Chapter 21. “A source of variation for which we cannot account”

The extreme skeptic pays no attention to the simple common sense, he knows that when the presumed order of the world seems threatened, the reason grants him an unlimited overdraft.”

B. Méheust[1]

Nature having successively killed the article of 1988 and the article written in association with A. Spira – mainly on statistical arguments – one could suppose that the article of S. Hirst *et al* was indisputable on this point. Of course, one would be curious to know the comments and remarks of the experts who reviewed this article. Unfortunately, we did not get the pleasure of reading them.

However, although the raw data were not released by the authors, we can find rather easily the experimental outcomes that allowed the statistical analysis, through the graphs of the article. Furthermore, a statistical report of J. Burridge from the department of Statistics was drafted in March 1992 and noticeably enlightens these results. [2, 3]

It is thus by confronting the text of the article, all results deducted from figures and the statistical report of J. Burridge that we are going to be analyzing this article and showing that this latter could be a textbook case.

The experimental protocol

Thirty-six working sessions were performed by Hirst *et al* for the 3 types of dilutions (Table 21.1). One session was performed in one day.[4] The first experiments began on early June 1990. The samples of blood were obtained from 11 blood donors who could participate in more than one session: five participated once and one participated nine times. Basophils were counted by a single “trained” experimenter.

	Type A Dilutions of Anti-IgE (shaken)	*Type B* Dilutions of anti-IgE (not shaken)	*Type C* Dilutions of solvant (shaken)
1 control + 1 anti-IgE dilution 10^2 + 8 high dilutions $1/10^{12}$, $1/10^{14}$ … $1/10^{26}$ = 30 counts	5 sessions	4 sessions	3 sessions
1 control + 1 anti-IgE dilution 10^2 + 8 high dilutions $1/10^{30}$, $1/10^{32}$ … $1/10^{44}$ = 30 counts.	5 sessions	4 sessions	3 sessions
1 control + 1 anti-IgE dilution 10^2 + 8 high dilutions $1/10^{46}$, $1/10^{48}$ … $1/10^{60}$ = 30 counts.	5 sessions	4 sessions	3 sessions

Table 21.1. Summary of the plan of experiments of Hirst *et al.* There were 36 sessions. Each session corresponds to one working day with the preparation of cells, preparation of the series of dilutions, incubation of cells with dilutions and finally counting of basophils. Each session was dedicated to the study of a series of 1/100-dilutions for one of the types of dilution (antiserum anti-IgE diluted with shaking, antiserum anti-IgE diluted without shaking, solvent diluted with shaking) and for one of the 3 ranges of dilutions (3 ranges were defined between $1/10^{12}$ and $1/10^{40}$).

Each session included 30 counts of basophiles. The report of J. Burridge gives an example for a session "diluted and shaken anti-IgE" (Figure 21.1). The continuous lines between tubes were intended to show how tubes were "connected" by the successive dilutions. Consequently, to each dilution (of anti-IgE or solvent), 3 counts of basophils corresponded. It is the means of these triple counts that are presented in the figures of the article.

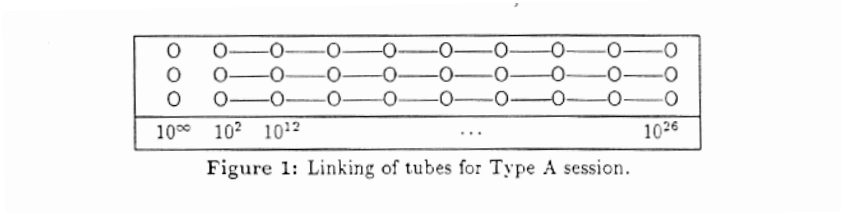

Figure 1: Linking of tubes for Type A session.

Figure 21.1. This figure extracted from the report of J. Burridge corresponds to an experiment of Type A (diluted and shaken anti-IgE) and of "rank 1" (i.e. high dilutions from $1/10^{12}$ to $1/10^{26}$). The sessions of Type B were identical except that tubes were not shaken between each dilution. For the sessions of type C, *the only tubes that contained dilutions of anti-IgE were 3 tubes at dilution $1/10^2$*; the 27 other tubes of the sessions of Type C contained the solvent serially diluted in solvent with shaking between each dilution.

A textbook case

Before going farther, we have to overcome the hurdle of Figure 1. Indeed, the first figure, which the reader saw on the first page after having read the title and possibly the lead paragraph, is reproduced in Figure 21.2.

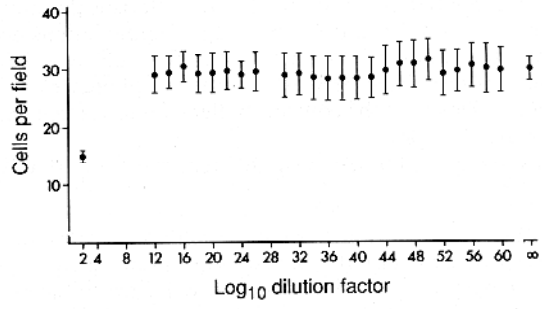

FIG. 1 Mean cell densities (with standard errors) as a function of dilution for all the data. The data for succussed anti-IgE, unsuccussed anti-IgE and succussed buffer have been combined. For the buffer control (10^{∞}) and anti-IgE dilution 10^{2}, $n=108$; for the other anti-IgE dilutions, $n=36$.

Figure 21.2. Comment: *succussed* = shaken according to the homeopathic wording that names "succession" the shaking of solution between each dilution.

At first sight, this first figure of the article was coherent with the title because a quick examination seemed to indicate that high dilutions hardly modified the counts of basophils. However, an attentive reading of the text and the figure legend showed that in fact the results of the three series of data were – contrary to usual and good scientific practices – presented by their *common means*! The legend of the figure indeed specified that the data "have been combined"!

But, this way of "presenting" results did not apply to anti-IgE with a classic dose (which is the point on the extreme left of the graph: $1/10^{2}$)! Indeed the latter was present in all sessions (whatever the type of session: A, B or C). The first impression was thus strengthened since the number of basophils in the presence of anti-IgE with classic dose ($1/10^{2}$), an active one, had considerably decreased. The reader who was accustomed to "normal" presentations of scientific graphs was deceived because he supposed that classic doses of anti-

IgE at high dilution of anti-IgE *were presented on an equal footing, which was not the case.*

It is useless to impute motives, but if one did not want to highlight a "signal" by flooding it in the "background noise", one would not have done so otherwise. [5]

Some very different clouds

Let us resume the results of the article which are represented in the form of the means of the 3 percentages of degranulation corresponding to each dilution (see figure 21.3).

To each dilution of anti-IgE (or solvent), there were 3 counts of basophils (= a triplicate). The article reported the averages of these triplicates. For 30 sessions, there were thus 8 × 30 = 240 percentages of degranulation with high dilutions. If there was no difference between a "control" well and a "high dilution" well, we should expect that the percentages of degranulation fluctuated around 0%, because, according to this hypothesis, high dilutions were supposed to have no effect. In statistics, this is known as the null hypothesis.

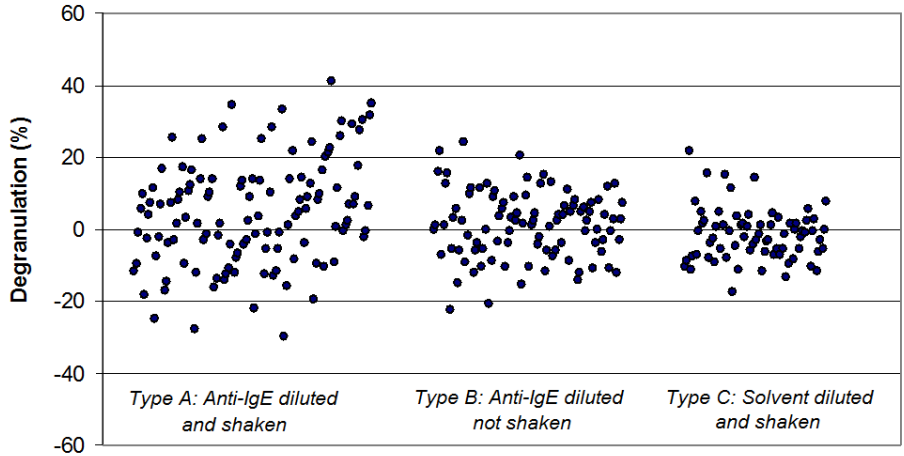

Figure 21.3. Degranulation corresponding to each experimental point was obtained from Figure 2 of the article of Hirst *et al* (1993). Each point is the mean of 3 experimental points with high dilution. We notice that the 3 "clouds" have very different shapes. This suggests that the type of high dilutions had an influence on the counts of basophils.

Nevertheless, even a non-experienced eye notices that the 3 series of experiments (diluted and shaken anti-IgE, diluted and not shaken anti-IgE and

diluted and shaken solvent) did not identically behaved (Figure 21.3). The percentages of degranulation of the diluted and shaken solvent seemed less scattered. In contrast, the cloud of points of diluted and shaken anti-IgE was very scattered and it seemed to contain more positive than negative percentages. The cloud of points corresponding to diluted and not shaken anti-IgE was in an intermediate situation. The calculation of the averages and the standard deviations confirms this impression:

Diluted and shaken anti-IgE:	4.5 ± 14.7 %
Diluted and not shaken anti-IgE:	1.1 ± 9.3 %
Diluted and shaken solvent:	-1.7 ± 7.1 %

Let us build the distributions of the results

As regards the evidence of an effect with high dilutions, the A and C series (diluted and shaken anti-IgE and diluted and shaken solvent) are enough for this analysis (indeed, series B tested the necessity of shaking tubes to obtain an effect). Let us classify each of the points: degranulation from 0 to 9%; 10 to 19%; 20 to 29%, etc. and let us count how many points belong to each class. We then obtain the distributions of Figure 21.4.

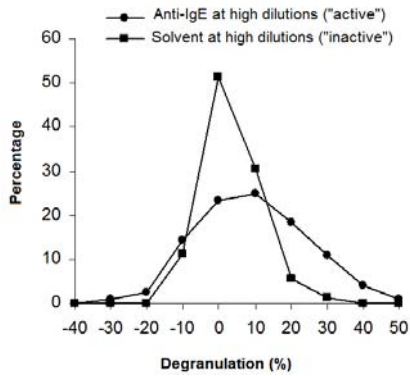

Figure 21.4. This figure is built from the results of Figure 21.3 for diluted and agitated anti-IgE (type A) and for diluted and agitated solvent (type C). The frequency of the counts is calculated for every class of percentage (every point on the *x*-axis corresponds to the upper limit of the interval). One clearly highlights here (whatever was the cause) a difference of the distribution of counts according to high dilutions of anti-IgE (supposed to be "active") or high dilutions of solvent (supposed to be "inactive") prepared in the same conditions. Nevertheless, the authors of the article refused to envisage the possibility that these differences were an argument in favour of an effect of high dilutions of anti-IgE.

We thus notice that the two populations are different. We observe a shift of the distribution towards higher dégranulations when basophils were in the presence of high dilutions of anti-IgE. A statistical analysis shows that both populations are significantly different. In other words, it looks like an effect of high dilutions of anti-IgE!

Furthermore, thanks to the series of diluted but not shaken anti-IgE (type B; not represented on the Figure 21.4), the authors showed that high dilutions obtained with shaking were more active than high dilutions, which were performed without shaking! It is difficult to completely homogenize a solution by not shaking it, what could explain the small "degranulating" activity – although statistically not significant – that seems to have been passed on during the process of "dilution without shaking".

The three series of measurements thus appeared to be very different and it was indeed what the statistical calculations of Hirst *et al* indicated! Indeed, the authors reported this table that summarized the p-values (statistical significance) after an analysis of variance according to the type of treatment (Table 21.2).

Table 6: ANOVA Tests*(*p*-values) For Differences Between Dilutions Separately For Each Session

Treatment Type	High Dilution Range		
	$10^{12} - 10^{26}$	$10^{30} - 10^{44}$	$10^{46} - 10^{60}$
A (combined "Fisher" *p*-value = 0.0027)	0.34	0.028	0.34
	0.066	0.018	0.21
	0.14	0.42	0.17
	0.0043‡	0.42	0.21
	0.70†	0.80	0.40
B (combined *p*-value = 0.086)	0.56	0.25	0.27
	0.27	0.97	0.27
	0.0073**	0.76	0.16
	0.084	0.48	0.44
C (combined *p*-value=0.85)	0.92	0.80	0.66
	0.25	0.25	0.21
	0.65	0.91	0.72

* ANOVA tests used the one-way F-test with (8, 18) degrees of freedom, using control and high dilution treatments in each session, ie the 10^2 dilutions were excluded.

† The missing value in this session means that the F-test used (8, 17) degrees of freedom.

‡ See Fig. 6.

**See Fig. 7.

Table 21.2. This table comes from the report of J. Burridge. It was reproduced in the article of Hirst *et al* without major change. There are the same data in Table 21.3 in a simplified and commented version.

Treatment	*p*-value (statistical significance calculated by J. Burridge)	Interpretation (not in the article)
Treatment Type A (anti-IgE diluted and shaken)	0.0027	Very significant
Treatment Type B (anti-IgE diluted but not shaken)	0.086	Not significant (trend)
Treatment Type C (solvent diluted and shaken)	0.85	Not significant

Table 21.3. Simplified and interpreted version of Table 21.2. The statistical tests (variance analysis) performed by J. Burridge indicated (whatever was the cause) that high dilutions of anti-IgE did not have the same effect – with a high statistical significance – compared to control dilutions performed in the same conditions. Interestingly, one notes that agitation appears necessary to observe an effect with high dilution. In spite of these results, Hirst *et al* concluded that the significance of these tests was probably the result of a "statistical artifact".

In other words, the percentages of degranulation were not null for high dilutions of anti-IgE. In contrast, high dilutions of solvent were not significantly different from 0. If there was no experimental bias, this indicates that cells did not have the same behavior in the presence of high dilution of anti-IgE or in the presence of a control.

In the statistical report, J. Burridge commented on these results:

> "According to conventional scientific theory there should, within a session, be no differences between the control treatment and the eight high dilutions "treatments". […] Such an hypothesis can be tested, separately for each session, by applying the conventional ANOVA F-test to the mean counts for each tube […]. The resulting p-values are given in Table 6. These results are curious. They should, if the null hypothesis is correct, to be uniformely distributed between 0 and 1. This does not seem to be the case for treatment A and B for which the p-values are collectively too small."

J. Burridge even considered a possible effect of high dilutions!:

> "Table 6 and Figure 6 and 7 suggest that triples differ from each other. The reasons for this are at present obscure. One interpretation is that there are, after all, differences between treatments – for some sessions and subjects at least."

Let the reader relish this "after all". Having quickly pushed aside this hypothesis which seemed obviously unthinkable to him – but what his statistical analysis was nevertheless supposed to assess! - J. Burridge pursued:

> "The most plausible explanation of these effects is that some as yet unidentified feature of the experimental procedure tends to make triples differ from each other in some random or haphazard way. It is possible that the serial dilution procedure is responsible for this effect – although it is hard to see how."

If we summarize the method of J. Burridge, our reason forbids us to consider the possibility of an effect with high dilutions, therefore another cause exists – but an "unidentified" one – related to the experimental procedure! Once again, the spirit of Descartes paradoxically crossed the Channel: "And the demonstrations are so certain that, even if experience seemed to show us the contrary, we would nevertheless be obliged to place more faith in our reason than in our senses." [6]

The criticisms (not published) of the statistician J. Burridge towards the article, of which he was co-author

Obviously, the expertise of the statistician J. Burridge was used *after* the results were completed. What seems certain is that he did not participate to the design of the experimental protocol. He repeatedly complains in his report about defects in the design of the protocol thus leading to a delicate analysis. His main criticisms were the following ones:

1) No randomization between sessions:

> "[…] there are some features which make analysis somewhat awkward and others which make the interpretation problematic at times. For example, no attempt appears to have been made either to randomise the time order of the sessions or to balance the ranges within each treatment type with respect to volunteers. Thus the type A sessions were done first, the type B followed by type C. Similarly, within each type, range 1 sessions were done first, then range 2 followed by range 3. The allocation of volunteers to sessions was unavoidably haphazard. The lack of randomisation of the order of the sessions is an inconvenient feature of the experiment and mean that certain "treatment" effects could be attributable to trends over time (due in particular, perhaps, to learning effects acquired by the experimenters during the course of the experiment".

2) The "links" between the dilutions were "broken":

> "[…] for most sessions the tubes were "linked" by the serial dilution procedure. Such linking means that the results for successive dilutions might be serially correlated within a series of nine dilutions so that ideally these series should be analysed as single entities with their own mean and covariance structure. However, the linking was not recorded during the subsequent randomisation procedure and so cannot be properly accounted for in the statistical analysis."

Yet, concerning this second point, one of the results highlighted in the article is that there was no periodicity for degranulations according to dilutions. Nevertheless, J. Burridge implicitly recognized that the authors of the article did not give themselves the means to analyze this precise point. Indeed, when tubes were blinded, it has been not noted to which of the three serial dilutions these tubes belonged (see the "links" between the dilutions for each of the three series of dilution on Figure 21.1). It is quite possible that the authors of the article did not think that they would need to analyze the periodicity because they did not envisage that a significant global effect would be observed. It is likely that the differences of effect of the various treatments surprised the authors of the article and that they tried "to sweep them under the carpet" by insisting on the absence of degranulation "waves" as described in the article of *Nature* in 1988.

How did the authors of the article overcome this difficulty to nevertheless "demonstrate" – in spite of their results – that high dilutions of anti-IgE were without effect?

A "statistical artefact"…

Of course, these criticisms of J. Burridge about the issues of methodology were not reported in the article. Even though the authors correctly summarized the description about the different behaviors of the 3 treatments (corresponding above to "According to conventional scientific theory… the p-values are collectively too small"), they never considered the possibility that the observed effect could be due to high dilutions of anti-IgE. To explain these unexpected differences between treatments, the authors introduced a new and curious notion: a "statistical artifact":

> "Although it is possible that these observed effects are a statistical artefact, some unidentified part of our experimental procedure might account for them. It is an interesting feature of our data but it does not, of course, lend any support to the findings of Davenas et al., and serves once again to underscore the complexity of the analysis of variance in an assay of this type."

The very original notion of "statistical artifact" played the role of a wild card. As unexpected results were obtained, the "statistical artifact" allowed refuting them without explanation. Furthermore, the authors seemed to complain that everything could not be checked, that the experimental protocol could play tricks on them and that these analyses were so complicated…

But all these considerations did not prevent the authors to impudently conclude the article with this sentence:

> "We have been unable to find any evidence that very high dilutions of anti-IgE, succussed or unsucussed, cause any reproducible effect on the degranulation of human basophil leukocytes."

This conclusion was thus in disagreement with the reported results. If the authors were aware that their experimental protocol contained weaknesses and was not thus capable of answering the question that was asked, they were free to explain it (by reporting the criticisms of the report of J. Burridge on which the article is tongue-tied). But would their results then have deserved to be published in the pages of *Nature*? In the end, it is the strategy "everything but the results of Benveniste" that has prevailed.

The arguments of J. Benveniste and A. Spira

It was on a unique column that J. Benveniste and A. Spira gave their comments in *Nature* and, as already said, eight long months after the publication of the article.[7] They noted fifteen differences between the methodology of Hirst *et al* and the one of the article of *Nature* of 1988 but, given the limited space, they only pointed out the most important differences. Thus, Hirst *et al* included in their analysis all experiments, including those for whom the percentage of degranulation with anti-IgE at a classic dose was low. J. Benveniste explained again that, if one did not observe a degranulation with anti-IgE at classic doses, there was little chance to observe an effect with high dilutions of anti-IgE.

Another criticism expressed by J. Benveniste and A. Spira was about assessing in separated experiments the various treatments that one wished to compare. It was the most importing reproach because it is at the heart of the reasoning in experimental biology to vary only one factor at the same time. The correct procedure would have been to compare the high dilutions of anti-IgE and high dilutions of the solvent in the same session. Finally, a step of centrifugation had been added after cell incubation with high dilutions. This procedure might have increased the variability of the counts.

Secondly, J. Benveniste and A. Spira questionned the statistical method and pointed out the "tactics" used to mask the statistically significant differences.

Finally, they insisted that the authors refused to communicate the raw data of the experiments.

The exercise of the criticism was however delicate. One could not indeed blame the authors for not having respected the original experimental protocol and asserting at the same time that the results proved nevertheless the existence of an effect of high dilutions.

In the press, J. Benveniste expressed himself in a more direct way by asserting that Hirst *et al* "committed several methodological and ethical errors".[8] For the famous "Figure 1" in particular, he considered that it was "a manipulation unprecedented in the history of science (combination of results of active and control samples)."

Then, he continued:

> "Also unethical […] is the fact that I was not approached for the adjustment of the numerous details necessary for the good practice of so complex experiments and that I learnt the existence of this article only by the press. It is extremely surprising to see a journal as *Nature*, which portrays itself as an archetype of the excellence and of the scientific integrity, be engaged in such a manipulation. The question is: what are the real motives?"

More technical, A. Spira declared:

> "All in all […] I do not think that these results are contradictory with ours and I think that it would be necessary that we can exchange our raw data so as to compare the results of both series of experiments by using the same strategy of statistical analysis".

Exchanging data, comparing results, is not this called doing research? But was it the concern of the authors and those who promoted the publication of this article? [9]

Comparison with the results of the article of the Comptes Rendus de l'Académie des Sciences

The comparison of these results with those of the study done with the collaboration of the team of A. Spira and published in the *Comptes Rendus de l'Académie des Sciences* is quite unexpected. Indeed, the presentation of the results as distributions of percentages of degranulation leads to similar profiles (Figure 21.5).

It is extremely strange to notice that results, which in fact are very similar, were interpreted with opposite conclusions. In both cases, high dilutions had a behavior different from that of control dilutions and the "sinusoidal" curves

were not present. We remember that J. Benveniste had said: "once again we obtain the results published in *Nature*. It is the same girl, as beautiful as ever. She only lacks a bit of makeup." [10] But for Hirst *et al* this absence of "makeup" is a decisive argument to state: "We can find no evidence for any periodic or polynomial change of degranulation as a function of anti-IgE dilution" and that consequently in no case one could assert that the results of J. Benveniste had been confirmed. [11]

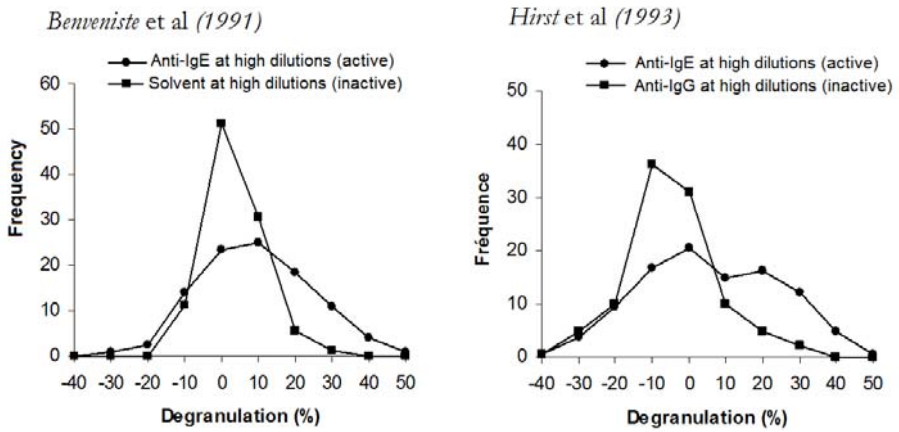

Figure 21.5. Comparison of the results of the article of J. Benveniste and A. Spira [12] of 1991 in the *Comptes Rendus de l'Académie des Sciences* and those of Hirst *et al* of 1993 in *Nature*. Very close results were obtained but the conclusions of the authors were diametrically opposed for a possible effect of anti-IgE at high dilutions.
(Each point on the *x*-axis corresponds to the upper limit of the interval).

J. Maddox and the "Popperian spirit"

To finish on the relationship of J. Benveniste and the journal *Nature*, we will conclude on a thought, in the form of a prediction, that J. Maddox had inserted into the long final comment of four pages that he wrote (not cosigned by the other investigators) in the issue of *Nature* of October 27th, 1988.

Indeed, in September 1988, J. Maddox wrote to the Israeli, Italian and Canadian teams to ask them if they wished to comment on the events of the previous months.[13] The team of Toronto answered in particular that it pursued the research on high dilutions in a "Popperian spirit". J. Maddox added in his final text of October 27th that he would be "glad to publish as Scientific Correspondance the general conclusion of any or all of these groups when they are ready." [14] Then, about J. Benveniste and his team, he formed the wish that

this latter "will now counting basophils in replicate, following the standard procedure for controlling sampling errors, and will be eliminating unavoidable observer bias by making blind experiment a routine".

Finally, he concluded:

> "I expect that these results will not differ substantially from those obtained in the three blind experiments (each with two observers) at Clamart on 9 and 10 July (*sic*)[15]; it will be extremely interesting if it should be otherwise, but no doubt that Dr Benveniste would prefer to publish in some other journal."

The first prediction of J. Maddox did not come true. Indeed, during the series of experiments done in association with A. Spira and his team an effect associated with high dilutions was observed in much more rigorous experimental conditions than for the experiments quoted by J. Maddox.

The second prediction did not come true either. Indeed, as we saw above, the manuscript reporting these experiments has been indeed proposed to *Nature*. We have noticed how J. Maddox again disqualified these results. We saw how in contrast the article of Hirst *et al* crossed the barrier of the experts apparently without too many difficulties despite legitimate questions that should have been raised if the same criteria had been applied as for the article of J. Benveniste and A. Spira.

The Popperian spirit consists in questioning and in testing one's own convictions in the light of the experiment. It seems that for J. Maddox the Popperian spirit applies to all scientists, except however for the director of *Nature*.

Notes of end of chapter

[1] B. Méheust. Somnambulisme et médiumnité, tome II, *Les Empêcheurs de Penser en Rond / Synthélabo*. 1998.

[2] J. Burridge. A repeat of the "Benveniste" statistical analysis. Department of Statistical Science. University College London. Research Report N° 100, March 1992.

[3] This report was kindly communicated to me in 2001 after simple request to the secretariat of the Department of Statistics of *Royal College of London*.

[4] The experimenters cut the experiments according to this design most probably because they considered that an experiment including 30 wells was sufficient for one working day. It is a pity that they did not benefit from any advice and assistance of W. Stewart who had a more Stakhanovite conception of the counting of basophils…

[5] A possible explanation of this statistical "approach" could be that the variance of the counts would be increased if some "active" dilutions modified the numbers of the basophils in comparison with the "inactive" dilutions. It remains that this approach is quite unusual.

[6] R. Descartes. Principes de la philosophie (1644).

[7] J. Benveniste, B. Ducot and A. Spira. Memory of water revisited. *Nature*, Auguts 4th, 1994, p. 322.

[8] F. Nouchi. Une équipe de chercheurs anglais n'a pu reproduire les travaux du docteur Benveniste sur la « mémoire de l'eau ». *Le Monde*, 11 décembre 1993.

[9] Nevertheless, it seems that the manuscript required some improvements since it was received by the journal on April 16th, 1993 and accepted on October 22nd, 1993.

[10] M. de Pracontal. Les mystères de la mémoire de l'eau, p. 200.

[11] This incoherence between the results and the conclusions of the authors was nevertheless noticed by some scientists. In particular independent analyses from Italo Vecchi as well as those of Jean-Pierre Pharabod could be read on Internet shortly after the publication of the article of Hirst *et al.* These analyses led to the same conclusions, namely that the null hypothesis (i.e. no difference between controls and "active" samples) must be rejected; consequently the results contradicted the title of this article.

[12] For the results of J. Benveniste and A. Spira, the percentages of degranulation within every experiment are calculated with the mean of the counts of highly diluted anti-IgG.

[13] The Israeli team answered in two waves. First of all, B. Robinzon explained that it was difficult for him to comment on an investigation which he did not attend (letter of September 18th, 1988 to J. Maddox). Then, while commenting on several technical points, he reaffirmed his conviction that "a biological phenomenon was observed, not an artifact, although it is one for which there is no explanation". He concluded: "I did not comment until now […] since I do not think that it is for us to prove if the phenomenon we observed was real or an artifact, especially after your report discredited us". He asked however to J. Maddox not to publish his letter. J. Amara and M. Oberbaum answered after at lenghth by resuming several technical points and they complained too about the treatment by the British journal of the works on the high

dilutions which according to them "is not such as befits a journal of *Nature*'s calibre regarding scientific work" (letter of J. Amara and M. Oberbaum to J. Maddox of December 11th, 1988).

The Italian team answered at the beginning of October to J. Maddox and blamed the "unfair" treatment of its contribution.

[14] J. Maddox. Waves caused by extreme dilution. *Nature*, 27 octobre 1988, p. 760.

[15] The investigators left on Friday, July 8th, 1988; J. Maddox meant about Thursday, 7th and of Friday, 8th.