to construct a non-Mendelian theory of inheritance for peas. Weldon saw Mendel's claims that the genotype determines discrete phenotypes and that the offspring's genotype is being determined solely by the parents as inconsistent with Darwinism. Mendel's view was seen as inconsistent with Darwinism not only by the pro-Darwin biometricians, but also by the anti-Darwin Mendelians (Sapp, 1990).

It is not within the scope of this paper to discuss the details of this conflict, or why Weldon's paper had little effect.<sup>1</sup> That said, part of the reason for the latter is that Weldon's paper did not actually claim anything too surprising about Mendel. Weldon did not say Mendel fabricated his results, merely that repeating his experiments was not likely to yield such good results. He accepted as given that Mendel was anti-Darwinian. Nor had he made any investigation of possible errors in Mendel's work.

By 1936, when Fisher published his famous paper, the conflict between Mendelians and Darwinists had been mostly resolved—in no small part due to the work of Fisher himself (together with J. B. S. Haldane and S. Wright), leading to the 'evolutionary synthesis' of Mendelism and Darwinism by 1930 (Haldane, 1924, 1926, 1932; Fisher, 1930; Wright, 1931; Olby, 1997). (Some claim, however, that the conflict was not in reality ever solved.) Despite being a key figure in this development, Fisher declared in his paper, after a similar (if significantly more thorough) statistical analysis of Mendel results using  $\chi^2$  tests:

A serious and almost inexplicable discrepancy has, however, appeared, in that in one series of results the numbers observed agree excellently with the two to one ratio, which Mendel himself expected, but differ significantly from what should have been expected had his theory been corrected to allow for the small size of his test progenies. To suppose that Mendel recognized this theoretical complication, and adjusted the frequencies supposedly observed to allow for it, would be to contravene the weight of the evidence supplied in detail by his paper as a whole. Although no explanation can be expected to be satisfactory, it remains a possibility among others that Mendel was deceived by some assistant who knew too well what was expected. (Fisher, 1936, p. 132)

Fisher's paper had a much stronger influence than Weldon's. Again, without attempting a complete historical analysis as to why, part of the reason is that Fisher reaches far stronger conclusions than Weldon: he says the data were cooked, or 'fictitious' (ibid, p. 129). Fisher's (rather careful) formulation of the possible explanations for this "cooking" was seen as a thinly veiled accusation of deliber-

ate cheating against Mendel. The result of Fisher's criticism (as well as, it should be said, of later critics who argued that Mendel cheated based on historical and biographical data) was to greatly hurt Mendel's reputation. Indeed, even Martin Gardner, perhaps the most famous science writer for the general public in the USA, included Mendel in his list of 'great fakes of science' (Gardner, 1977).

Defenders of Mendel divide into two main camps. One camp is "statistical": it denies the reliability of Fisher's statistical analysis, arguing that the results were not actually too good to be true (Nissani, 1994; Weiling, 1986, 1989, etc.). More radically, some say that it is a non sequitur to argue that results that are very close to expectation should be any reason to suspect error in the first place, as such suspicion amounts to an unjustifiable "two-tailed" rejection test for the hypothesis (Pilgrim, 1984). The other camp is "historical": it accepts Fisher's criticism, but provides historical evidence about either Mendel's life or the conditions in his garden that shows cheating was impossible, and therefore that there must be some other explanation for Mendel's results. These include, inter alia, the claim that Mendel openly, with no intent to deceive, presented only his best results for pedagogical purposes (Fairbanks and Rytting, 2001); that he had unconsciously misclassified some plants in a way that favored his expectation (Wright, 1966); or that other errors, such as a non-random pollenization, are the culprit (Sturtevant, 1965; Weiling, 1991; Seidenfeld, 1998).

### 3. Did Fisher accuse Mendel of cheating?

The first question to settle is whether Fisher accusing Mendel of cheating. Historical evidence does not seem to support this view. At the same time that he was writing his famous paper, he wrote to his colleague, E. B. Ford, that:

Now, when data had been faked, I know very well how generally people underestimate the frequency of wide chance deviations, so that the tendency is always to make them agree too well with expectations . . . the deviations [in Mendel's data] are shockingly small . . . I have divided the data in several different ways to try to get a further clue, e.g. by years and by the absolute sizes of the numbers, but . . . can get no clue to the method of doctoring. (Box, 1978, p. 297)

Soon afterwards, when submitting the manuscript of his paper to the editor of the *Annals of Science* (D. McKie), he added a note saying:

I had not expected to find the strong evidence which has appeared that the data had been cooked. This makes my paper far more sensational than I ever had intended . . . but I cannot help it if circumstances proceed to emphasize so strongly my main point, that Mendel's own

<sup>1</sup> For a summary of the conflict from the Kuhnian point of view, see Farrall (1975). See also Froggatt & Levin (1971). Numerous other papers exist.

work, in spite of the immense publicity it has received, has only been examined superficially, and that the rediscoverers of 1900 were just as incapable as the non-discoverers of 1870 of assimilating from it any idea other than those which they were already prepared to accept. (Ibid.)

These two notes make clear that Fisher in private, as well as in public, does not accuse Mendel of faking the data. He is convinced that the data were cooked, but confesses that he has no idea how—and, a fortiori, by whom—it was done. Fisher does not take Mendel off his list of suspects, but not concluding Mendel did not cheat is a far cry from claiming that Mendel *did* cheat.

Internal evidence also argues against Fisher accusing Mendel of cheating. The statistical analysis of Mendel's results only begins in Section 3 ('An attempted reconstruction'). Its conclusion, quoted above, is at the beginning of Section 4 ('The nature of Mendel's discovery'). Before it lie two other sections ('The polemic use of the rediscovery' and 'Should Mendel be taken literally?') and, after it, the rest of Section 4, and Section 5 ('The contemporary reaction to Mendel's views'). Fisher's statistical analysis and its conclusion about Mendel's results should be seen in this context. The real goal of Fisher's 1936 paper is not to accuse Mendel of cheating, but to rescue Mendel from the various misinterpretations forced upon him by the prejudices of his rediscoverers and promoter in the early twentieth century, who could not view Mendel's work objectively owing to their own (anti-Darwinian) prejudices. In particular, Fisher disagrees on two points with Bateson, who is the real target of Fisher's paper (Sapp, 1990).

Bateson (1909) claims that Mendel 'embarked on his experiments with peas' after 'not finding himself in full agreement' with 'the views of Darwin which were at that time coming into prominence'. Also, Mendel's results were actually summaries. Otherwise, Mendel must have been 'incredibly wasteful' with his data. He must have considered, in every experiment, only one or two 'factors' (seed color, height of stem, etc.) and ignored the others.

Fisher's reply to Bateson is extremely ingenious. He starts by considering whether Mendel's results are summaries, or to be taken literally. He quickly concludes that they *should* be taken literally, despite the 'incredible wastefulness':

The Great objection to the view suggested by Bateson ... that Mendel's 'experiments' are fictitious, and that the paper is a didactic exposition ... lies in the words which Mendel himself used in introducing the successive steps of his account, e.g., at the beginning of the eight section ... 'The next task consisted in ascertaining ...' ... and the opening sentence of the ninth section ... 'The results of the experiments previously described led to further experiments.' ... if Mendel is not to be taken literally, when he implies that one set of data was available when the next experiment was planned, he is taking, as *redacteur*, excessive and unnecessary lib-

erties with the facts. Moreover, the style throughout suggests that he expects to be taken entirely literally. (Fisher, 1936, pp. 119–120)

But if Mendel's results are to be taken literally, what do they show us? Here, Fisher embarks on the 'attempted reconstruction'—the statistical analysis—at the end of which he concludes that Mendel was 'possibly deceived by an assistant'. (Clearly, there is no point in such statistical analysis of results that are not intended as more than summaries or pedagogical tools.) Yet this is just the starting point of Fisher's real argument.

If Mendel *could* be biased towards these expected results, it follows that Mendel knew in 1859, at the latest, what experimental results to expect from his laws of genetics: one cannot be biased towards an expected result unless one knows it *is* expected. In particular, Mendel already knew in 1859 what he wrote in 1866, that '*n* factors would give rise to  $3^n$  different genotypes, of which  $2^n$  would be capable of breeding true' (ibid., p. 133). Furthermore, he probably realized that even for small *ns*, these numbers would be sufficient to allow, for both inter- and intra-species crosses, enough varieties that 'breed true' (i.e., potential new species) to satisfy Darwin.

Thus, for Fisher, the surprising discovery that the results were "cooked" is (further) evidence that Mendel was a good Darwinian—which is the main point of his paper! Clearly Fisher would have no interest in using those very same "cooked" results as evidence that Mendel, whom Fisher is defending against what he considers an unfair misrepresentation of his views by the anti-Darwinists, was himself a cheat. The evidence, then, is strong that Fisher meant just what he said: he concluded that *someone* cooked the results, but he doesn't know who, nor does he accuse Mendel.

#### 4. Fisher's belief change process

We can now move on to our main task: the analysis of Fisher's criticism of Mendel from the point of view of rational belief change theory. Fisher begins his investigation about Mendel with a corpus of beliefs, **K**, about the experiments. Initially, Fisher considered Mendel's results free of error ( $\sim E$ ); as both his paper and the letters quoted above make clear, the discovery of the 'serious and almost inexplicable discrepancy' was a surprise. So  $\sim E \in \mathbf{K}$ .

In the first and second sections, Fisher is concerned with whether Mendel's work is to be taken literally (**L**) or is a summary (**S**). He decided to add **L** to his original corpus of beliefs, **K**, reaching a new  $\mathbf{K}_L^+$ . But if this is so, it follows that the results can be subjected to statistical analysis, and the conclusions of this analysis (**P**<sub>1</sub> and **P**<sub>2</sub> below) must be taken literally as well. So **P**<sub>1</sub>, **P**<sub>2</sub>,  $\sim E$ , and **L**  $\in \mathbf{K}_L^+$ .

**P**<sub>1</sub> is summarized in Fisher's Table III (ibid., p. 128). Fisher considers Mendel's bifactorial experiments, which test Mendel's 1:2:1 law. For each of the seven characteristics Mendel first crossed pure-bred dominant plants (**AA**)

with recessives (aa), to create a generation (F1) of hybrids. These were self-fertilized to create an F2 generation, where Mendel expected (according to Fisher) a 1 (AA):2(Aa):1 (aa) ratio. But how can one tell the difference between an AA and Aa plant, since both show the same dominant phenotype? Mendel's method was to self-fertilize "suspect" plants and consider their progeny. If all of the latter show the dominant phenotype, the original plant is considered AA; if at least one recessive plant appears, the original must be Aa. This method risks misclassifying Aa plants as AA, since some heterozygotes might produce only dominant-phenotype offspring purely by chance. In the case of Mendel's first two properties (seed form and color), the chance of such misclassification is negligible, since each plant produced at least 30 seeds; the probability of all of them showing the dominant phenotype when the original plant was a heterozygote is at most  $0.75^{30} \sim 0.00018$ .<sup>2</sup> However, Mendel (according to Fisher) used a less reliable method in the case of the other five properties. In those cases Mendel grew only ten plants from each of the "suspect" plants. About 5.6% ( $0.75^{10}$ ) of the heterozygotes would produce ten dominant-looking plants by pure chance, and thus be misclassified as homozygotes.<sup>3</sup>

The result is that Mendel should have classified the 600 F2 plants grown to test the genotype of the other five properties at a 1.056:1.944:1 proportion; yet the reported results fit the expected 1:2:1 ratio almost perfectly. They are very close to what Mendel *wrongly* expected the results to be, while being quite far from the results Mendel should have gotten. Similar considerations apply to Mendel's method of classifying certain plants used in his trifactorial experiments (Mendel, 1866, §8). With the 5.6% correction, Mendel should have classified about 175 of 473 dominant-phenotype plants as homozygous, yet actually Mendel so classified only 152: close to the uncorrected value of 157.7. As Fisher puts it, 'a total deviation of the magnitude observed, and in the right direction [towards the outcome Mendel wrongly expected], is only to be expected once in 444 trials; there is therefore here a serious discrepancy' (Fisher, 1936, p. 129).

$P_1$  is enough to cause Fisher to give up  $\sim BE$  (and therefore also to give up  $\sim E$ ): the belief that there is no bias error in Mendel's experiments, since by 'bias error' I mean error due to conscious or unconscious fudging of the results, at any stage of the experiment, to fit better with the outcome the experimenter expects. Bias error can occur in all stages of an experiment, from unconsciously choosing

plants that "fit better" with the theory when recording the results, to deliberate falsification or suppression of the records after the fact. Fisher becomes open to the possibility that such error exists; his corpus of belief becomes the new, smaller,  $((K_L^+)_{\sim BE})^-$ .

Fisher, however, does not claim that any particular type of bias error occurred. He considers from this new corpus various types of bias error that could explain the results, such as Mendel 'changing his previous practice and ... backcrossing the 473 doubtful plants' (ibid.); that the 'selection of plants for testing favored the heterozygotes ... unconsciously' (ibid.); or the 'tendency to give the theory the benefit of the doubt when objects such as seeds, which may be deformed or discolored by a variety of causes, are being classified' (ibid., pp. 130–131); and, finally, 'the possibility that the data ... do not represent objective counts, but are the product of some process of sophistication' (ibid., p. 129)—that is, that the results were "cooked".<sup>4</sup> The latter is not only bias error, but also *deliberate* bias error: deliberate cheating.

Fisher checks these possibilities one by one and dismisses all as insufficient, except for the "cooked data" possibility. To test it, Fisher checks *all* of Mendel's data, and discovers, to his shock, that  $P_2$  is the case, summarized in Fisher's Table v (ibid., p. 131): Mendel's results taken as a whole are so close to his expectations, that  $\chi^2$  tests show the probability of repeating the experiments and exceeding the deviation from the reported results is 0.999 93. These two statistical results,  $P_1$  and  $P_2$ , taken together, settle it for Fisher: he becomes convinced not only that there *might* be some bias error in the experiment, but that there *is* bias error in the experiment and, in particular, that it is due to deliberate cheating (DC), though Fisher does not say by whom. His final corpus is  $((K_L^+)_{\sim BE})_{DC}^+$ : in effect, adding L to the corpus caused Fisher to replace  $\sim BE$  (and  $\sim E$ ) with DC.

## 5. Are the changes justified? Rational belief change theory and Fisher

Three criticisms of this interpretation can be made. First of all, how do we know that Fisher gives up  $\sim BE$  in response to discovering  $P_1$ , when he does not explicitly say so? Second, it is tempting to suggest that Fisher actually did not first give up  $\sim BE$  and then add DC to the corpus, but rather that he underwent a "conversion" from believing Mendel's results are error-free to believing the

<sup>2</sup> It should be noted that here, and in other experiments, Mendel is susceptible to another source of error. Mendel tests whether the peas' fertilization is independent, but the method for distinguishing AA from Aa plants by looking at their progeny is itself *based* on this very assumption. If, for example, Aa plants tended to create fewer aa plants than expected through self-fertilization, the number of Aa plants incorrectly classified as AA plants would be even larger. Still, the effect would be small, for any deviation in the number of aa plants created that is large enough to materially effect the proportion of Aa plants mistakenly identified as AA ones would be evident in the original proportion of aa plants.

<sup>3</sup> Seidenfeld (1998) and others disagree with this reading, claiming that Fisher's experimental setup has many further controls Fisher ignored, which make the proportion of mistakenly identified plants insignificant. But even if true this does not affect the second part of Fisher's criticism, following.

<sup>4</sup> In Hon's typology, the first of these is an error in the *reporting* of the data (Mendel neglecting to mention the change in his methods); the second is an error in the *background assumptions* (that the selection does not favor heterozygotes over homozygotes); the third and fourth are errors in the *recording* of the data (conscious or unconscious). Note that the first of these is not actually a bias error as I define it; we shall come back to this point later.

results are cooked: that he *replaced*  $\sim\text{BE}$  with **DC** in one step. Finally, there is the question of what justification Fisher had for giving up  $\sim\text{BE}$  in the first place. In the answer to all of these objections we can see that Fisher is implicitly committed to important basic principles of rational belief change theory.

The evidence that  $\sim\text{BE}$  was given up is clear. Had  $\sim\text{BE}$  not been given up, Fisher would not have bothered to investigate types of bias error as potential explanations for the puzzling data. In addition, if Fisher seriously considered giving up *more* than  $\sim\text{BE}$ —that is, had he seriously considered the possibility that the puzzling results were due to some other type of error (such as nonrandom pollenization)—he would have considered whether or not such errors could be potential explanations of Mendel's results. But apart from raising—and immediately dismissing—the possibility of Mendel 'changing his previous practice and ... backcrossing the 473 doubtful plants', which is never taken seriously, all the other possible explanations of Mendel's puzzling results that Fisher considers are some type of bias error, from unconsciously selecting plants in ways that "fit" the theory or consciously manipulating results to so do.

Also, it is clear that  $\sim\text{BE}$  was given up first, and only then replaced with **DC**, instead of this occurring "in one fell swoop". Fisher looks at the various forms of bias error and determines in each *whether or not* each one could explain the surprising results of Mendel's experiments (ibid., pp. 129–130). It is not, a priori, necessarily the case that some sort of bias error would succeed in explaining  $\text{P}_1$  &  $\text{P}_2$ . It is a logical possibility that *all* possible bias errors considered would be found wanting. Indeed, even before looking at  $\text{P}_2$ , Fisher already dismisses all other bias error possibilities apart from **DC**. And there is nothing to suggest that just because **DC** is the last type of bias error considered, it must a priori be accepted as correct.

It could have been—logically—that the testing of all of Mendel's data, which discovered  $\text{P}_2$  is the case, would not have shown anything remarkable, which would lower the likelihood of deliberate data "cooking". Or it could be that even if  $\text{P}_2$  is the case, that there would be some other reason to reject the possibility of deliberate cheating. Indeed, Fisher himself rejected Bateson's view that Mendel's results are a didactic summary—surely the easiest way to explain (or explain away) Mendel's "too good to be true" results—by appealing to the fact that Bateson's hypothesis must be wrong for other reasons, no matter how puzzling and in need of further explanation the data look if one rejects Bateson's view.

In rejecting  $\sim\text{BE}$  without automatically embracing **DC** (or anything else that implies **BE**). Fisher complies with two basic desiderata of rational belief change, as seen in Levi's and Peirce's formulations. First, one should not,

says Peirce, 'put roadblocks in the path of inquiry' (Peirce, 1955, p. 54) and, just because **BE** should be considered, jump to the conclusion that it *had* occurred. One should rather first become open about whether or not **BE** occurred by removing  $\sim\text{BE}$ , and conclude that **BE** itself—or the stronger **DC**, which implies **BE**—is the case only if there is some further good reason for this. From this smaller corpus, Fisher indeed discovered that some sort of bias error is the best explanation for the puzzling experimental result. But it also might be that Fisher would add back  $\sim\text{BE}$  after all, and return to the "original"  $\text{K}_L^+$ : this might have occurred if none of the bias errors considered "worked out" (as may have been the case if conclusive proof surfaces that cheating could not have happened, for example). Or he might conclude there is not enough justification to accept *either* that there was a bias error in the experiment *or* that there was no such error, and remain open between the two possibilities—remain, that is, in the corpus  $(\text{K}_L^+)_{\sim\text{BE}}$ . One cannot, a priori, decide which one is the case purely from general principles of belief change.<sup>5</sup>

Second, as Levi (1980, 1988, and elsewhere) notes, from Fisher's current point of view,  $\text{K}_L^+$ , **BE** is surely false. One must never knowingly add a falsehood to one's corpus, for this violates the basic rational belief change desideratum to avoid error. On the other hand, merely giving up  $\sim\text{BE}$  involves no risk of error: only of loss of information. If, later, Fisher adds **BE** (or the stronger **DC**) to his views from the new corpus,  $(\text{K}_L^+)_{\sim\text{BE}}$ , then it would not be a violation of a desideratum of inquiry since, from that new corpus, **BE** is *not* certainly false.

The next question, from the point of view of rational belief change theory, is how does Fisher justify giving up  $\sim\text{BE}$  in the first place. Assuming  $\text{L} \& \text{P}_1 \& \text{P}_2 \& (\sim\text{E})$ , it only follows (logically) that Mendel's results are highly improbable. Fisher has not actually reached a logical contradiction, and low-probability events are by no means necessarily problematic. For example, the mere fact that somebody won the lottery—a priori an event with a far lower probability than Mendel's results—does not require a special explanation, nor does it require one to conclude that there must have been some other reason than dumb luck for it. It could be that Mendel *was* simply lucky; why look for another explanation at all? In rational belief change theory terms, a basic desideratum of rational belief change theory is that one should not give up information without good reason, and mere low probability is not enough.

Fisher, it seems to me, is quite aware that some good reason for giving up  $\sim\text{BE}$  is needed; and, second, his (implicit) reason is not a mere psychological hunch, or obeying a statistical test, but *epistemic*, and in line with the recommendations of rational belief change theory. To understand Fisher's justification, we need to go back to his views as he expressed them in 1925:

<sup>5</sup> For example, the answer crucially depends on whether Fisher believes he had considered *all* possible types of bias error. If he isn't sure that he had considered all of them, he might tend to remain in the "open" corpus. If he is certain that he had considered all possible types of bias error, on the other hand, he might conclude that he had shown  $\sim\text{BE}$  must be the case after all.

The term Goodness of Fit has caused some to fall into the fallacy of believing that the higher the value of  $P$  the more satisfactorily is the hypothesis verified. Values over .999 have sometimes been reported which, if the hypothesis were true, would only occur once in a thousand trials. Generally such cases have proved to be due to the use of inaccurate formulæ, but occasionally small values of  $\chi^2$  beyond the expected range do occur, as in Ex. 4 with the colony numbers obtained in the plating method of bacterial counting. In these cases the hypothesis considered is as definitely disproved as if  $P$  had been .001. (Fisher, 1925, p. 80)

It seems, *prima facie*, as if Fisher is applying in 1936 a “two-tailed” Neyman–Pearson-like (N–P) test for accepting or rejecting hypotheses, which he seems to have been advocating in 1925. That is, that  $H_0$  (Mendel’s hypotheses) should be rejected not only if  $p$  of the experiment made to verify it is very low, but also if  $p$  is very high; in the latter case at least, some error should be suspected in the experiment as well.

However, this clearly could not be the case. For one thing Fisher’s writings predate the N–P tests. For another, Fisher’s statistical analysis usually does not deal with accepting or rejecting one of two alternative hypotheses, as the N–P method does, but only with whether one can have confidence in a single hypothesis. Fisher never offers  $H_1 =$  ‘There is a bias error in the experiment’ (or any other specific hypothesis that offers a reduced variance: see Edwards, 1972, §9.3) as an alternative (which would be impossible in any case since  $H_0$  and  $H_1$  are neither exclusive nor exhaustive). Finally, Fisher most definitely does *not* reject Mendel’s hypotheses about genetics, of course, despite the high  $p$  values.<sup>6</sup>

So what does Fisher mean? Consider why a *low*  $p$  value disproves a hypothesis. The obvious answer is that a low  $p$  value ‘strongly indicate[s] that the hypothesis fails to account for the whole of the facts’ (Fisher, 1925, p. 79). It shows there is likely to be some other cause than the hypothesis for the observed results. *But the same is true*, Fisher says, *in the case with exceptionally high  $p$  values*. He gives us two such possible causes: ‘inaccurate formulæ’—that is, the type of experimental error Hon (1989, p. 191) would call an ‘error of interpretation’; and (implicitly) bias or cheating by the observers, which skews the results towards what they wanted to get. So a high  $p$  value is, typically, a good reason to suspect that some other cause than the hypothesis will give us a better explanation of the data (Fisher, 1925).

This is just what Fisher (1936) applies in practice. The high  $p$  value means that the possibility of bias error should be seriously considered, since a bias error of some sort would better explain such a  $p$  value than mere luck on Mendel’s part. In Levi’s terminology (Levi, 1988), by accepting  $P_1$ ,  $P_2$ ,  $\sim E$ , and  $L$  Fisher can at most explain how it is possible (though highly unlikely) that Mendel could have reached just those results merely by chance. But assuming bias error would allow Fisher to explain not only why Mendel’s results are possible, but also why they are not surprising: if Mendel had, consciously or unconsciously, skewed the results towards his expectations, it is no longer puzzling they match so closely.

What motivates Fisher, then, to consider the possibility of bias error is not undue fear of the mere improbability of Mendel’s results. Nor is it the application of an N–P-like statistical test that *forces* him to reject  $\sim BE$  and accept that bias error existed if  $p$  is high enough (as we saw, Fisher considers it possible that none of the bias error explanations would work and  $\sim BE$  would be added back after all). It is that the high  $p$  value is a good reason to suspect that a better explanation—namely, the possibility of bias error—of Mendel’s results might (though not necessarily) be the case.

This hope of a better explanation of puzzling phenomena is very often the reason that scientists give up old theories for new, at least for the sake of the argument. This is an essentially epistemic, not statistical or logical, justification for belief change: it is based not on a statistical test’s recommendations of accepting or rejecting an hypothesis, nor on the fact that one reached a logical contradiction, but on the rational belief change theory’s desideratum to find informative, explanatory theories.

## 6. Where Fisher disagrees with the theory of belief change

Fisher, then, accepts many basic desiderata of rational belief change theory. Nevertheless, I argue, Fisher “jumped to conclusions”: there was another possible source of error he ignored. A significant cause of Fisher ignoring the other source of error is that he was using a “rule of thumb” for rational belief change that is *usually* accurate, but not in this case.

Fisher did, as we saw, have good reason to suspect that  $BE$  is a possible culprit. But in this case, Fisher also had good reason to think that there is another possible problem, due to the conditions of the peas. He was suspicious of the low number of peas per plant reported on average

<sup>6</sup> Fisher (1935, Ch. 10, §62) knew a statistical test could be interpreted as the data rejecting what a set of hypotheses ( $H_0$  possibly included) says about the mean or variance of the population. A low  $\chi^2$  (high  $p$ ) arouses suspicion, not because the data rejects  $H_0$ ’s *mean*, but because the  $\chi^2$  test shows it rejects  $H_0$ ’s *variance*. Fisher is not offering a “two-tailed” test concerning the mean, as Pilgrim (1984) and others argue. Crucially, ‘all “rejecting the null hypothesis” means’ is that ‘we tend to look for another hypothesis’ (Edwards, 1972, §9.3), because either a “rare chance” has occurred, or the null hypothesis is wrong (Fisher, 1956, p. 39). This is *not* the same as being *forced* to give up  $H_0$  and replace it with some  $H_1$ . The data rejecting  $H_0$  might *suggest* an alternative  $H_1$  and giving up  $H_0$  to investigate  $H_1$  from a neutral position; but one might then add back  $H_0$ , or remain in doubt between  $H_0$  and  $H_1$ , as well as accept  $H_1$ .

in Mendel's work: suspicious enough, in fact, to ask Dr Rasmussen (an expert on *Pisum*) if Mendel's numbers are possible (Fisher, 1936, pp. 122–123). Furthermore, statistical analysis of at least some repeats of Mendel's experiments from the early 1900s, available to Fisher, show the "too good to be true" results in similarly low-yielding peas (Seidenfeld, 1998).

With this information available to him, Fisher might well have come to suspect it is possible that the small number of peas indicates *some* basic error is afoot in the experimental setup, which makes our underlying assumption about the statistical distribution of the peas suspect.<sup>7</sup> He should have also considered removing  $\sim\text{AE}$ , 'there isn't another (i.e., non-bias) type of error', becoming open about it, in order to consider whether or not some such error could explain  $\text{P}_1\&\text{P}_2$ . Fisher had good reason to remove  $\sim\text{BE}$  and to remove  $\sim\text{AE}$ . In such a case, rational belief change theory recommends *giving up both*. Had Fisher done so, he would probably not have concluded Mendel was biased or misled: it would remain a possibility, even if the only possible type of bias error is deliberate falsification, that the "too good to be true" results might not be due to bias error at all, but rather to the peas' conditions.

What is the decision-theoretic justification for such a course? Again, it deals with basic principles: in this case, Peirce's dictum to 'not put roadblocks in the path of inquiry'. If it becomes necessary to remove something from one's corpus, and there are two reasonable ways to do it, both of which remove roughly the same amount of information, then one should suspend judgment and become open about both instead of arbitrarily choosing to remove one or the other. The reason is that arbitrarily keeping one of the two "blocks inquiry" in the sense of being arbitrary, of giving artificial preference to keeping one bit of information over another.

A more technical explanation is due to Levi (1991, 2004). As we saw, information is a desideratum of inquiry and should not be given up without good cause. But if one is presystematically committed to *always* removing the least amount of information possible, absurd results occur. Consider  $\mathbf{J}$  = 'a fair coin was tossed', and  $\mathbf{H}$  = 'it landed "heads"'. Suppose an agent accepts both. This means she also accepts, as a matter of logic,  $\mathbf{J}\rightarrow\mathbf{H}$ . Suppose further that she removes that the coin was tossed; then she must also obviously remove the belief that the coin landed "heads". But while  $\mathbf{J}$  and  $\mathbf{H}$  are removed, if she retains the maximal amount of information possible, she should keep  $\mathbf{J}\rightarrow\mathbf{H}$ ; but then, if she adds back that the coin *was* tossed, later on, she must also add back that it landed "heads". This seems clearly wrong.

What made Fisher go wrong? Fisher's disagreement with Levi and Peirce in this particular case is not due to his lack of concern for rational belief change. Indeed, what

Fisher did wrong here shows that he usually did what was right from the point of view of rational belief change theory. In this case, it is reasonable for Fisher to come to doubt both  $\sim\text{BE}$  and  $\sim\text{AE}$  since there is good reason to suspect both. But in many experiments where the results are suspiciously close to expectations, there is no particular reason to suspect that any other error apart from experimenter's bias is involved. In such cases, removing  $\sim\text{BE}$  while retaining  $\sim\text{AE}$ —considering only the possibility of bias error—is, indeed, a good idea as a first step in inquiry. Fisher used a good rule of thumb in a situation where it is not applicable.

## 7. Conclusion

Fisher, it seems, was at least implicitly aware of the desiderata of rational belief change theory and followed its recommendations quite closely in the actual belief changes he went through in his famous 1936 paper. In particular, he was following his rule from 1925 that a high  $p$  value is as good a reason for rejecting a hypothesis as a low  $p$  value. If my interpretation is correct, the rule in Fisher (1925) and its application in Fisher (1936) should not be seen as a formal statistical test of the N–P type, but as an epistemic rule of thumb as to when there is a good reason to consider the possibility of bias error in an experiment and give up  $\sim\text{BE}$ .

The problem with the 1936 paper, on this view, is that Fisher might have "jumped to conclusions": i.e., he might not have given up *enough*, since in this particular case there was also good reason to suspect *other* error might have been possible. This, if correct, shows that statistical analysis, such as  $\chi^2$  tests, is not in itself sufficient to determine what type of error exists in an experiment. At most it can give us reason to suspect *some sort* of error exists. It is only through carefully considering different possible *types* of error that Fisher could have come to the conclusion that one of the possible types of error to consider in this case is a non-deliberate flaw in the experiment's design.

It is true that often, as Fisher's rule of thumb shows, statistical analysis showing a very high  $p$  value should make one suspect bias error; but it is not *necessarily* true. There often are other possible sources of error that cause similar outcomes, such as the low yield of the peas in this case, or the 'incorrect formulae', which Fisher notes (Fisher, 1925, p. 80). And it just *might* turn out that there is no error at all, that the high  $p$  value was just dumb luck. This lends credence to the view held by Hon (1989) and others that a typology of experimental error has value over and above the statistical analysis of error. The latter can show that some error exists, and perhaps what it is likely to be; but the former plays a crucial role as well.

<sup>7</sup> Seidenfeld (1998) suggests that in such conditions, fertilization is 'non-independent' in a way that drastically lowers its variance.

## Acknowledgements

My thanks to Giora Hon for reading earlier drafts of this paper and making many helpful comments, and to an anonymous reviewer for helpful criticism.

## References

- Bateson, W. (1909). *Mendel's principles of heredity*. Cambridge: Cambridge University Press.
- Box, J. F. (1978). *R. A. Fisher: The life of a scientist*. New York: Wiley & Sons.
- Correns, C. G. (1900). Mendels Regel über das Verhalten der Nachkommenschaft der Rassenbastarde. *Berichte der deutschen botanischen Gesellschaft*, 18, 158–168.
- Edwards, A. W. F. (1972). *Likelihood*. Cambridge: Cambridge University Press.
- Fairbanks, D. J., & Rytting, B. (2001). Mendelian controversies: A botanical and historical review. *American Journal of Botany*, 88(5), 737–752.
- Farrall, L. A. (1975). Controversy and conflict in science: A case study—The English biometric school and Mendel's laws. *Social Studies of Science*, 5, 269–301.
- Fisher, R. A. (1925). *Statistical methods for research workers*. London: Oliver and Boyd.
- Fisher, R. A. (1930). *The genetical theory of natural selection*. Oxford: Clarendon Press.
- Fisher, R. A. (1935). *The design of experiments*. Edinburgh: Oliver and Boyd.
- Fisher, R. A. (1936). Has Mendel's work been rediscovered? *Annals of Science*, 1, 115–137.
- Fisher, R. A. (1956). *Statistical methods and scientific inference*. Edinburgh: Oliver and Boyd.
- Froggatt, P., & Levin, N. C. (1971). The 'law of ancestral heredity' and the Mendelian-ancestral controversy in England, 1886–1906. *Journal of Medical Genetics*, 8, 1–36.
- Gardner, M. (1977). Great fakes of science. *Esquire*, 87(10), 88–92.
- Haldane, J. B. S. (1924). A mathematical theory of natural and artificial selection, Part I. *Proceedings of the Cambridge Philological Society*, 23, 19–41.
- Haldane, J. B. S. (1926). A mathematical theory of natural and artificial selection, Part IV. *Proceedings of the Cambridge Philological Society*, 23, 607–615.
- Haldane, J. B. S. (1932). *The causes of evolution*. London: Longman.
- Hon, G. (1989). Towards a typology of experimental errors: An epistemological view. *Studies in History and Philosophy of Science*, 20, 469–504.
- Levi, I. (1980). *The enterprise of knowledge: An essay on knowledge, credal probability, and chance*. Cambridge: MIT Press.
- Levi, I. (1988). Four themes in statistical explanation. In W. L. Harper, & B. Skyrms (Eds.), *Causation in decision, belief change, and statistics* (2 vols.) (Vol. 2, pp. 195–222). Dordrecht: Kluwer.
- Levi, I. (1991). *The fixation of belief and its undoing: Changing beliefs through inquiry*. New York: Cambridge University Press.
- Levi, I. (2004). *Mild contraction: Evaluating loss of information due to loss of belief*. New York: Oxford University Press.
- Mayo, D. (1996). *Error and the growth of experimental knowledge*. Chicago: University of Chicago Press.
- Mendel, G. (1866). Versuche über Pflanzen-Hybriden. *Verhandlungen des naturforschenden Vereines, Abhandlungen, Brünn*, 4, 3–47.
- Nissani, M. (1994). Psychological, historical, and ethical reflections on the Mendelian paradox. *Perspectives in Biology and Medicine*, 37, 182–196.
- Olby, R. C. (1997). Mendel, Mendelism, and genetics. In *Mendelweb* (ed. 97.1). <http://www.mendelweb.org/MWolby.html>. (Accessed 31 October 2006)
- 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 could be reasonably supposed to have arisen from random sampling. *Philosophical Magazine, Series 5*, 50(302), 157–176.
- Peirce, C. S. (1955). *Philosophical writings of Peirce* (J. Buchler, Ed.). New York: Dover Publications.
- Pilgrim, I. (1984). The too-good-to-be-true paradox and Gregor Mendel. *Journal of Heredity*, 75, 501–502.
- Sapp, J. (1990). The nine lives of Gregor Mendel. In H. E. Le Grand (Ed.), *Experimental inquiries* (pp. 137–166). Dordrecht: Kluwer.
- Seidenfeld, T. (1998). P's in a pod: Some recipes for cooking Mendel's data. In *PhilSci Archive*. <http://philsci-archive.pitt.edu/archive/00000156/>. (Accessed 31 October 2006)
- Sturtevant, A. H. (1965). *A history of genetics*. New York, NY: Harper and Row.
- Tschermak, E. von (1900). Über künstliche Kreuzung bei *Pisum sativum*. *Berichte der deutschen botanischen Gesellschaft*, 18, 232–239.
- Vries, H. de (1900). Das Spaltungsgesetz der Bastarde. Vorläufige Mitteilung. *Berichte der deutschen botanischen Gesellschaft*, 18, 83–90.
- Weiling, F. (1986). What about R. A. Fisher's statement of the 'too good' data of J. G. Mendel's *Pisum* paper? *Journal of Heredity*, 77, 281–283.
- Weiling, F. (1989). Which points are incorrect in R.A. Fisher's statistical conclusion: Mendel's experimental data agree too closely with his expectations? *Angewandte Botanik*, 63, 129–143.
- Weiling, F. (1991). Historical study: Johann Gregor Mendel 1822–1884. *American Journal of Medical Genetics*, 40, 1–25.
- Weldon, W. R. F. (1902). Mendel's laws of alternative inheritance in peas. *Biometrika*, 1, 228–254.
- Wright, S. (1931). Evolution in Mendelian populations. *Genetics*, 16, 97–159.
- Wright, S. (1966). Mendel's ratios. In Stern, C., & Sherwood, E. R. (Eds.), *The origins of genetics: A Mendel source book* (pp. 173–175). San Francisco: W. H. Freeman.