PATRICK SUPPES

FROM THEORY TO EXPERIMENT AND BACK AGAIN

In this article I consider two substantive examples of the way in which there is continuing interaction in science between theory and experiment. The picture of theory often presented by philosophers of science is too austere, abstract and self-contained. In particular, the picture of theory that is painted is much too removed from the shock effects of new experiments. Perhaps even more to the point, in many parts of science the actual formulation of theory is much driven by the latest experiments.

The first example comes from scientific research I am currently doing on language and the brain. I begin by describing the work in broad terms. I then present the response of new experiments and new theoretical statistical analysis of the data to answer claims that the recognition rates for brain-wave representation of words and sentences is not significant, because of the large amount of information available. Here the use of the concept of an extreme statistic is used to answer this criticism in a detailed way. Discussion of this example will end with some brief remarks on how this use of more detailed statistical methods is now generating new experiments, and having an impact on the design of the experiments.

The second example deals with experiments and physical theory on the entanglement of particles, and the consequent nonlocality of standard quantum mechanics. After some general remarks on this area of research in quantum mechanics and its philosophical importance for our basic physical concepts, I turn to the theoretical work of Greenberg, Horne and Zeilinger and their proposed "GHZ-type" experiments.

First the purely theoretical result, formulated in probability-one terms, is stated. Then the question is asked, how can such probability-one theoretical results be tested, given the inevitable inefficiences of particle detectors.

This prompts a new theoretical effort to derive inequalities, like those of Bell for other experiments, to deal with GHZ-type experiments. What comes out of the analysis is that better experimental results should be achievable with very careful design and use of current photon detectors. But the proof

M C. Galavotti (ed), Observation and Experiment in the Natural and Social Sciences, 1–41

of this is rather detailed and relies on theory in critical ways at several points. These examples are but current illustrations, but the lesson is meant to be universal. The continual interaction between theory and experiment occurs in nearly every developed branch of science.

LANGUAGE AND THE BRAIN[1]

Some historical background

Aristotle said that the distinguishing feature of man as an animal is that he is a rational animal, but, in more biological and psychological terms, it is that of being a talking animal. Language is, in ways that we have not yet fully explored, the most distinguishing mark of man as an animal. Its processing is centered, above all, in the brain, not just for the production of speech, but for the intentional formation of what is to be said or for the comprehension of what has been heard or read. So it is the brain's processing of language that is the focus of this section. I begin with a historical sketch of the discovery of electrical activity in the brain.

An early reference to electricity being generated by muscles or nerves of animals comes from a study by Francesco Redi (1671), who describes in this way an experiment he conducted in 1666: "It appeared to me as if the painful action of the *torpedine* (electric ray) was located in these two sickle-shaped bodies, or muscles, more than in any other part." Redi's work was done in Florence under the Medici's. These electrical observations were fragmentary and undeveloped. But the idea of electrical activity in the muscles or nerves of various animals became current throughout the eighteenth century (Whittaker 1951, Galvani 1791). Yet it was more than 100 years after Redi before the decisive step was taken in Bologna by Luigi Galvani. He describes his first steps in the following manner:

> The course of the work has progressed in the following way. I dissected a frog and prepared it ... Having in mind other things, I placed the frog on the same table as an electric machine. When one of my assistants by chance lightly applied the point of a scalpel to the inner crural nerves of the frog, suddenly all the muscles of the limbs were seen so to contract that they appeared to have fallen into violent tonic convulsions. Another assistant who was present when we were performing electrical experiments thought he observed that this phenomenon occurred when a spark was discharged from the conductor of the electrical machine. Marvelling at this, he immediately brought the unusual phenomenon to my attention when I was completely engrossed and contemplating other things. Hereupon I became extremely enthusiastic and eager to repeat the experiment so as to clarify the obscure phenomenon and make it known I myself, therefore, applied the point of the scalpel first to one then to the other crural

[1]This section is taken from my forthcoming book *Representation and Invariance in Scientific Structures,* Stanford, CA: CSLI Publications.

nerve, while at the same time some one of the assistants produced a spark; the phenomenon repeated itself in precisely the same manner as before.

(Galvani 1791/1953, pp. 45–46)

Galvani's work of 1791 was vigorously criticized by the well-known Italian physicist Alessandro Volta (1745–1827), who was born in Como and was a professor of physics at the University of Pavia. Here are his words of criticism, excerpted from a letter by Volta to Tiberius Cavallo, read at the Royal Society of London:

The name of animal electricity is by no means proper, in the sense intended by Galvani, and by others; namely, that the electric fluid becomes unbalanced in the animal organs, and by their own proper force, by some particular action of the vital powers. No, this is a mere artificial electricity, induced by an external cause, that is, excited originally in a manner hitherto unknown, by the connexion of metals with any kind of wet substance. And the animal organs, the nerves and the muscles, are merely passive, though easily thrown into action whenever, by being in the circuit of the electric current, produced in the manner already mentioned, they are attacked and stimulated by it, particularly the nerves.

(Volta 1793/1918, pp. 203–208)

Galvani was able to meet these criticisms directly and in 1794 published anonymously a response containing the detailed account of an experiment on muscular contraction without the use of metals (Galvani 1794). The original and important nature of Galvani's work came to be recognized throughout Europe. The prominent German physicist Emil Du Bois-Reymond (1848) summarized in the following way Galvani's contribution:

1. Animals have an electricity peculiar to themselves, which is called Animal Electricity.
2. The organs to which this animal electricity has the greatest affinity, and in which it is distributed, are the nerves, and the most important organ of its secretion is the brain.
3. The inner substance of the nerve is specialized for conducting electricity, while the outer oily layer prevents its dispersal, and permits its accumulation.
4. The receivers of the animal electricity are the muscles, and they are like a Leyden jar, negative on the outside and positive on the inside.
5. The mechanism of motion consists in the discharge of the muscular fluid from the inside of the muscle via the nerve to the outside, and this discharge of the muscular Leyden jar furnishes an electrical stimulus to the irritable muscle fibres, which therefore contract.

(Du Bois-Reymond 1848/1936, p. 159)

A next event of importance was the demonstration by Carlo Matteucci (1844) that electrical currents originate in muscle tissue. It was, however,

almost 100 years after Galvani, that Richard Caton (1875) of Liverpool detected electrical activity in an exposed rabbit brain, using the Thomson (Lord Kelvin) reflecting telegraphic galvanometer. In 1890, Adolf Beck of Poland detected regular electrical patterns in the cerebral cortex of dogs and rabbits. Beginning at the end of the nineteenth century Villem Einthoven, a Dutch physician and physiologist, developed a new electrocardiograph machine, based on his previous invention of what is called the string galvanometer, which was similar to the device developed to measure telegraphic signals coming across transatlantic cables. Using Einthoven's string galvanometer, significant because of its sensitivity, in 1914, Napoleon Cyblusky and S. Jelenska Macieszyna, of the University of Cracow in Poland, recorded a dog's epileptic seizures. Beginning about 1910, Hans Berger in Jena, Germany began an extensive series of studies that detected electrical activity through intact skulls. This had the great significance of being applicable to humans. His observations were published in 1929, but little recognized. Recognition came, however, when his findings were confirmed by Edward Douglas Adrian and B. H. C. Matthews of the University of Cambridge, who demonstrated Berger's findings at the Physiological Society in Cambridge in 1934, and the International Congress of Psychology in 1937. In the late 1930s and the early 1940s research on electrical activity in brains, or what we now call electroencephalography (EEG), moved primarily to North America—W. G. Lennox and Erna and F. A. Gibbs at the Harvard Medical School, H. H. Jasper and Donald Linsley at Brown University, and Wilder Penfield at McGill University. One of the first English-language reports to verify Berger's work was by Jasper and Carmichael (1935). Nearly at the same time, Gibbs et al. (1935) began using the first ink-writing telegraphic recorder for EEG in the United States, built by Garceau and Davis (1935). By the 1950s, EEG was widely used clinically, especially for the study of epilepsy, and for a variety of research on the nature of the electrical activity in the brain. This is not the place to summarize in any serious detail the work by a wide variety of scientists from 1950 to the present, but an excellent review of EEG, that is, of electrical activity, pertinent especially to cognition, is to be found in Rugg and Coles (1995).

Observing the brain's activity

The four main current methods of observing the brain are easy to describe. The first is the classical electroencephalographic (EEG) observations already mentioned, which, and this is important, have a time resolution of at least one millisecond. The second is the modern observation of the magnetic field rather than the electric field, which goes under the title of magnetoen-

cephalography (MEG). This also has the same time resolution of approximately one millisecond. The third is positron emission tomography (PET), which has been widely used in the last several decades and is good for observing location, in some cases, of brain activity, but has a time resolution of only one second. Finally, the most popular current method is functional magnetic resonance imaging (fMRI), which does an excellent job of observing absorption of energy in well-localized places in the brain, but unfortunately, also has a time resolution of no less than a second.

Although many excellent things can be learned from PET and fMRI, they are not really useful if one wants to identify brain waves representing words or sentences, for the processing, although slow by modern computer standards, is much too fast to be able to accomplish anything with the time resolution of observation no better than one second. The typical word, for example, whether listened to or read, will be processed in not more than 4 or 5 hundred milliseconds, and often faster. My own serious interest, focused on the way the brain processes language, began from the stimulus I received by hearing a brilliant lecture in 1996 on MEG by Sam Williamson, a physicist who has been prominent from the beginning in the development of MEG. I was skeptical about what he said, but the more I thought about it, the more I realized it would be interesting and important to try using MEG to recognize the processing of individual words. This idea suggested a program of brain-wave recognition, as recorded by MEG, similar in spirit to speech recognition. I was familiar with the long history of speech recognition from the 1940s to the present, and I thought maybe the same intense analytical effort could yield something like corresponding results. So, in 1996, assisted especially by Zhong-Lin Lu, who had just taken a Ph.D. with Sam Williamson and Lloyd Kaufman at New York University, we conducted an MEG experiment at the Scripps Institute of Research in San Diego, California. When we proceeded to analyze the results of the first experiment, the problem of recognizing which one of seven words was being processed on the basis of either having heard the word or having read it on a computer screen, we were not able to get very good recognition results from the MEG recordings. Fortunately, it was a practice at the Scripps MEG facility, which is technically very much more expensive and complicated to run than standard EEG equipment, to also record the standard 20 EEG sensors used for many years. We proceeded to analyze the EEG data as well, and here we had much better recognition results (Suppes, Lu and Han 1997).

In the standard EEG system, widely used throughout the world for observing electrical activity in the brain, sensors to record the electrical activity are arranged in what are commonly called the 10-20 system, as shown in Figure 1, with the location on the surface of the skull of the head shown in the

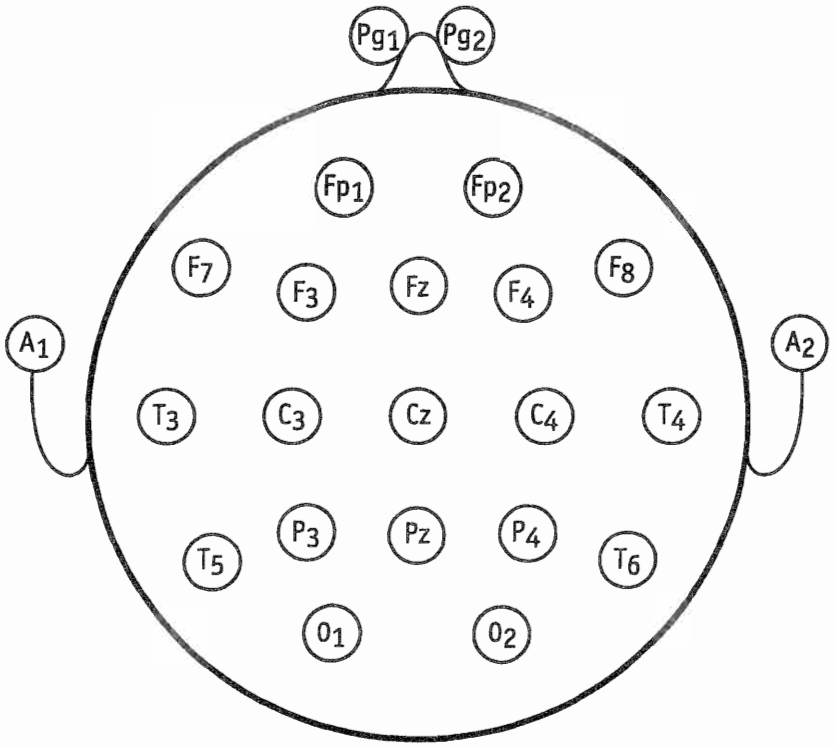

Figure 1· The 10-20 system of EEG sensors.

approximate form of a circle, with ears left and right and eyes at the top in the figure. The first letters of the initials used in the locations correspond to references to the location in the part of the brain, for example, F for frontal, C for center, T for temporal, P for parietal, O for occipital. Second, you will note that the odd-numbered sensors are located on the skull on top of the left hemisphere and the even-numbered sensors are located over the right hemisphere, with three sensors located approximately along the center line. There are more opinions than deeply articulated and established facts about what takes place in the left hemisphere or in the right hemisphere, possibly both, in the processing of language. My own view is that there is probably more duality than has been generally recognized, but I will not try to make an empirical defense of that view in the present context, although I have published data supporting duality (Suppes, Han and Lu 1998, Table 2).

Figure 2 shows a typical trial, in which the subject was given a visual, i.e., printed, sentence, one word at a time, on a computer screen. The trial

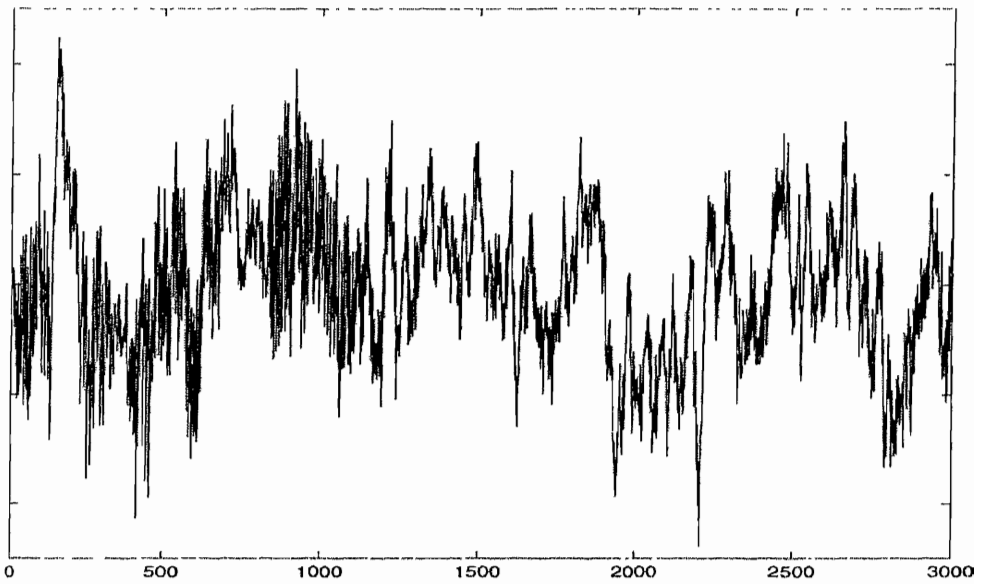

Figure 2: Unanalyzed EEG data from one trial and one sensor.

lasted for over 3000 milliseconds, with an observation of an amplitude of the observed wave plotted on the y-coordinate in microvolts every millisecond. Given that so much data are observed in a little over three seconds, just from one sensor, out of 20, it is easy to see that EEG recordings of language activity are rich in data and, in fact, we might almost say, swamped by data. It is not difficult to run an experiment with several subjects, each with a number of hours of recording, and have at the end between five and ten gigabytes of data of the kind to be seen in Figure 2. This means that the problem of trying to find waves corresponding to particular words and sentences is not going to be a simple matter. It is very different from behavioral experiments dealing with language, in which the observed responses of subjects are easily recorded in a few megabytes and then analyzed without anything like the same amount of computation.

Methods of data analysis

There is no royal road to finding words and sentences in the kind of data just described, so I will describe here the approach that I and my colleagues have used with some success in the past few years. The basic approach is very much taken from digital signal processing, but the application is a very

different one from what is ordinarily the focus of electrical engineers or others using the extensive mathematical, quantitative and statistical techniques that have been developed in the last 50 years in digital signal processing. An excellent general reference is Oppenheim and Schafer (1975).

The approach is easily sketched, although the technical details are more complicated and will only be outlined. Generally, but with important exceptions to be noted, the first thing to do is to average the data, for example, for a given condition, a typical case being the brain response to a particular verbal stimulus, given either auditorily or visually. The purpose of this averaging is to eliminate noise, especially high-frequency noise, on the assumption that the signal is of much lower frequency. The next step is then to perform a Fourier transform on the averaged data, passing from the time domain to the frequency domain, perhaps the most characteristic feature of signal processing analysis. The third step, which can be done simultaneously with the second step, is to filter in the frequency domain, to reduce still further the bandwidth of the signal used for identifying the word or sentence being processed. An alternative, which we have explored rather thoroughly in some work, is to select in the frequency domain the frequencies with the highest energies, as measured by the absolute value of their amplitudes, and then to superpose the sine functions in the time domain. (Actually, in "selecting a frequency ω_i with amplitude A_i", we are in fact selecting a sine wave $A_i \sin(\omega_i t + \varphi_i)$, where φ_i is its phase. More on this on the next page.) In either case, by filtering or superposition, we get a much simpler signal as we pass by an inverse Fourier transform back to the time domain.

Speaking now of filtering, and ignoring for the moment the superposition approach, we go back to the time domain with a bandpass filter fixed by two parameters, the low frequency L and the high frequency H of the filter. We now select two more parameters that are of importance. Namely, what should we take to be the beginning s and the ending e of the signals for the words or sentences in a given experimental condition. As in other brain matters, it is not at a glance obvious from the recorded brain wave when the representation of a particular word or sentence begins or ends in the continually active electrical waves that are observed. When a quadruple (L, H, s, e) is selected, we then use that quadruple of parameters to make a classification of the brain waves of the words or sentences that are the stimuli in a given experimental condition.

Our aim is to optimize or maximize the number of brain waves correctly classified. We keep varying the quadruple of parameters until we get what seems to be the best result that can be found. We are doing this in a four-dimensional grid of parameters, but I show in Figure 3 the optimal surface for two parameters of the filter, although here what is used as a measure on

the ordinate or y-axis is the difference between the high-frequency and low-frequency filter rather than the high filter itself. So we have on the abscissa the low filter measured in Hz and on the ordinate the difference W, which is the width in Hz of the bandpass filter. The smoothness of the surface shown in Figure 3 is characteristic of what we always observe and would be expected, of observations of an electrical field outside the skull, the place where we are observing them. (We are of course very fortunate that the electrical fields are strong enough to be observed without complete contamination by noise.) This isocontour map is for the first experiment with 48 geographic sentences discussed in more detail below, but the map shows clearly that the best recognition rate was 43 of the 48 sentences (approximately 90%). As the parameters of the bandpass filter are changed, the contour map shows definitely lower recognition rates.

Fourier analysis of EEG data

In the background of the Fourier analysis is the standard theory of the Fourier integral, but in practice our data are finite. The finite impulse data that we are interested in observing usually last for no more than a few seconds. For example, the sentences studied will ordinarily last not more than three or four seconds when spoken at a natural rate. To analyze frequencies with given amplitudes and phases, we use the discrete Fourier transform. As indicated, our goal is to find the frequencies that contain the signal and eliminate the noise. (The artifacts generated by eye blinks or other such events are discussed at the end of this section.)

Let N be the number of observations, equally spaced in time, usually one millisecond apart. We then represent the finite sequence of observations $x(n), 0 \leq n \leq N-1$ by Fourier series coefficients $\tilde{X}(k)$ as a periodic sequence of period N, so we have the dual pair

$$\tilde{X}(k) = \sum_{n=0}^{N-1} \tilde{x}(n) e^{-i(\frac{2\pi}{N})kn} \tag{1}$$

$$\tilde{x}(n) = \frac{1}{N} \sum_{k=0}^{N-1} \tilde{X}(k) e^{i(\frac{2\pi}{N})kn} \tag{2}$$

I first note the following:

1. The periodic sines and cosines are represented by the standard exponential terms.

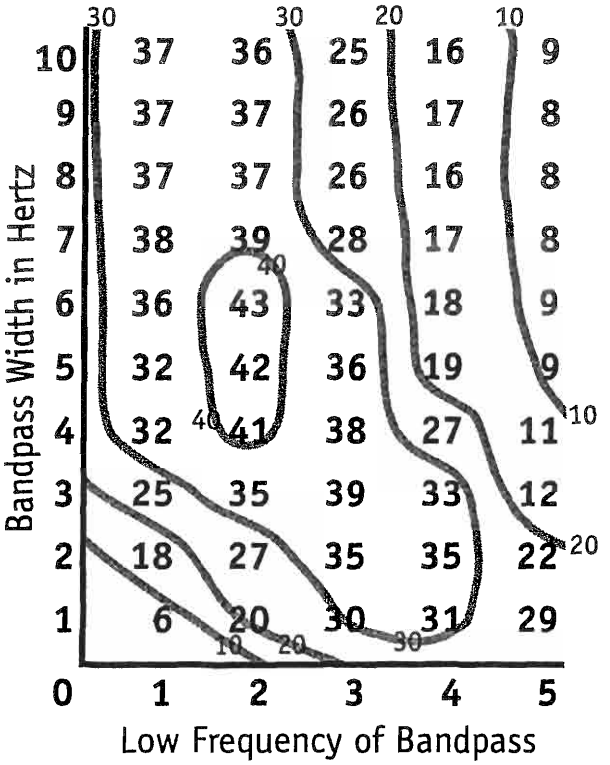

Figure 3: Typical contour map of recognition rate surface for bandpass-filter parameters L and W.

2. $\tilde{x}(n) = x(n) = \tilde{x}(n + kN)$, the tilde shows periodicity of length N and is for duality of time and frequency.

3. The kn part of the exponent gives us distinct exponentials, and thus sine and cosine terms for integer submultiples of the period N. This way we get in the representation frequencies that are an integer multiple of $\frac{2\pi}{N}$.

4. Using the periodicity N gets us duality between the time and frequency domains.

The properties of the two equations (1) and (2) are:

1. Linearity: If $\tilde{x}_1(n)$ and $\tilde{x}_2(n)$ have period N, so does

$$\tilde{x}_3(n) = a\tilde{x}_1(n) + b\tilde{x}_2(n)$$

and

$$\tilde{X}_3(k) = a\tilde{X}_1(k) + b\tilde{X}_2(k).$$

2. Invariance under shift of a sequence

$$n \to n + m$$

3. Various symmetry properties, e.g., $|\tilde{X}(k)| = |\tilde{X}(-k)|$.

4. Convolution of $\tilde{x}_1$ and $\tilde{x}_2$ of period N has period N:

$$\tilde{x}_3(n) = \sum_{m=0}^{N-1} \tilde{x}_1(m)\tilde{x}_2(n - m).$$

Of importance is the efficient fast discrete Fourier transform, an algorithm due to Cooley and Tukey (1965) and others, a variant of which was used in the computations reported below.

Filters. The principle of filter construction is simple. Details are not. A bandpass filter, e.g., 1-20 Hz simply "filters all the frequencies below 1 Hz and above 20 Hz." There are many developments in the electrical engineering literature on the theory and art of designing filters, which it is not possible to survey here. The important point is always to design a filter with some criterion of optimality.

If the signal is known, then the engineering objective is to optimize its transmission. Our problem, as already mentioned, is that, in our experiments, the signal carrying the word or sentence in question is unknown. So our

solution is to optimize the filter to predict the correct classification. The parameters we used have been discussed above. In addition we often make a smoothing correction around the edges of the filter by using a 4th-order Butterworth filter, although in the work reported here, something simpler would serve the purpose just about as well.

Three experimental results

I turn now to three of the most important results we have obtained so far.

Invariance between subjects

In the first experiment, we presented 48 sentences about the geography of Europe to 9 subjects. The subjects were asked to judge the truth or falsity of the sentences, and while they were either listening to or reading the sentences displayed one word at a time on a computer screen, we made the typical EEG recordings. The semantic task was simple, but because the sentences were separated by only four seconds, the task of judging their truth or falsity was not trivial. Typical sentences were of the form *The capital of Italy is not Paris,* and *Warsaw is not the largest city in Austria.* Taking now the data from five subjects to form prototypes of the 48 sentences, by averaging the data from the five subjects, and taking the other four subjects to form corresponding averaged test samples of each sentence, we applied the Fourier methods described above and found an optimal bandpass filter from a predictive standpoint. (The data are for the visual condition of reading the sentences, one displayed word at a time.) We were able to recognize correctly 90% of the test samples, using as a criterion for selection a classical least-squares fit between a test sample and each of the 48 prototypes, after filtering (Suppes, Han, Epelboim and Lu 1999a).

Let $x_i(n), 0 \leq n \leq N-1$, be the ith prototype (in the time domain), and $y_j(n), 0 \leq n \leq N-1$, the jth test sample. Then the sum of squared differences is S_{ij}, where

$$S_{ij} = \sum_{n=0}^{N-1} (x_i(n) - y_j(n))^2.$$

The test sample $y_j(n)$ is correctly classified if

$$S_{jj} = \min_i S_{ij},$$

with the minimum being unique.

The surprising invariance result is that the data for prototypes and for test samples came from different subjects. There was no overlap in the two groups. Theoretically this is an efficient aspect of any much used communication system. My brain-wave representation of words and sentences is much like yours, so it is easy to understand you. But it is a theoretical point that needs strong empirical support to have it accepted. Another angle of comment is that the electric activity in the cortex is more invariant across subjects performing the same task than is the detailed anatomical geometry of their brains. I return to this invariance between subjects a little later, when I respond to some skeptical comments.

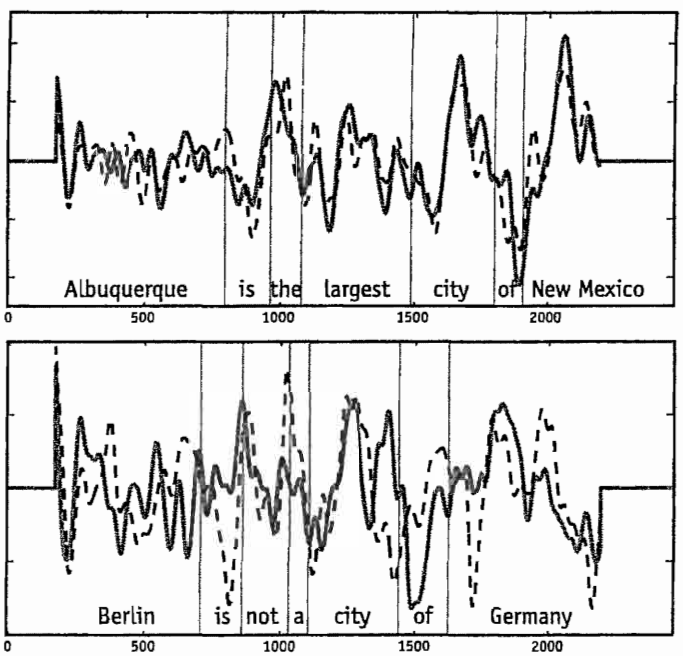

Figure 4: Prototypes (grey lines) and test samples (dashed black lines) generated by the best fitting sentence (upper panel) and worst fitting sentence (lower panel) correctly classified for subject S32. Time measurements after the onset of the sentence are shown in milliseconds on the abscissa.

One hundred sentences

I now turn to a second more recent experiment in which subjects were visually presented 100 different geography sentences (Suppes, Wong, et al., to

appear). I concentrate here only on the remarkable result of correct recognition of 93 of the 100 sentences for one subject (S32). Using the methods described, the best recognition rate achieved for a single subject (S32) was 93%, i.e., 93 of the 100 test samples. These results were achieved with $L = 1.25$ Hz, $W = 21.25$ Hz, $s = 180$ ms after onset of the visual presentation of the first word of each sentence, and $e = 2200$ ms, marking the ending of the recordings used for the least-squares criterion of fit. The best bipolar sensor was C4-T6. In Figure 4 we show at the top the best and at the bottom the worst fit, as measured by the least-squares criterion, for the 93 sentences correctly recognized. The sum of squares for the worst was more than three times that for the best.

Invariance between visual images and their names

The third experiment showed that the visual images generated on a computer screen, of a familiar shape, such as a circle or triangle, were very similar to the brain images generated by the corresponding word (Suppes, Han, Epelboim and Lu 1999b). This surprising result very much reinforced a classical solution of how the mind has general concepts. It is a famous episode in the history of philosophy in the eighteenth century that Berkeley and Hume strongly criticized Locke's conception of abstract or general ideas. Berkeley has this to say in *A New Theory of Vision* (1709/1901):

> It is indeed a tenet, as well of the modern as the ancient philosophers, that all general truths are concerning universal abstract ideas; without which, we are told, there could be no science, no demonstration of any general proposition in geometry. But it were no hard matter, did I think it necessary to my present purpose, to shew that propositions and demonstrations in geometry might be universal, though they who make them never think of abstract general ideas of triangles or circles.
>
> After reiterated efforts and pangs of thought to apprehend the general idea of a triangle, I have found it altogether incomprehensible. And surely, if any one were able to let that idea into my mind, it must be the author of the *Essay concerning Human Understanding*: he, who has so far distinguished himself from the generality of writers, by the clearness and significancy of what he says. Let us therefore see how this celebrated author describes the general or which is the same thing, the abstract idea of a triangle. "It must be", says he, "neither oblique nor rectangle, neither equilateral, equicrural, nor scalenum; but all and none of these at once. In effect it is somewhat imperfect that cannot exist; an idea, wherein some parts of several different and inconsistent ideas are put together." (*Essay on Human Understanding*, B. iv. ch. 7. s. 9.) This is the idea which he thinks needful for the enlargement of knowledge, which is the subject of mathematical demonstration, and without which we could never come to know any general proposition concerning triangles. Sure I am, if this be the case, it is impossible for me to attain to know even the first elements of geometry: since I have not the faculty to frame in my mind such an idea as is here described
>
> (Berkeley, pp 188–189.)

Hume, in a brilliant exposition and extension of Berkeley's ideas, in the early pages of *A Treatise of Human Nature,* (1739/1951) phrased the matter beautifully in the opening paragraph of Section VII, entitled *Of Abstract Ideas*:

> A very material question has been started concerning *abstract* or *general* ideas, *whether they be general or particular in the mind's conception of them.* A great philosopher has disputed the receiv'd opinion in this particular, and has asserted, that all general ideas are nothing but particular ones, annexed to a certain term, which gives them a more extensive signification, and makes them recall upon occasion other individuals, which are similar to them. As I look upon this to be one of the greatest and most valuable discoveries that has been made of late years in the republic of letters, I shall here endeavour to confirm it by some arguments, which I hope will put it beyond all doubt and controversy. (Hume, *Treatise*, p. 17.)

Although not discussed by Berkeley and Hume, we also confirmed that the same is true of simple patches of color. In other words, a patch of red and the word "red" generate similar brain images in the auditory part of the cortex.

The specific significant results were these. By averaging over subjects as well as trials, we created prototypes from brain waves evoked by stimuli consisting of simple visual images and test samples from brain waves evoked by auditory or visual words naming the visual images. We correctly recognized from 60% to 75% of the test-sample brain waves. Our general conclusion was that simple shapes and simple patches of color generate brain waves surprisingly similar to those generated by their verbal names. This conclusion, taken together with extensive psychological studies of auditory and visual memory, support the solution conjectured by Berkeley and Hume. The brain, or, if you prefer, the mind, associates individual visual images of triangles, e.g., to the word *triangle*. It is such an associative network that is the likely procedural replacement for the mistaken attempt by Locke to introduce abstract ideas.

Comparisons of averaged and filtered brain waves generated by visual images and spoken names of the images are shown in Figure 5. Time after the onset of the stimulus (visual image or word) is shown in milliseconds on the abscissa. In the upper panel the solid curved line is the prototype brain wave generated by the color blue displayed as a blank computer screen with a blue background. The dotted curved line is the test-sample brain wave generated by the spoken word *blue*. In the lower panel are the prototype brain wave (solid line), generated by display of a triangle on the screen, and the test-sample brain wave (dotted line), generated by the spoken word *triangle*. In neither case is the match perfect, for even when the same stimulus is repeated, the filtered brain waves do not match exactly, since the brain's

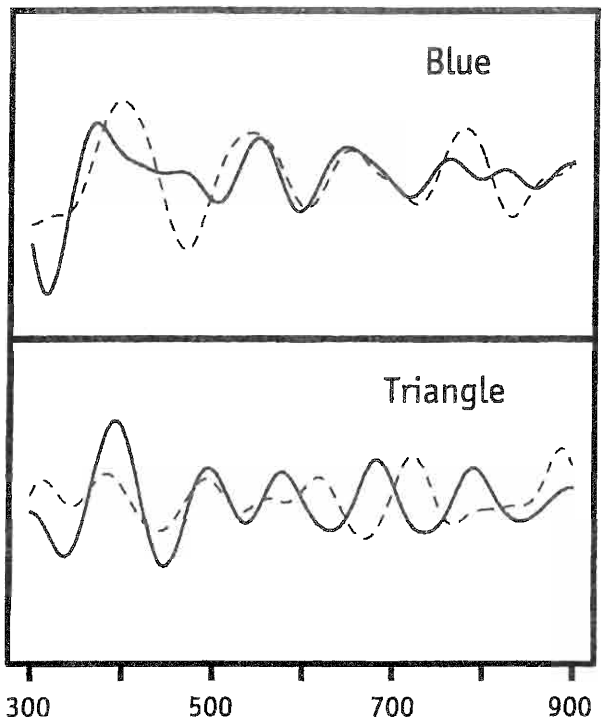

Figure 5. Comparison of filtered brain waves generated by visual images (solid curves) with those generated by the spoken names (dotted curves) of the images

electric activity continually changes from moment to moment in numerous ways. But, all the same, there are many invariances necessary for human communication, and even at this early stage we can identify some of them.

Criticisms of results and response

I first sketch the general nature of the criticisms. In many brain imaging experiments the data are very rich and complex. Consequently, a complicated procedure may also be used to find, for given conditions, an optimal value. The search for this optimal value, which here is the best correct recognition rate, is analogous to computing an extreme statistic for a given probability distribution. The basis of the analogy is that the search corresponds to possibly many repetitions of a null-hypothesis experiment. These repetitions require computation of the appropriate extreme statistic. Moreover, if several parameters are estimated in finding such an optimal value, the significance

of the value found may be challenged. The basis of such a challenge is the claim that for rich data and several estimated parameters, even a random assignment of the meaningful labels in the experiment may still produce a pretty good predictive result with a large enough number of chance repetitions.

Extreme statistics

We meet this criticism in two ways. The first is to derive the extreme-statistic distribution under the null hypothesis of a binomial distribution arising from such random assignments of labels. We compute how many standard deviations the scientifically interesting predictive result is from the mean of the extreme-statistic distribution. Physicists usually accept at least three or four standard deviations as a significant result. Other scientists and statisticians usually prefer a p value expressing the probability of the observed result under the null hypothesis. Here we report both measures.

The second approach is meant to answer those who are skeptical that the null hypothesis of a binomial distribution with a single parameter for the chance probability of a correct classification will adequately characterize the structure of the data even after a random permutation of the labels. To respond to such possible skeptics, we also compute a recognition rate for a sample of 50 random permutations of labels. We then fit a beta distribution to each such sample for a given experimental condition to compare with the corresponding extreme statistic distribution arising from the null hypothesis.

We first derive the extreme statistic under the null hypothesis.

Let p = probability of a success, a correct classification in our case, on a single trial, and $q = 1 - p$. Let $\mathbf{X}$ be the random variable whose value is the number k of successes in n independent trials. The probability of at least k successes is:

$$P(\mathbf{X} \geq k) = \sum_{j=k}^{n} P(\mathbf{X} = j) = \sum_{j=k}^{n} \binom{n}{j} p^j q^{n-j}. \tag{3}$$

Now we repeat the experiment governed by a binomial distribution. So we have r independent repetitions of the n independent trials. For r repetitions ($r = 21,000$ in the 100-sentences experiment), the random variable representing the extreme statistic is

$$\mathbf{Y} = \max(\mathbf{X}_1, \ldots, \mathbf{X}_r). \tag{4}$$

Let $P(\mathbf{Y} \geq k)$ be the probability that $\mathbf{Y}$ is at least k in at least one of the r

repetitions, the extreme statistic of interest. Then clearly

$$\begin{aligned} P(\mathbf{Y} \geq k) &= 1 - P(\mathbf{X} < k)^r, \\ &= 1 - \left[\sum_{j=0}^{k-1} \binom{n}{j} p^j q^{n-j}\right]^r. \end{aligned} \tag{5}$$

We also need the theoretical density distribution of $\mathbf{Y}$, to compare to various empirical results later. This is easy to compute from (3).

$$\begin{aligned} P(\mathbf{Y} = k) &= P(\mathbf{Y} \geq k) - P(\mathbf{Y} \geq k+1), \\ &= P(\mathbf{X} < k+1)^r - P(\mathbf{X} < k)^r, \\ &= \left[\sum_{j=1}^{k} \binom{n}{j} p^j q^{n-j}\right]^r - \left[\sum_{j=1}^{k-1} \binom{n}{j} p^j q^{n-j}\right]^r. \end{aligned} \tag{6}$$

From (4) we can compute the mean and standard deviation of the extreme statistic $\mathbf{Y}$.

Beta distribution fitted to empirical sample

Second, we report results for the beta distribution on $(0, 1)$ fitted to the empirical sample of extreme statistics. The density $f(x)$ of the beta distribution is:

$$f(x) = \begin{cases} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}, \ a, b, > 0, & 0 < x < 1, \\ 0 & \text{otherwise}, \end{cases} \tag{7}$$

where $\Gamma(a)$ is the gamma function. If $\mathbf{Z}$ is a random variable with a beta distribution, then its mean and variance are given as simple functions of the parameters a and b.

$$\mu_{\mathbf{Z}} = E(\mathbf{Z}) = \frac{a}{a+b}, \tag{8}$$

$$\sigma^2_{\mathbf{Z}} = \text{Var}(\mathbf{Z}) = \frac{ab}{(a+b)^2(a+b+1)}. \tag{9}$$

The probability that the random variable $\mathbf{Z}$ has a value equal to or greater than $\frac{k}{n}$ is:

$$P(\mathbf{Z} \geq \frac{k}{n}) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_{\frac{k}{n}}^{1} x^{a-1}(1-x)^{b-1} dx. \tag{10}$$

The computation of $P(\mathbf{Z} \geq \frac{k}{n})$ is difficult for the extreme tail of the distribution. In some cases we use a mathematically rigorous upper bound that is not the best possible, but easy to compute, namely, just the area of the rectangle with height $f(\frac{k}{n})$ containing the tail of the distribution to the right of $f(\frac{k}{n})$:

$$P(\mathbf{Z} \geq \frac{k}{n}) \leq f(\frac{k}{n})(1 - \frac{k}{n}), \tag{11}$$

where $f(\frac{k}{n})$ is defined by (5).

Computation of extreme statistics

I begin with the second experiment using 100 sentences. As a check on the null hypothesis, we constructed an empirical distribution of the extreme statistic by sampling 50 random permutations. Several points are to be noted.

1. A permutation of the 100 sentence "labels" is randomly drawn from the population of 100! possible permutations, and the sentence test samples are relabeled using this permutation.

2. Exactly the same grid of parameters (L, W, s, e) is now run for each bipolar pair of sensors, as for the correct labeling on the data of subject S32, to obtain, by Fourier analysis, filtering and selection of temporal intervals (s, e), a best rate of recognition or classification for the random label assignment. For the 100-sentences experiment, the number of points on the grid tested for each random permutation is $7 \times 10 = 70$ for $L \times W$, $5 \times 4 = 20$ for $s \times e$ and 15 for the number of sensors, so the number of repetitions r, from the standpoint of the null hypothesis, is $70 \times 20 \times 15 = 21,000$.

3. This random sampling of label permutations is repeated, and the recognition results computed, until a sample of 50 permutations has been drawn.

In Figure 6 I show the cumulative computation of the mean m and standard deviation s for the sample of 50 label permutations for the data of subject S32. For the full sample of 50 the mean $m = 6.04$ and the standard deviation $s = 0.77$. In Figure 7 I show: (i) the frequency distribution of the null-hypothesis extreme statistic $\mathbf{Y}$ with $n = 100, p = 0.01$ and $r = 21,000$, (ii) the empirical histogram of the maximum number of successes obtained for the 50 sample points with $r = 21,000$, and (iii) the fitted beta distribution as well. From Figure 7 it is visually obvious that the correct classification of more than 80 of the 100 sentences for S32 is not compatible with either

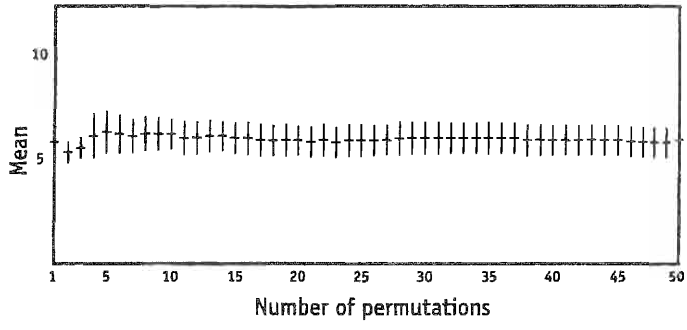

Figure 6: Cumulative mean and standard deviation of the recognition rate of the sample of random permutations.

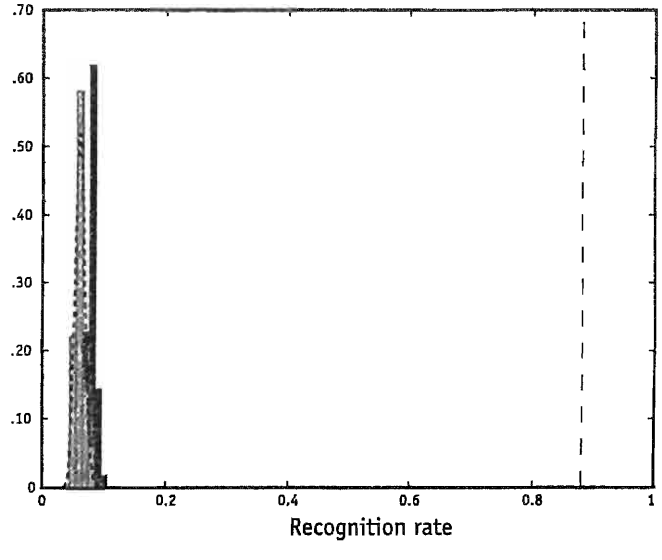

Figure 7: The frequency distribution (dark area) of the null-hypothesis extreme statistic $\mathbf{Y}$, the histogram of the sample of 50 random permutations (lightly shaded areas), the fitted beta distribution (dotted line), and on the right (dashed vertical line) the recognition rate for S32.

the distribution of the extreme statistic $\mathbf{Y}$ or the estimated beta distribution for the sample grid computations based on 50 random permutations of the labels. The fact that the beta distribution fits slightly better than the distribution of $\mathbf{Y}$ is not surprising, since no free parameters were estimated for the latter. A finer search, with much larger r, yielding the higher result of 93 out of 100, is discussed in the next paragraph.

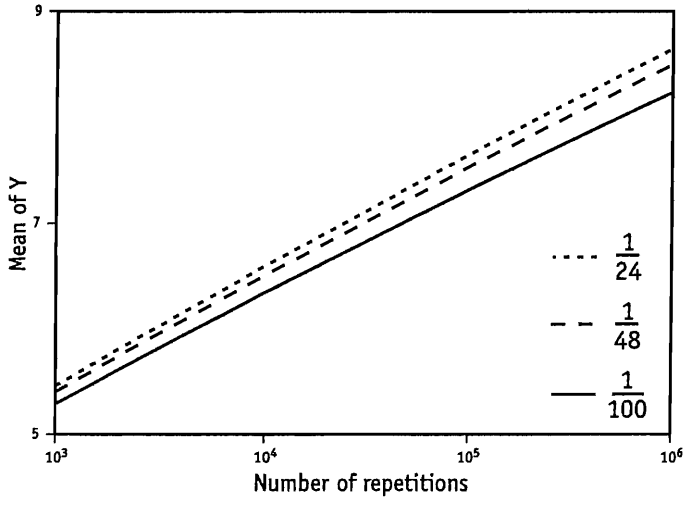

Figure 8: Mean of the null-hypothesis extreme statistic $\mathbf{Y}$ as a function of the number r of repetitions.

What is perhaps surprising is that the mean $\mu = 6.95$ of the null-hypothesis extreme statistic $\mathbf{Y}$ is slightly larger than the mean $m = 6.04$ of the empirical sample distribution. Three points are worth noting. First, the standard deviation $s = 0.77$ of the empirical sample is larger than the standard deviation $s = 0.67$ of the extreme statistic $\mathbf{Y}$. I comment on this difference below. Second, I show in Figure 8 the rate of growth of the recognition rate for the null-hypothesis extreme statistic $\mathbf{Y}$, for $n = 100$ and $p = 0.01$, and some other values of p and r used later, as r is increased by one or more orders of magnitude. As can be seen, under the null hypothesis the correct-recognition growth rate is slow. As an important example, we refined by extensive search the grid for the data of S32. We did not use a complete grid, but refined and extended only in promising directions. Extended comparably in all directions, the order of magnitude of r would be 10^7, i.e., 10,000,000 repetitions. So we computed the null-hypothesis distribution of $\mathbf{Y}$ for this large value of r, which is much larger than any actual computation we made. Even for this

large grid, the mean of the extreme statistic $\mathbf{Y}$ for $r = 10^7$ only moved to $\mu = 9.60$. With the standard deviation now reduced to 0.62, the number of standard deviation units of the distance between 93 and 9.60 is 134.5, larger than before.

In reflecting on these results, it is important to keep in mind that physicists are usually very happy with a separation of 6 or 7 standard deviations between the classical or null-hypothesis prediction and new observed results, e.g., in quantum entanglement experiments of the sort discussed in Section 7.2. The significance levels obtained in our brain experiments and the very large distance in standard deviations of the observed results from the expected null-hypothesis results are as yet seldom found in psychological or neuroscience experiments.

The third point concerns the level of significance, or p value, we report for rejecting the null hypothesis. The p value of the result of $k = 93$, which is 134.5 standard deviations from the mean of the null-hypothesis extreme statistic $\mathbf{Y}$ is extravagantly low, at the very least $p < 10^{-100}$. Every other aspect of the experiment would have had to be perfect to support such a p value. (We did check the computation of 93 on two different computers running different programs.) So here, and in other cases later, we report only the inequality $p < 10^{-10}$ for such very small values, but the actual number of standard deviations from the mean is reported.

We also checked that using the rigorous upper bound of inequality (9), $P(\mathbf{X} \geq 93)$, computed for the fitted beta distribution, is also on the order of $p < 10^{-100}$. This is further support for the view that the p value inequality used later, namely, $p < 10^{-10}$, is highly conservative.

Analysis of earlier studies

I have emphasized the gain in predictive results from averaging across subjects as well as trials. The best result of the second experiment of 93% for one individual subject prompted us to review the best individual results in earlier experiments. In each experiment we have performed, the analysis of at least one individual subject's brain waves yielded a correct classification greater than 90%, with the exception of the 48-sentence experiment, mentioned already, which was 77%. (In Suppes, Han, Epelboim and Lu (1999a), this 77% was reported as 79%, because a finer grid was used.) Results for the best subject in the various experiments are summarized in Table 1.

With one exception, the p values shown in Table 1 are highly significant, by most standards of experimental work, extravagantly so. The exception is for the visual-image experiment in which 8 simple visual images were presented as stimuli. For four of the subjects, as shown in Table 1, we were

Table 1: *Exceptional Recognition Rates*[a]

Experiment	Subj.	Number of Successes	Chance Prob.	% Cor.	Repet. r	Statistic $\mathbf{Y}$ μ	m	σ	s	Significance $\#\sigma$	$\#s$	p value
7 visual words[1]	S1	32 of 35	1 of 7	91	2925	13.40		0.92		20.2		$< 10^{-10}$
7 auditory words[1]	S3	34 of 35	1 of 7	97	3600	13.55		0.91		22.5		$< 10^{-10}$
12 sentences[2]	S8	56 of 60	1 of 12	93	60,480	16.10		0.90		44.3		$< 10^{-10}$
24 visual sent [3]	S18	24 of 24	1 of 24	100	30,800	6.83	5.64	0.63	0.79	27.3	23.2	$< 10^{-10}$
48 visual sent.[3]	S26	38 of 48	1 of 48	79	30,800	7.04	6.20	0.63	0.94	49.1	33.8	$< 10^{-10}$
8 visual images[4]	4 Ss	8 of 8	1 of 8	100	95,550	6.28	4.92	0.46	0.69	3.7	4.5	$< .01$
100 visual sent.	S32	88 of 100	1 of 100	88	21,000	6.95	6.04	0.67	0.77	121.0	106.4	$< 10^{-10}$
100 visual sent.	S32	93 of 100	1 of 100	93	10^7	9.60		0.62		134.5		$< 10^{-10}$

1. Suppes, Lu, and Han (1997); 2. Suppes, Han and Lu (1998)
3. Suppes, Han, Epelboim and Lu (1999a); 4. Suppes, Han, Epelboim and Lu (1999b)

[a]The first column lists the experiment, with the last two entries being for the 100-sentence one. The subjects, listed in the second column, are numbered continuously from the experiments first reported in Suppes, Lu and Han (1997). The third column shows the maximum number of test samples successfully recognized out of the total presented. The fourth column shows the chance probability of a correct classification, which is simply 1 divided by the number of prototypes. The fifth column records the percent correct, as computed from the third column. The sixth columns shows the number r of repetitions used in the particular experiment to compute the extreme statistic. The number r is also the number of repetitions originally used in the grid for the initial search with correct labels. The seventh column records the mean μ of the null-hypothesis extreme statistic $\mathbf{Y}$. The eighth column records the mean of the empirical samples of extreme statistics for the experiments for which we made this computation. The ninth column shows the standard deviation σ of the extreme statistic $\mathbf{Y}$, and the tenth column the corresponding standard deviation s of the empirical samples. The eleventh column records the number $\frac{k-\mu}{\sigma}$, which is the number of standard deviations that the number k of successes recorded in column three is from the mean μ of the null-hypothesis distribution of the extreme statistic $\mathbf{Y}$, and the twelfth column the corresponding number for the empirical sample. The thirteenth column shows a conservative bound for the p value of the observed number k of successes, with respect to the distribution of the extreme statistic $\mathbf{Y}$, as given by equation (6). In the case of the four subjects in the visual-image experiment, m and s are the average for the four. The superscript on the description of each experiment is the reference to the published study. (The EEG sensor or bipolar pair of sensors and the optimal filter for each subject, except the four subjects of Suppes, Han, Epelboim and Lu (1999b), were as follows: S1:T6, 1-10 Hz; S3:T3, 3-11 Hz; S8:C4-T6, 2.5-9 Hz; S18:P4-T6, 0.5-10 Hz; S26:C4-C6, 1-15 Hz; S32:C4-T6, 1.25-22.5 Hz. The optimal parameters were often not unique.)

able to classify all 8 brain waves correctly, but this perfect result of 100 percent was significant only at the level of $p < 0.01$ for the null hypothesis, because with enough repetitions the best guesses under the null hypothesis do pretty well also, with $\mu = 6.28$.

The lesson for experimental design of this last point is obvious. If the data are massive and complex, as in the brain experiments described, and extensive search for optimal parameters is required, then the probability p of a correct response under the null hypothesis should be small. Figure 8 graphically makes the point. When p is small the number of repetitions can be very large, without affecting very much the mean of $\mathbf{Y}$, the extreme statistic of r repetitions. As can be seen also, from Table 1, when $p = 0.01$, the binomial parameter of the 100-sentences experiment, even 10,000,000 repetitions under the null hypothesis of 100 trials, increases $E(\mathbf{Y})$ only slightly to 9.60. To put the argument dramatically, at the rate of 1 second per trial, it would take more time than the present estimated age of the universe to have enough repetitions to obtain $E(\mathbf{Y}) \geq 93$.

I say in the preceding paragraph that p should be small, but that is too simple. The other way out, used in the first two experimental conditions of Table 1, 7 visual words and 7 auditory words, is to increase the number of test samples. In those two conditions, $p = \frac{1}{7}$, but the number of test samples was 35, and, as can be seen from the table, the null hypothesis was rejected at a level better than 10^{-10}. Reanalysis of the data from the visual-image experiment with $p = \frac{1}{8}$, in a similar approach by increasing the number of test samples from 8 to 24 yielded some better levels of rejection of the null hypothesis. The details are reported below.

More skeptical questions

As in all regimes of detailed experimentation, there is no sharp point after which further experiments need not be conducted, because all relevant questions have been answered. Galison (1987) made a detailed study of several important research programs of experimentation in physics. It seems likely that the main aspects of his analysis apply to many other areas of science.

> Amidst the varied tests and arguments of any experimental enterprise, experimentalists must decide, implicitly or explicitly, that their conclusions stand *ceteris paribus:* all other factors being equal. And they must do so despite the fact that the end of an experimental demonstration is not, and cannot be, based purely on a closed set of procedures. ... Certain manipulations of apparatus, calculations, assumptions, and arguments give confidence to the experimentalist: what are they? ... When do experimentalists stake their claim on the reality of an effect? When do they assert that the counter's pulse or the spike in a graph is more than an artifact of the apparatus or environment? In short: How do experiments end?
>
> (Galison 1987, pp. 3–4)

In the context of the present brain experiments, the question is not really when do they end, but when do the computations on the experimental data come to an end? I examine three more different, but typical skeptical questions that are asked about new research with strong statistical support for its validity.

Other pairs in the first experiment with 48 sentences

Some skeptics commented to us that we were just lucky in the particular 2-element partition of the subjects we analyzed. So, we ran all 510 2-element partitions of the 9 subjects, with the same optimal values as in Table 1, without trying all points on the grid (Suppes, Wong, et al., to appear). In Figure 9 we show the histogram of these 510 partitions. The level of significance of the results is $p < 10^{-10}$ for all but 4 of the 510 possibilities, and one of these 4 has $p < 10^{-7}$. The best result is 46 out of 48, which holds for several partitions. So the brain-wave invariance between subjects argued for in the earlier study is robustly supported by the present more thorough statistical analysis. Another view of the same data is shown in Figure 10, where the number of subjects in the prototype of each 2-element partition is plotted on the abscissa and on the ordinate is shown the mean number of correct classifications of the 48 sentences for each type of prototype. Surprisingly, the mean results are good ($p < 10^{-10}$) when the prototype has only 1 subject or all but 1 subject, i.e., 8 subjects. The evidence is pretty convincing that our original choice of a partition was not just some happy accident.

Test of a timing hypothesis for the experiment with 100 sentences

In discussing with colleagues the high recognition rate of 93% obtained in the second experiment reported above, and the earlier results summarized in Table 1, several persons skeptically suggested that perhaps our recognition rates are just coming from the different timing of the visual presentation of words in different sentences. Sentences were presented one word at a time in the center of the computer screen, with the onset time of each visual word the same as the onset time of the corresponding auditory presentation of the sentence. Visually displaying one word at a time avoids many troublesome eye movements that can disrupt the brain waves, and it has also been shown to be an effective fast way to read for detailed content (Rubin and Turano 1992). The duration of each visual word of a sentence also matched the auditory duration within a few milliseconds.

To test this timing idea, which is supported by the presence of an evoked response potential at the onset of most visual words, we used a recognition

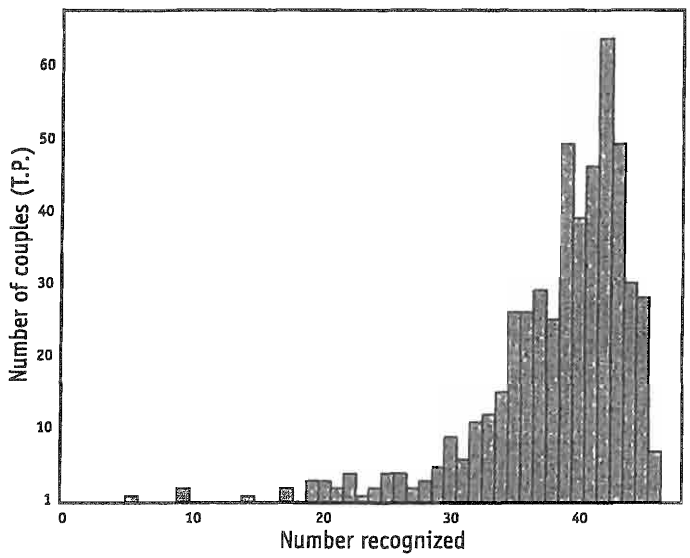

Figure 9: Histogram of the correct recognition rates for the 510 2-element partitions of the 9 subjects in the 48-sentences experiment.

model that depended only on an initial segment of the brain-wave response to each word in a sentence (Suppes, Wong, et al., to appear). The model replaces the two parameters s and e for the temporal interval by two different parameters. The first is α, which is the estimated time lag between the onset of each word in every sentence and the beginning of the corresponding brain wave in the cortex. The second is β, which is the proportion of the displayed length of each word, starting from its onset, used in the prototype for recognition after the delay time α for the signal to reach the cortex. Because of the variable length of words and sentences, we normalized the least squares computation by dividing by the number of observations used. If only timing, and not the full representation of the word, matters in recognition, then only a small portion of the initial segment of a word is needed, essentially the initial segment containing the onset-evoked response potential. On the other hand, if the full representation of the word is used in successful recognition, in terms of our least squares criterion, then the larger β is, the better for recognition. To adjust β to the temporal length of each word displayed, we expressed β as a decimal multiple of the temporal display length of word i of each sentence. The best predictive result was for $\alpha = 200$ ms and $\beta = 1.25$, with a recognition rate of 92%. The recognition rate as a function of $0.125 \leq \beta \leq 2.00$ is shown in Figure 11. The rate of correct recognition in-

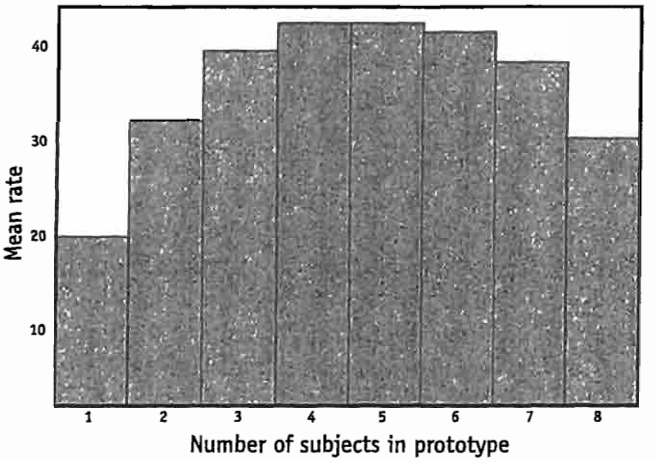

Figure 10: Histogram of mean number of the correct recognition rate for the 510 2-element partitions, indexed from 1-8 for the number of subjects in the prototype. So, e.g., 1 on the abscissa corresponds to all the 2-element partitions having exactly one subject used for the prototype.

creases monotonically with β up to $\beta = 1.25$ and then declines slowly after $\beta = 1.50$. These results support two conclusions. First, timing is important. The recognition rate of 45% for $\beta = 0.125$ is much greater than a chance outcome. But, second, the more complete the brain-wave representation of the words in a sentence, the better for recognition purposes.

Censoring data in the visual-image experiment

One kind of skeptical question that keeps the computations going is about artifacts. Perhaps the remarkable levels of statistical significance are due to some artifacts in the data. Now there is a long history of the problems of artifacts in EEG research. A main source is eye blinks and saccadic or pursuit eye movements, another is ambient current in the environment, mainly due to the 60 Hz oscillation of the standard alternating current in any building in which experiments are ordinarily conducted, and still another source is in the instrumentation for observing and recording the electric brain waves. This list is by no means exhaustive. There is a large literature on the subject from several different angles, but it would be too much to survey it here.

Given that the extreme statistics of the random permutations had a mean close to the low mean of the null hypothesis for the second experiment, it is extremely unlikely that any artifacts could account for the correct classifica-

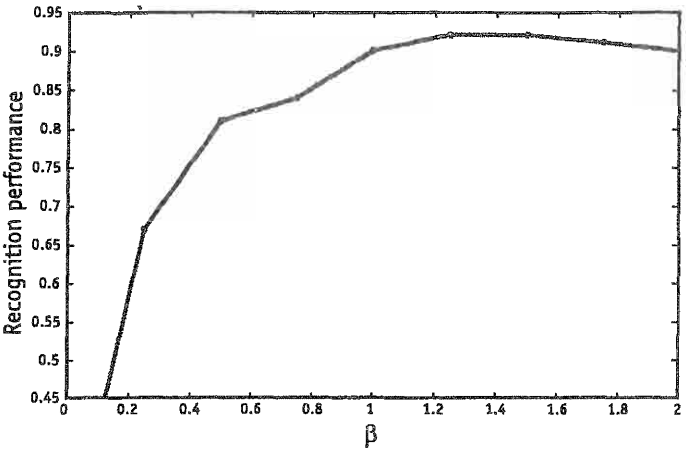

Figure 11: Classification results using different initial segments of brain-wave data after onset of each word in a sentence. Segment i begins after α ms following onset of word i in a sentence and segment i ends after $\beta \cdot l_i$ ms, where l_i is the length in ms of the visual display of word i

tion of 93% in the second experiment, or, in fact, in any of the others with $p < 10^{-10}$. But artifact removal remains an important topic, not so much to meet ill-informed skeptical questions, but to improve the classification results, as is the case for an example that follows.

I restrict myself to describing how we used a rather familiar statistical approach, rather than any visual inspection of the recorded data for eye blinks or other artifacts. In our larger experiments with more than 40,000 trials it is impractical to try to use traditional observational methods to detect artifacts. The approach was to censor the data, but to introduce a free parameter to optimize the censoring—optimize in the sense already described of maximizing the correct recognition rate.

Let $\mathbf{X}_{ik}$ = observation i on trial k, and let ω be the number of trials averaged to create a prototype or test sample. Then

$$m_i = \overline{\mathbf{X}}_i = \frac{1}{\omega} \sum_{k=1}^{\omega} \mathbf{X}_{ik}.$$

In similar fashion, we compute the variance s_i^2. Let α be the free parameter for censoring such that if

$$|\mathbf{X}_{ik} - m_i| > \alpha s_i$$

eliminate obervation i of trial k from the data being averaged. The computational task is to find the $\hat{\alpha}$ that optimizes classification. In the example reported here, we ran a one-dimensional grid of 20 values of s_i to approximate the best value $\hat{\alpha}$.

The experiment for which the extreme statistics were not highly significant was the visual-image one already described, the one that was relevant to the eighteenth-century controversy about abstract ideas. The data, as I said earlier, supported in a direct way the skeptical views of Berkeley and Hume, but the statistical support was not very strong. So we reanalyzed the data, creating 24 rather than 8 test samples, and we also ran the experiment with four monolingual Chinese (Mandarin) speakers to confirm the verbal part, with auditory or visual words, in Chinese as well as English. The details are reported in Suppes, Wong, et al. (to appear).

Table 2 shows the significant results for cross-modal classification. The first two conditions are for the original experiment using English. For the feminine-auditory-voice representing brain waves (AWF) as prototypes and the visual-image brain waves as test samples, 15 of the 24 test samples were correctly classified after censoring, an improvement from 11 of 24 without censoring, for a resulting significance level of $p < 0.016$. When the roles of prototype and test sample were reversed the results of censoring were better, 16 of the 24 test samples correctly classified after censoring, with $p < 0.001$. Note that the significance levels here are conservative, based on the complete grid search equal to $r = 1,470,000$, the number of repetitions under the extreme-statistic null hypothesis.

In the case of the Chinese, the best results were in the comparison of the auditory and visual presentation of the eight words, with the best result being for the visual Chinese words (VW) as prototypes and the auditory Chinese words (AW) as test samples, in the censored case, 17 of 24 correctly classified, with p approximately 0.0001, and with, as before, the number of repetitions r, under the null hypothesis, greater than a million. So we end by strengthening the case for Berkeley and Hume, without claiming the evidence is as yet completely decisive.

An appropriate stopping point for this analysis is to emphasize that censoring does not guarantee improvement in classification. Most of the other results in Table 1 showed little improvement from censoring, but then for all of the experiments reported there, except for the visual-image one, the results were highly significant without censoring.

As in many areas of science, so with EEG recordings, statistical and experimental methods for removing artifacts and other anomalies in data constitute a large subject with a complicated literature. I have only reported a common statistical approach here, but I am happy to end with this one exam-

Exper.	R	Y μ	σ	Significance $\#\sigma$	p-value
English					
AWF/VI	11	11.57	.692	0.8	$\approx .500$
censor	15	13.01	.642	3.1	$< .016$
VI/AWF	11	11.57	.692	0.8	$\approx .500$
censor	16	13.01	.642	4.7	$< .001$
Chinese					
AW/VW	9	11.42	.690		
censor	16	12.86	.662	4.7	$< .001$
VW/AW	13	11.42	.690	2.3	$\approx .05$
censor	17	12.86	.662	6.3	$\approx .0001$

Table 2: Cross-modal results in visual-image experiment with censored data.

ple of a typical method of "cleaning up" data. It is such censored data that should be used to form a representation suitable for serving as a test of some theory or, as is often the case, some congerie of theoretical ideas.

QUANTUM MECHANICAL ENTANGLEMENT

The literature on hidden variables in quantum mechanics is now enormous. This section covers mainly the part dealing with probabilistic representation theorems for hidden variables, even when the hidden variables may be deterministic. Fortunately, this body of results can be understood without an extensive knowledge of quantum mechanics, which is not developed *ab initio* here. Many of the results given are taken from joint work with Acacio de Barros and Gary Oas (Suppes, de Barros and Oas 1998; de Barros and Suppes 2000).

First, I state, and sketch the proof, of the fundamental theorem that there is a factoring hidden variable for a finite set of finite or continuous observables, i.e., random variables in the language of probability theory, if and only if the observables have a joint probability distribution. The physically important aspect of this theorem is that under very general conditions the existence of a hidden variable can be reduced completely to the relationship between the observables alone, namely, the problem of determining whether or not they have a joint probability distribution compatible with the given data, e.g., means, variances and correlations of the observables.

I emphasize that although most of the literature is restricted to no more than second-order moments such as covariances and correlations, there is no

necessity to make such a restriction. It is in fact violated in the third-order or fourth-order moments that arise in the well-known Greenberger, Horne and Zeilinger (1989) three- and four-particle configurations providing new Gedanken experiments on hidden variables, which are discussed later.

Factorization

In the literature on hidden variables, the principle of factorization is sometimes baptized as a principle of locality. The terminology is not really critical, but the meaning is. We have in mind a quite general principle for random variables, continuous or discrete, which is the following. Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be random variables, then a necessary and sufficient condition that there is a random variable $\boldsymbol{\lambda}$, which is intended to be the hidden variable, such that $\mathbf{X}_1 \ldots, \mathbf{X}_n$ are conditionally independent given $\boldsymbol{\lambda}$, is that there exists a joint probability distribution of $\mathbf{X}_1, \ldots, \mathbf{X}_n$, without consideration of $\boldsymbol{\lambda}$. This is the general fundamental theorem relating hidden variables and joint probability distributions of observable random variables.

THEOREM 1. (Suppes and Zanotti 1981, Holland and Rosenbaum 1986) *Let n random variables $X_1, \ldots, X_n$, finite or continous, be given. Then there exists a hidden variable $\boldsymbol{\lambda}$ such that there is a joint probability distribution F of $(\mathbf{X}_1, \ldots, \mathbf{X}_n, \boldsymbol{\lambda})$ with the properties*

(*i*) $F(x_1, \ldots, x_n \mid \lambda) = P(\mathbf{X}_1 \leq x_1, \ldots, \mathbf{X}_n \leq x_n \mid \boldsymbol{\lambda} = \lambda)$

(*ii*) *Conditional independence holds, i.e., for all $x_1, \ldots, x_n, \lambda$,*

$$F(x_1, \ldots, x_n \mid \lambda) = \prod_{j=1}^{n} F_j(x_j|\lambda),$$

if and only if there is a joint probability distribution of $\mathbf{X}_1, \ldots, \mathbf{X}_n$. Moreover, $\boldsymbol{\lambda}$ may be constructed so as to be deterministic, i.e., the conditional variance given $\boldsymbol{\lambda}$ of each $\mathbf{X}_i$ is zero.

To be completely explicit in the notation

$$F_j(x_j|\lambda) = P(\mathbf{X}_j \leq x_j|\boldsymbol{\lambda} = \lambda). \tag{12}$$

Idea of the proof. Consider three ± 1 random variables $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$. There are 8 possible joint outcomes $(\pm 1, \pm 1, \pm 1)$. Let p_{ijk} be the probability of outcome (i, j, k). Assign this probability to the value λ_{ijk} of the hidden

variable λ we construct. Then the probability of the quadruple (i, j, k, λ_{ijk}) is just p_{ijk} and the conditional probabilities are deterministic, i.e.,

$$P(\mathbf{X} = i, \mathbf{Y} = j, \mathbf{Z} = k \mid \lambda_{ijk}) = 1,$$

and factorization is immediate, i.e.,

$$P(\mathbf{X} = i, \mathbf{Y} = j, \mathbf{Z} = k \mid \lambda_{ijk}) = P(\mathbf{X} = i \mid \lambda_{ijk})P(\mathbf{Y} = j \mid \lambda_{ijk})P(\mathbf{Z} = k \mid \lambda_{ijk}).$$

Extending this line of argument to the general case proves the joint probability distribution of the observables is sufficient for existence of the factoring hidden variable. From the formulation of Theorem 1 necessity is obvious, since the joint distribution of $(\mathbf{X}_1, \ldots, \mathbf{X}_n)$ is a marginal distribution of the larger distribution $(\mathbf{X}_1 \ldots, \mathbf{X}_n, \lambda)$.

It is apparent that the construction of λ is purely mathematical. It has in itself no physical content. In fact, the proof itself is very simple. All the real mathematical difficulties are to be found in giving scientifically interesting criteria for observables to have a joint probability distribution.

Locality

The next systematic concept to discuss is locality. What John Bell meant by locality is made clear in the following quotation from his well-known 1964 paper (Bell 1964).

> It is the requirement of locality, or more precisely that the result of a measurement on one system be unaffected by operations on a distant system with which it has interacted in the past, that creates the essential difficulty. ... The vital assumption is that the result B for particle 2 does not depend on the setting **a**, of the magnet for particle 1, nor A on **b**.

To make the locality hypothesis explicit, we need to use additional concepts. For each random variable $\mathbf{X}_i$, we introduce a vector M_i of parameters for the local apparatus (in space-time) used to measure the values of random variable $\mathbf{X}_i$.

DEFINITION 1. (LOCALITY CONDITION I)

$$E(\mathbf{X}_i^k | M_i, M_j, \lambda) = E(\mathbf{X}_i^k | M_i, \lambda),$$

where $k = 1, 2$, *corresponding to the first two moments of* $\mathbf{X}_i$, $i \neq j$, *and* $1 \leq i, j \leq n$.

Note that we consider only M_j on the supposition that in a given experimental run, only the correlation of $\mathbf{X}_i$ with $\mathbf{X}_j$ is being studied. Extension to more variables is obvious. In many experiments the direction of the measuring apparatus is the most important parameter that is a component of M_i.

DEFINITION 2. (LOCALITY CONDITION II) *The distribution of λ is independent of the parameter values M_i and M_j, i.e., for all functions g for which the expectation $E(g(\lambda))$ and $E(g(\lambda)|M_i, M_j)$ are finite,*

$$E(g(\lambda)) = E(g(\lambda)|M_i, M_j).$$

Here we follow Suppes and Zanotti (1976). In terms of Theorem 3, locality in the sense of Condition I is required to satisfy the hypothesis of a fixed mean and variance for each $\mathbf{X}_i$. If experimental observation of $\mathbf{X}_i$ when coupled with $\mathbf{X}_j$ were different from what was observed when coupled with $\mathbf{X}_{j'}$, then the hypothesis of constant means and variances would be violated. The restriction of Locality Condition II must be satisfied in the construction of λ and it is easy to check that it is.

These remarks are summarized in Theorem 2.

THEOREM 2. *Let n random variables $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be given satisfying the hypothesis of Theorem 3. Let M_i be the vector of local parameters for measuring $\mathbf{X}_i$, and let each $\mathbf{X}_i$ satisfy Locality Condition I. Then there is a hidden variable λ satisfying Locality Condition II and the Second-Order Factorization Condition if there is a joint probability distribution of $\mathbf{X}_1, \ldots, \mathbf{X}_n$.*

The next theorem states two conditions equivalent to an inequality condition given in Suppes and Zanotti (1981) for three random variables having just two values.

THEOREM 3. *Let three random variables $\mathbf{X}, \mathbf{Y}$ and $\mathbf{Z}$ be given with values ± 1 satisfying the symmetry condition $E(\mathbf{X}) = E(\mathbf{Y}) = E(\mathbf{Z}) = 0$ and with covariances $E(\mathbf{XY}), E(\mathbf{YZ})$ and $E(\mathbf{XZ})$ given. Then the following three conditions are equivalent.*

(*i*) *There is a hidden variable λ satisfying Locality Condition II and equation (a) of the Second-Order Factorization Condition holds.*

(*ii*) *There is a joint probability distribution of the random variables $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$ compatible with the given means and covariances.*

(*iii*) *The random variables $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$ satisfy the following inequalities.*

$$-1 \leq E(\mathbf{XY}) + E(\mathbf{YZ}) + E(\mathbf{XZ})$$
$$\leq 1 + 2\text{Min}(E(\mathbf{XY}), E(\mathbf{YZ}), E(\mathbf{XZ})).$$

There are several remarks to be made about this theorem, especially the inequalities given in (iii). A first point is how do these inequalities relate to Bell's well-known inequality (Bell 1964).

$$1 + E(\mathbf{YZ}) \geq \mid E(\mathbf{XY}) - E(\mathbf{XZ}) \mid . \tag{13}$$

Bell's inequality is in fact neither necessary nor sufficient for the existence of a joint probability distribution of the random variables $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$ with values ± 1 and expectations equal to zero.

The next well-known theorem states two conditions equivalent to Bell's Inequalities for random variables with just two values. This form of the inequalities is due to Clauser et al. (1969), referred to as CHSH. The equivalence of (ii) and (iii) is due to Fine (1982).

THEOREM 4. (BELL'S INEQUALITIES) *Let n random variables be given satisfying the locality hypothesis of Theorem 4. Let $n = 4$, the number of random variables, let each $\mathbf{X}_t$ be discrete with values ± 1, let the symmetry condition $E(\mathbf{X}_i) = 0$, $i = 1, \ldots, 4$ be satisfied, let $\mathbf{X}_1 = \mathbf{A}$, $\mathbf{X}_2 = \mathbf{A}'$, $\mathbf{X}_3 = \mathbf{B}$, $\mathbf{X}_4 = \mathbf{B}'$, with the covariances $E(\mathbf{AB})$, $E(\mathbf{AB}')$, $E(\mathbf{A}'\mathbf{B})$ and $E(\mathbf{A}'\mathbf{B}')$ given. Then the following three conditions are equivalent.*

(*i*) *There is a hidden variable λ satisfying Locality Condition II and equation (a) of the Second-Order Factorization Condition holds.*

(*ii*) *There is a joint probability distribution of the random variables $\mathbf{A}$, $\mathbf{A}'$, $\mathbf{B}$ and $\mathbf{B}'$ compatible with the given means and covariances.*

(*iii*) *The random variables $\mathbf{A}$, $\mathbf{A}'$, $\mathbf{B}$ and $\mathbf{B}'$ satisfy Bell's inequalities in the CHSH form*

$$-2 \leq E(\mathbf{AB}) + E(\mathbf{AB}') + E(\mathbf{A}'\mathbf{B}) - E(\mathbf{A}'\mathbf{B}') \leq 2$$

$$-2 \leq E(\mathbf{AB}) + E(\mathbf{AB}') - E(\mathbf{A}'\mathbf{B}) + E(\mathbf{A}'\mathbf{B}') \leq 2$$

$$-2 \leq E(\mathbf{AB}) - E(\mathbf{AB}') + E(\mathbf{A}'\mathbf{B}) + E(\mathbf{A}'\mathbf{B}') \leq 2$$

$$-2 \leq -E(\mathbf{AB}) + E(\mathbf{AB}') + E(\mathbf{A}'\mathbf{B}) + E(\mathbf{A}'\mathbf{B}') \leq 2.$$

GHZ-type Experiments

Changing the focus, I now first consider GHZ-type experiments. Good references are Greenberger, Horne and Zeilinger (1989), the more extended discussion in Greenberger, Horne, Shimony and Zeilinger (1990) and Mermin (1990).

I follow the quantum-mechanical argument given in Mermin (1990). We start with the three-particle entangled state

$$|\psi\rangle = \frac{1}{\sqrt{2}}(|+\rangle_1|+\rangle_2|+\rangle_3 + |-\rangle_1|-\rangle_2|-\rangle_3), \tag{14}$$

This state is an eigenstate of the following spin operators:

$$\hat{\mathbf{A}} = \hat{\sigma}_{1x}\hat{\sigma}_{2y}\hat{\sigma}_{3y}, \quad \hat{\mathbf{B}} = \hat{\sigma}_{1y}\hat{\sigma}_{2x}\hat{\sigma}_{3y}, \tag{15}$$

$$\hat{\mathbf{C}} = \hat{\sigma}_{1y}\hat{\sigma}_{2y}\hat{\sigma}_{3x}, \quad \hat{\mathbf{D}} = \hat{\sigma}_{1x}\hat{\sigma}_{2x}\hat{\sigma}_{3x}. \tag{16}$$

If we compute quantum mechanically the expected values for the correlations above, we obtain at once that $E_Q(\hat{\mathbf{A}}) = E_Q(\hat{\mathbf{B}}) = E_Q(\hat{\mathbf{C}}) = 1$ and $E_Q(\hat{\mathbf{D}}) = -1$. (To exhibit all the details of this setup is too lengthy to include here, but the argument is elementary and standard, in the context of quantum mechanics.)

Now we note that

$$E_Q(\mathbf{ABC}) = (s_{1x}s_{2y}s_{3y})(s_{1y}s_{2x}s_{3y})(s_{1y}s_{2y}s_{3x}) \tag{17}$$

$$= s_{1x}s_{2x}s_{3x}(s_{1y}^2 s_{2y}^2 s_{3y}^2), \tag{18}$$

but since the s_{ij} can only be 1 or -1, we obtain at once that

$$\hat{s}_{1y}^2 = s_{2y}^2 = s_{3y}^2 = 1, \textit{and} \tag{19}$$

$$E_Q(\mathbf{ABC}) = s_{1x}s_{2x}s_{3x} = E_Q(\hat{\mathbf{D}}) = -1. \tag{20}$$

In stark contrast, we have the following elementary theorem of classical probability.

THEOREM 5. *Let* $\mathbf{A}, \mathbf{B}$, *and* $\mathbf{C}$ *be* ± 1 *random variables having a joint probability distribution such that* $E(\mathbf{A}) = E(\mathbf{B}) = E(\mathbf{C}) = 1$. *Then* $E(\mathbf{ABC}) = 1$.

Proof. Since $E(A) = 1$, $P(\overline{a}) = P(\overline{a}bc) = P(\overline{a}b\overline{c}) = P(\overline{a}\overline{b}c) = P(\overline{abc}) = 0$, where $P(\overline{a}bc) = P(\mathbf{A} = -1, \mathbf{B} = 1, \mathbf{C} = 1)$, etc. By similar argument for $E(\mathbf{B})$ and $E(\mathbf{C})$, we are left with $P(abc) = 1$, which implies at once the desired result.

So, rather than inequalities, we have a flat contradiction. Classically

$$E(\mathbf{ABC}) = 1.$$

but, as shown above, quantum mechanically

$$E_Q(\mathbf{ABC}) = -1$$

Of course, using now also Theorem 1 we infer at once that there can be no factoring hidden variable for the quantum mechanical case.

This striking characteristic of GHZ's theoretical predictions, however, has a major problem. How can one verify experimentally predictions based on probability-one statements, since experimentally one cannot in the relevant experiments obtain events perfectly correlated? Fortunately, the correlations present in the GHZ state are so strong that even if we allow for experimental errors, the non-existence of a joint distribution can still be verified, as we show in the following theorem and its corollary.

THEOREM 6. (deBarros and Suppes 2000) *If* $\mathbf{A}$, $\mathbf{B}$, *and* $\mathbf{C}$ *are three* ± 1 *random variables, a joint probability distribution exists for the given expectations* $E(\mathbf{A})$, $E(\mathbf{B})$, $E(\mathbf{C})$, *and* $E(\mathbf{ABC})$ *if and only if the following inequalities are satisfied:*

$$-2 \leq E(\mathbf{A}) + E(\mathbf{B}) + E(\mathbf{C}) - E(\mathbf{ABC}) \leq 2, \tag{21}$$

$$-2 \leq E(\mathbf{A}) + E(\mathbf{B}) - E(\mathbf{C}) + E(\mathbf{ABC}) \leq 2, \tag{22}$$

$$-2 \leq E(\mathbf{A}) - E(\mathbf{B}) + E(\mathbf{C}) + E(\mathbf{ABC}) \leq 2, \tag{23}$$

$$-2 \leq -E(\mathbf{A}) + E(\mathbf{B}) + E(\mathbf{C}) + E(\mathbf{ABC}) \leq 2. \tag{24}$$

Proof. First we prove necessity. Let us assume that there is a joint probability distribution consisting of the eight atoms abc, $ab\overline{c}$, $a\overline{b}c$, ...$\overline{a}\overline{b}\overline{c}$. Then,

$$E(\mathbf{A}) = P(a) - P(\overline{a}),$$

where

$$P(a) = P(abc) + P(a\overline{b}c) + P(ab\overline{c}) + P(a\overline{b}\overline{c}),$$

and

$$P(\overline{a}) = P(\overline{a}bc) + P(\overline{a}\overline{b}c) + P(\overline{a}b\overline{c}) + P(\overline{a}\overline{b}\overline{c}).$$

Similar equations hold for $E(\mathbf{B})$ and $E(\mathbf{C})$. For $E(\mathbf{ABC})$ we obtain

$$\begin{aligned} E(\mathbf{ABC}) &= P(\mathbf{ABC}=1) - P(\mathbf{ABC}=-1) \\ &= P(abc) + P(a\bar{b}\bar{c}) + +P(\bar{a}\bar{b}c) + P(\bar{a}b\bar{c}) \\ &\quad -[P(a\bar{b}c) + P(ab\bar{c}) + P(\bar{a}bc) + P(\bar{a}\bar{b}\bar{c})]. \end{aligned}$$

Corresponding to the first inequality above, we now sum over the probability expressions for the expectations

$$F = E(\mathbf{A}) + E(\mathbf{B}) + E(\mathbf{C}) - E(\mathbf{ABC}),$$

and obtain the expression

$$\begin{aligned} F &= 2[P(abc) + P(\bar{a}bc) + P(a\bar{b}c) + P(ab\bar{c})] \\ &\quad -2[P(\bar{a}\bar{b}\bar{c}) + P(\bar{a}\bar{b}c) + P(\bar{a}b\bar{c}) + P(a\bar{b}\bar{c})], \end{aligned}$$

and since all the probabilities are nonnegative and sum to ≤ 1, we infer at once inequality (21). The derivation of the other three inequalities is very similar.

To prove the converse, i.e., that these inequalities imply the existence of a joint probability distribution, is slightly more complicated. We restrict ourselves to the symmetric case

$$P(a) = P(b) = P(c) = p,$$

$$P(\mathbf{ABC}=1) = q$$

and thus

$$E(\mathbf{A}) = E(\mathbf{B}) = E(\mathbf{C}) = 2p - 1,$$

$$E(\mathbf{ABC}) = 2q - 1.$$

In this case, (21) can be written as

$$0 \leq 3p - q \leq 2,$$

while the other three inequalities yield just $0 \leq p + q \leq 2$. Let

$$x = P(\bar{a}bc) = P(a\bar{b}c) = P(ab\bar{c}),$$

$$y = P(\bar{a}\bar{b}c) = P(\bar{a}b\bar{c}) = P(a\bar{b}\bar{c}),$$

$$z = P(abc),$$

and

$$w = P(\bar{a}\bar{b}\bar{c}).$$

It is easy to show that on the boundary $3p = q$ defined by the inequalities the values $x = 0$, $y = q/3$, $z = 0$, $w = 1 - q$ define a possible joint probability distribution, since $3x + 3y + z + w = 1$. On the other boundary, $3p = q + 2$ so a possible joint distribution is $x = (1 - q)/3$, $y = 0$, $z = q$, $w = 0$. Then, for any values of q and p within the boundaries of the inequality we can take a linear combination of these distributions with weights $(3p - q)/2$ and $1 - (3p - q)/2$, chosen such that the weighed probabilities add to one, and obtain the joint probability distribution:

$$\begin{aligned}
\mathbf{x} &= \left(1 - \frac{3p - q}{2}\right)\frac{1 - q}{3}, \\
y &= \left(\frac{3p - q}{2}\right)\frac{q}{3}, \\
z &= \left(1 - \frac{3p - q}{2}\right)q, \\
w &= \frac{3p - q}{2}(1 - q),
\end{aligned}$$

which proves that if the inequalities are satisfied a joint probability distribution exists, and therefore a noncontextual hidden variable as well, thus completing the proof. The generalization to the asymmetric case is tedious but straightforward.

As a consequence of the inequalities above, the correlations present in the GHZ state can be so strong that even if we allow for experimental errors, the non-existence of a joint distribution can still be verified (deBarros and Suppes 2000), as is shown in the following.

COROLLARY 1. *Let* $\mathbf{A}$, $\mathbf{B}$, *and* $\mathbf{C}$ *be three* ± 1 *random variables such that*
(i) $E(\mathbf{A}) = E(\mathbf{B}) = E(\mathbf{C}) \geq 1 - \epsilon$,
(ii) $E(\mathbf{ABC}) \leq -1 + \epsilon$,
where ϵ *represents a decrease of the observed GHZ correlations due to experimental errors. Then, there cannot exist a joint probability distribution of* $\mathbf{A}$, $\mathbf{B}$, *and* $\mathbf{C}$ *if*

$$\epsilon < \frac{1}{2}. \tag{25}$$

Proof. To see this, let us compute the value of F define above. We obtain at once that

$$F = 3(1 - \epsilon) - (-1 + \epsilon).$$

But the observed correlations are only compatible with a noncontextual hidden variable theory if $F \leq 2$, hence $\epsilon < \frac{1}{2}$. Then, there cannot exist a joint probability distribution of $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ satisfying (i) and (ii) if

$$\epsilon < \frac{1}{2}. \tag{26}$$

From the inequality obtained above, it is clear that any experiment that obtains GHZ-type correlations stronger than 0.5 cannot have a joint probability distribution. For example, the recent experiment made at Innsbruck (Bouwmeester et al. 1999) with three-photon entangled states supports the quantum mechanical result that no noncontextual hidden variable exists that explains their correlations. Thus, with this reformulation of the GHZ theorem it is possible to use strong, yet imperfect, experimental correlations to prove that a noncontextual hidden-variable theory is incompatible with the experimental results.

On the other hand, as is shown in de Barros and Suppes (2000), the mean result of the Innsbruck experiment is not far from the classical regime. The distance is slightly less than two standard deviations from the classical boundary, so a more refined experiment, with mean results further from the boundary, would be desirable as a next step, and should be possible without any major technological changes in the experimental instruments.

Ventura Hall, Stanford University
Stanford, California

REFERENCES

de Barros, J. A. and Suppes, P. 2000. "Inequalities for Dealing with Detector Inefficiencies in Greenberger-Horne-Zeilinger-Type Experiments". *Physical Review Letters* 84: 793–797.

Bell, J. S. 1964. "On the Einstein-Podolsky-Rosen Paradox". *Physics* 1: 195–200.

Berkeley, G 1901. "An Essay Towards a New Theory of Vision". In A. C. Fraser (ed.), *Berkeley's Complete Works*. London: Oxford University Press, vol. 1, pp. 93–210. First published in 1709.

Bouwmeester, D., Pan, J. W., Daniell, M., Weinfurter, H. and A., Z. 1999. "Observation of Three-Photon Greenberger-Horne-Zeilinger Entanglement". *Physical Review Letters* 82: 13–45.

Caton, R. 1875. "The Electric Currents of the Brain". *British Medical Journal* 2: 278.

Clauser, J. F., Horne, J. F., Shimony, A. and Holt, R. A. 1969. "Proposed Experiment to Test Local Hidden-Variable Theories". *Physical Review Letters* 23: 880–884.

Cooley, J. W. and Tukey, J. W. 1965. "An Algorithm for the Machine Computation of Complex Fourier Series". *Math. Computation* 19: 297–301.

Du Bois-Reymond, E. 1848. *Untersuchungen über Thierische Elektricität.* Berlin: Verlang von G. Reiner. Passage quoted translated by Hebbel E. Hoff, "Galvani and the Pre-Galvanic Electrophysiologists". *Annals of Science* 1 (1936), 157–172.

Fine, A. 1982. "Hidden Variables, Joint Probability, and the Bell Inequalities". *Physical Review Letters* 48: 291–295.

Galison, P. 1987. *How Experiments End.* Chicago: University of Chicago Press.

Galvani, L. 1791. "De Viribus Electricitatis in Motu Musculari". *De Bononiensi Scientiarum et Artium Instituto atque Academia, Comm.* 7: 363–418. Translated by Margaret Glover Foley, as *Luigi Galvani, Commentary on the Effects of Electricity on Muscular Motion.* Notes and introduction by I. Bernard Cohen. Norwalk, Connecticut, USA: Burndy Library, 1953.

Galvani, L. 1794. *Dell'uso, e dell'attività, dell'arco conduttore nelle contrazioni dei muscoli.* Bologna: A. S. Tommaso d'Aquino. Published anonymously.

Greenberger, D. M., Horne, M. A., Shimony, A. and Zeilinger, A. 1990. "Bell's Theorem without Inequalities". *American Journal of Physics* 58: 1131–1143.

Greenberger, D. M., Horne, M. A and Zeilinger, A. 1989. "Going Beyond Bell's Theorem". In M. Kafatos (ed.), *Bell's Theorem, Quantum Theory, and Conceptions of the Universe,* Dordrecht: Kluwer Academic Press.

Holland, P. W. and Rosenbaum, P. R 1986. "Conditional Association and Unidimensionality in Monotone Latent Trait Models". *Annals of Statistics* 14: 1523–1543.

Hume, D 1739. *A Treatise of Human Nature.* London: John Noon. Quotations taken from L A. Selby-Bigge's edition (1951), Oxford University Press, London.

Matteucci, C. 1844 *Traité des Phenomènes Electrophysiologiques des Animaux.* Paris: Fortin Masson.

Mermin, N D. 1990. "Quantum Mysteries Revisited". *American Journal of Physics* 58(8): 731–734.

Oppenheim, A. V. and Schafer, R. W. 1975. *Digital Signal Processing.* Englewood Cliffs, NJ: Prentice-Hall.

Redi, F. 1671. *Esperienze intorno a diverse cose naturali, e particolarmente a quelle, che ci son portate dall'Indie.* Florence: Piero Matini. First published in 1671.

Rubin, G. S. and Turano, K. 1992. "Reading without Saccadic Eye Movements". *Vision Research* 32(5): 895–902.

Rugg, M. D. and Coles, M. G. 1995. *Electrophysiology of Mind: Event-Related Brain Potentials and Cognition.* New York: Oxford University Press.

Suppes, P., de Barros, J. A. and Oas, G. 1998a. "A Collection of Probabilistic Hidden-Variable Theorems and Counterexamples". In R. Pratesi and L. Ronchi (eds.), *Waves, information and foundations of physics. Conference proceedings, Vol. 60*, Bologna: Società Italiana Di Fisica, pp. 267–291.

Suppes, P., Han, B., Epelboim, J. and Lu, Z. L. 1999a. "Invariance Between Subjects of Brain Wave Representations of Language". *Proceedings of the United States National Academy of Sciences* 96: 12953–12958.

Suppes, P., Han, B., Epelboim, J. and Lu, Z. L. 1999b. "Invariance of Brain-wave Representations of Simple Visual Images and Their Names". *Proceedings of the United States National Academy of Sciences* 96: 14658–14663.

Suppes, P., Han, B. and Lu, Z. L. 1998b. "Brain-wave Recognition of Sentences". *Proceedings of the National Academy of Sciences* 95: 15861–15866.

Suppes, P., Lu, Z. L. and Han, B. 1997. "Brain-wave Representations of Words". *Proceedings of the United States National Academy of Sciences* 94: 14965–14969.

Suppes, P., Wong, D. K., Perreau-Guimaraes, M., Uy, E. T. and Yang, W. Forthcoming. "High Statistical Recognition Rates for Some Persons' Brain-wave Representations of Sentences" .

Suppes, P. and Zanotti, M. 1976. "Necessary and Sufficient Conditions for Existence of a Unique Measure Strictly Agreeing with a Qualitative Probability Ordering". *Journal of Philosophical Logic* 5: 431–438.

Suppes, P. and Zanotti, M. 1981. "When Are Probabilistic Explanations Possible?" *Synthese* 48: 191–199.

Volta, A. 1793/1918. "Letter to Tiberius Cavallo, 22 May 1793". In *Le opere di Alessandro Volta*. Milan: Ulrico Hoepli, vol. 1, pp. 203–208.

Whittaker, E. T. 1951. *The History of the Theories of Aether and Electricity*, vol. 1. London: Nelson.