Testing Dowsing

The Failure of the Munich Experiments

German physicists concluded from their massive experimental study that water dowsers unquestionably have a remarkable, mysterious skill. Those results, however, provide the most convincing disproof *imaginable that dowsers can do what they claim.*

J. T. ENRIGHT

The notion that certain skilled individuals can discover underground water by using a mysterious talent known as "dowsing" (or "witching" or "divining") is widely regarded among serious scientists as no more than a superstitious relic from medieval times. No plausible physical or physiological mechanism has ever been proposed by which such detection might be possible. Nevertheless, the worldwide persistence of this practice through the centuries might lead open-minded people to wonder whether there could be a germ of truth behind the folklore. After all, valuable additions to the modern pharmacopoeia have sometimes been derived from folk medicine, thus proving that not all folklore is unmitigated superstition.

Planning an Experimental Study

Many dowsers in Germany think that the stimuli to which they claim to respond ("earthrays," said to be a subtle form of radiation not otherwise known to science) are potentially hazardous to human health, perhaps even inducing cancer. Hence, in the middle 1980s, the German government brought together a committee to consider how a proper study might be conducted to investigate the possibility that dowsing is a genuine skill. If dowsers can indeed detect (dangerous?) radiation, perhaps they might be able to contribute to research in public health issues.

The outcome of those deliberations was a grant of 400,000 German marks (about $250,000), in 1986, to university physicists in Munich. Generous funding assures a large-scale project, so that even weak effects might become evident; the reputation of university-based researchers for open-minded integrity means that their participation provides credibility that a project managed only by dowsers themselves would not have.

For an open-minded test of claimed extraordinary abilities, the claimants deserve a fair opportunity for success by providing conditions they regard as suitable, and in this regard, the Munich researchers seem to have bent over backward. Experiments designed only by doubters might, of course, leave dowsers with convenient reasons to discount a disappointing outcome. Enthusiasts for dowsing were therefore involved in the planning sessions. When practitioners of various occult "skills" have, in the past, been unsuccessful under controlled testing, they have at times claimed that the research was conducted in a skeptical (by implication, hostile) atmosphere, which interfered with their performances and invalidated the studies. That potential problem did not arise in the Munich experiments because the principal investigators, from the University of Munich and the Technical University of Munich, had publicly gone on record as thinking that dowsing is probably a genuine phenomenon. No hostility there!

Water dowsing ordinarily takes place out of doors, and this raises potential difficulties for meaningful experiments, because no two outdoor locations can be considered fully equivalent replicates; and the essence of proper scientific research is replicated testing to examine reproducibility. Most German dowsing practitioners, however, also claim to be able to dowse the location of water piping in a garden or within a structure, so indoor testing was decided upon.

Another potential problem is that among those who *think* that they have dowsing skill, some may be mistaken or perhaps are even deliberate frauds. To avoid these potential pitfalls, some 500 candidate dowsers were recruited for preliminary testing. That group was winnowed down to forty-three individuals for the final, critical experiments: those who seemed to be most successful in the preliminary tests. Those dowsers all freely participated in the carefully controlled final experiments, which they accepted as suitable to their abilities. There could thus be no basis for subsequent claims that the test program was inappropriate or unfair.

J.T. (Jim) Enright is a professor of behavioral physiology, Scripps Institution of Oceanography, University of California at San Diego, La Jolla, CA 92093, where he emphasizes data analysis in teaching critical evaluation of scientific literature. He has conducted research on "biological clocks" and sensory physiology of both crustaceans and humans and has spent several years at research labs in Germany and Austria.

Experimental Design

The detailed test procedure was a very simple one. On the ground floor of a large vacant barn near Munich, a ten-meter test line was established, along which a small wagon could be moved; and atop the wagon was a short length of pipe, perpendicular to the line and connected by hoses to a pump that could provide circulating water. Circulating water was chosen rather than still water because the traditions of dowsers postulate that useful underground water supplies are mainly to be found as flowing streams that they refer to as "water arteries" and not just within extensive deposits of permeable sediment, as geologists would tell them. The location of the pipe for each single test was to be determined by a computer-generated random number (although the settings actually used turned out to be decidedly nonrandomly located along the line).

On the upper floor of the barn, directly above the experimental line, a ten-meter test line was established. For the critical final experiments, a dowser was re-admitted to the upper-floor arena each time that the pipe had been repositioned, and was required, with his or her witching stick (or pendulum or other tool of choice), to guess where the pipe on the ground floor was located. A given dowser was tested in a sequence of from 5 to 15 single tests (typically 10), which typically took about an hour. During the two-year program in the barn, the forty-three selected dowsers participated in 843 single tests, grouped into 104 test-series of this sort. Some dowsers undertook only a single test series, selected others underwent more than ten test series.

It would seem that such indoor testing should appreciably simplify the dowsers' task. Out of doors, the critical stimuli might be deflected or refracted by intervening layers of soil and rock, but in the barn, the only obstruction was the flooring between stories. Furthermore, in an outdoor setting, the detection of "water arteries," as dowsers envision them, should require remarkable precision. If, say, a 3-meter-diameter stream of water were to be located at a depth of 100 meters, the dowser must achieve precision of less than 1° around the vertical in determining the point of maximal stimuli for drilling, and this apparently implies detection of minuscule changes in stimulus direction and/or intensity. Comparable 1°-precision around the vertical in the barn, however, with the target only, say, about five meters away, would result in uncertainty of less than 10 centimeters around the pipe's actual location.

Before the experiments began, a professional magician was brought in to inspect the entire arrangement for the potential for deception or cheating by the dowsers. As an additional precaution against cheating (such as peeking through cracks in the floor), an experimenter/observer was also present to supervise

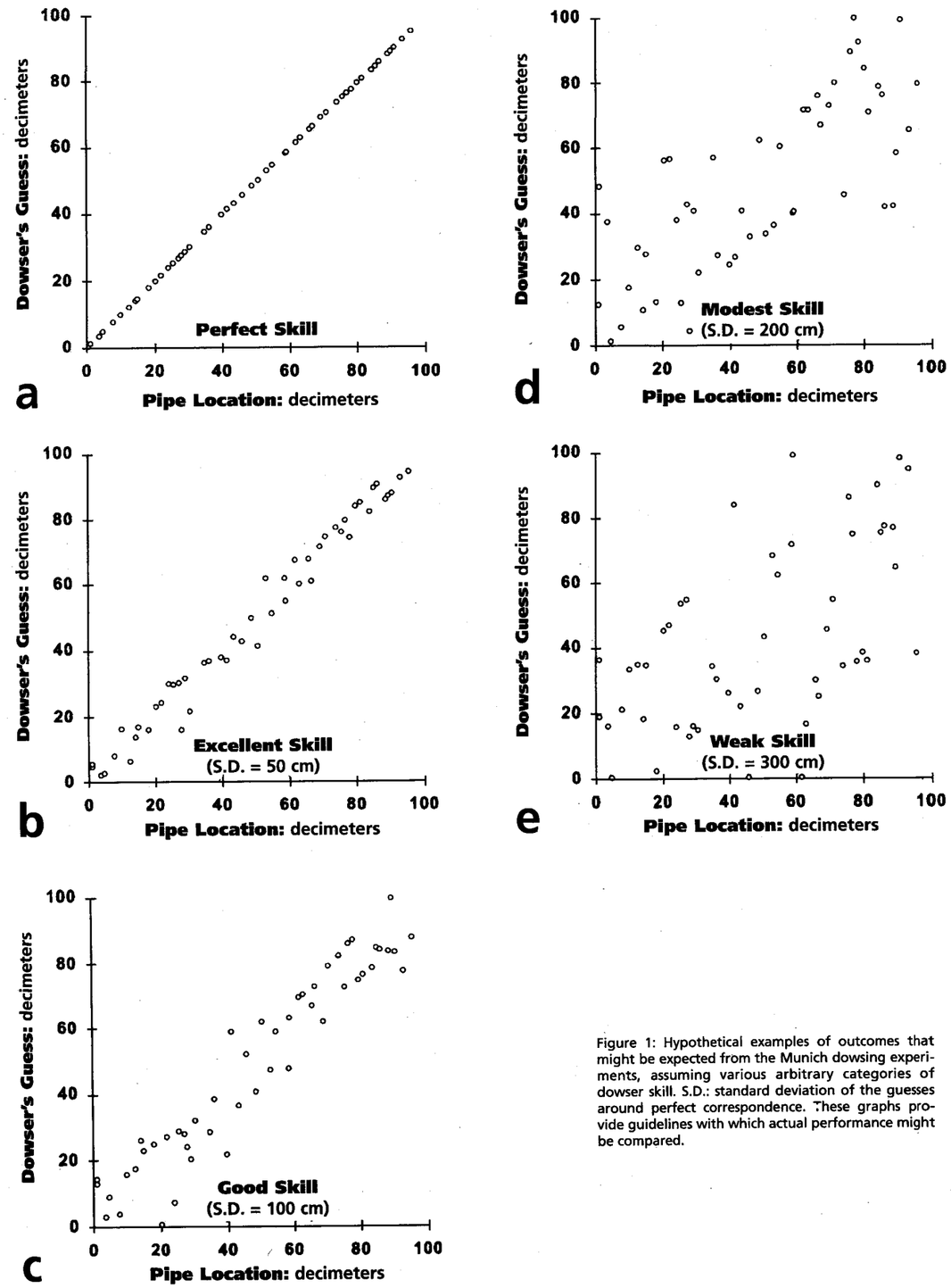

Figure 1: Hypothetical examples of outcomes that might be expected from the Munich dowsing experiments, assuming various arbitrary categories of dowser skill. S.D.: standard deviation of the guesses around perfect correspondence. These graphs provide guidelines with which actual performance might be compared.

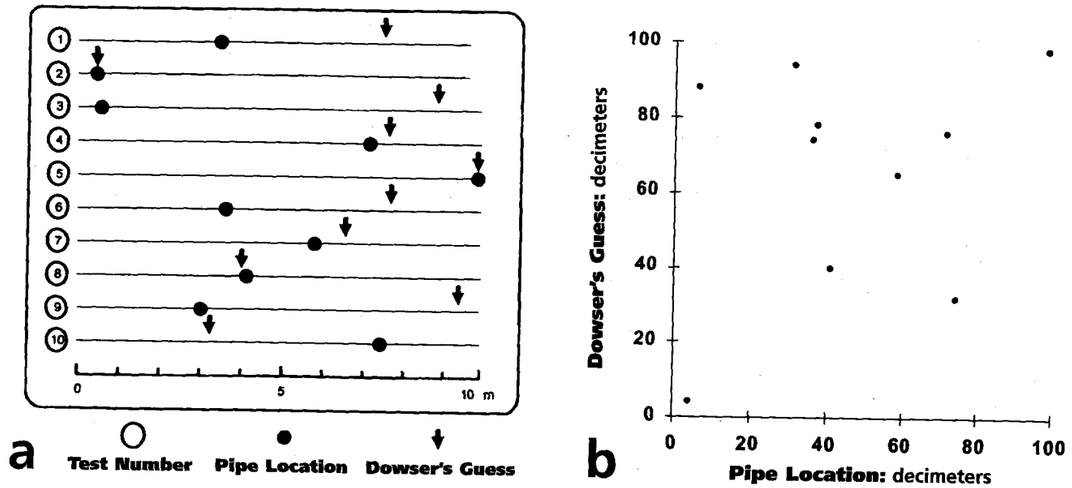

Figure 2: Dowsing results from the second of the four test series undertaken by dowser #99. These results were evaluated by the researchers as the "best" of all 104 series undertaken, and represent the only set of dowsing data from the barn presented graphically in the final research report. A: presentation in the format used by Wagner, Betz, and König (1990); b: presentation in the format of Figure 1 here.

the dowsers' performances, and to record the guesses. Double-blind procedures assured that neither the observer nor the dowser knew the pipe's location, even after a guess had been made; thus, no feedback was provided during the critical testing.

The study involved many thousands of preliminary tests, in which the careful controls of the final critical experiments were relaxed. Often, for example, feedback about success or failure was given in those preliminary tests. Sometimes the pipe was filled with fresh water, sometimes salt water, sometimes even empty; sometimes flow was turbulent, sometimes not; sometimes sand or gravel was mixed in with the water, and so on. The preliminary testing had two purposes: as indicated above, to eliminate those candidates whose trials showed no appreciable dowsing skill (more than 90 percent of the candidates!); and to choose for the selected participants those aspects of the preliminary tests (fluid, flow rate, etc.) that had led to their best initial results. Each individual's final critical testing could thus be based on his or her "optimal stimuli."

Before a critical test series, each dowser was asked to determine whether there were any places along the test line (without pipe present) that seemed to provide stimuli that could be mistaken for the target (presumably indicating natural sources of "earth rays"). In quite a large number of cases, two or three such locations were reported along the 10-meter line. "Earth rays" are seemingly nearly everywhere! When such non-target stimuli were reported, the surrounding regions of the line (typically one meter wide) were then excluded for that dowser's test series as potential test locations. (A given dowser often reported different artifact locations on different days; natural sources of "earth-ray" stimuli are apparently transient.)

An ideal experimental design was frequently compromised, because two dowsers arrived at the barn at the same time. Instead of testing those individuals one after the other, the two dowsers were tested alternately, each pipe setting being used twice in succession. It was assumed that their guesses could be treated as independent because the two individual dowsers were not simultaneously present in the test arena.

If a dowser felt that his or her concentration was waning during testing, the test series could be interrupted or terminated, which apparently happened quite often. Thus, it seems quite obvious that many accommodations were made to the wishes and whims of the dowsers and the experimenters. Nevertheless, many aspects of sound experimental design were built into the critical testing: double-blind protocol, no feedback about success or failure, randomized (well, sort of!) pipe settings, replicated testing of the same dowsers on different days, and a large-scale program (843 critical tests) so that small sets of "good" results would not deserve undue attention.

It is conceivable that the noise of water turbulence (sometimes with gravel in the water) could have provided localized auditory information during the testing. Another concern is that the experimenter supervising the dowsers may not have been properly "blinded," when aware that identical pipe settings were being used for two dowsers in alternation. If truly remarkable dowsing success had been achieved in the experiments, such concerns would deserve careful attention. In fact, however, the overall negative outcome suggests that any residual defects in the experimental design usually had no important impact on the outcome.

Expectations

Proper planning requires that one consider in advance what sort of results might arise from experiments of this design, if dowsing were to be a real, reproducible phenomenon. Several examples of hypothetical outcome are shown in Figure 1.

"Perfect" skill (the equivalent of using Superman's X-ray vision to look through the flooring) would lead to a perfect correlation between the dowsers' guesses and pipe locations (Figure 1a); weak skill (with a standard deviation of 3 meters along a 10-meter test line) would produce broad scatter around the same diagonal line (Figure 1e), and intermediate skill levels would involve lesser scatter about that diagonal line in this kind of plot. Note that while the "Weak-Skill" results are very scattered, they do not look random (nor are they: $r = 0.57$, $p < 0.001$) because of the vacant regions of the graph in the upper-left and lower-right corners, an issue that arises again below.

Results and Interpretation by the Experimenters

In the final report on their dowsing experiments, submitted to the granting agency (Wagner, Betz, and König 1990), the researchers concluded that most dowsers did not do particularly well in the experiments. That report, however, still painted a very positive picture of the overall outcome. The following quotation is a translation of the German text:

> Some few dowsers, in particular tasks, showed an extraordinarily high rate of success, which can scarcely if at all be explained as due to chance . . . a real core of dowser-phenomena can be regarded as empirically proven . . . (5)

The evidence provided for this interpretation consisted of a plot of results from a single test series (out of 104 available), data shown here as Figure 2; and a table summarizing the purported statistical significance of each of the 104 test series. (This summary is based on nonstandard statistical methods that were conspicuously fitted to the data. More conventional statistical tests suggest less interesting conclusions [Enright 1995].) The peculiar plot in the report (Figure 2a) gives the visual impression of very good correspondence between observed and expected results. The re-plot in Figure 2b places those data in a more revealing context. Half of the results in Figure 2b (5 tests of 10) do indeed resemble an ideal hypothetical outcome (Figure 1a or 1b), but it deserves emphasis that Figure 2 cannot be considered "typical" but instead represents the very "best" results, consisting only of ten tests out of 843, from one test series out of 104. (In 843 spins of a roulette wheel, at least one sequence of 10 results that includes several seemingly exceptional events might be expected to arise by chance alone.)

A Broader Look at *All* the Data

Presented in Figure 3 is a plot of all 843 test results. The human eye is remarkably adept at detecting pattern in plots like this. Note, for example, groups of points that seem to follow curvy lines through certain regions of the graph. Resemblance to the expectations of Figure 1, however, is decidedly absent in Figure 3. Instead, the visual impression is that these results seem to be distributed more or less at random. In order to examine that interpretation, the actual results can be compared with the outcome of genuine randomization. To that end, the dowsers' actual choices were randomly paired with pipe settings from other test series. (The x and y values of

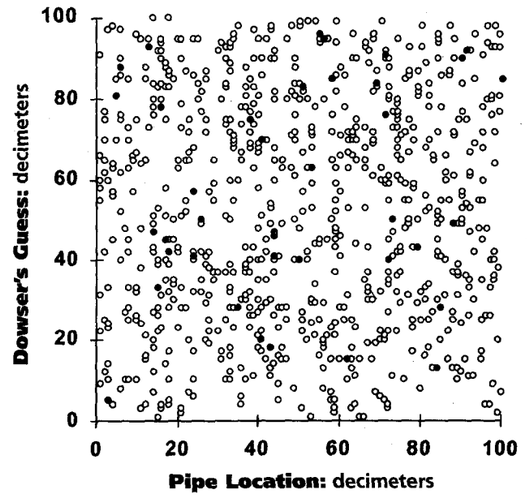

Figure 3: Results from all of the 843 tests in the Munich barn, plotted in coordinates like those of Figure 1. Filled symbols represent two data points at identical coordinates.

points shown in Figure 3 were thus randomly intermixed.) The results of two such randomizations are shown in Figures 4a and 4b. It would be difficult to defend a claim that the actual results (Figure 3) show better concordance with expectations (Figure 1) than do the randomizations of Figure 4.

To depend on what seems evident by inspection, however, may seem like an unrigorous approach. As an elementary form of quantification, the coordinate system of Figure 3 can be divided into squares, each 2.5 meters on a side, and the enclosed data points counted. Some of those counts are presented in Figure 5. Recalling that even weak skill should produce very few observations in the upper left and lower right regions of such a graph (Figure 1e), those two corner quadrats in Figure 5 can be summed. The total (90) turns out to be greater than the sum of counts in the upper right and lower left (87), so these counting data could even be interpreted, if one were so inclined, as suggestive of weak *anti*-dowsing skill.

A Few Unusually Talented Individuals?

The researchers in the Munich study would probably protest against this treatment of the data by noting that outstanding performances like these shown in Figure 2 have been obscured by results from unskilled candidates. That objection falters when one recalls that Figure 3 includes only the final tests of those forty-three dowsers (out of some 500 candidates) who were selected on the basis of preliminary testing as being the most skillful. Nevertheless, the possibility of unusual skill by only a very few individuals deserves careful scrutiny. In the tabulation of the final report, there were two other test series (in addition to that shown in Figure 2) that the researchers themselves evaluated as being particularly impressive. Results from all

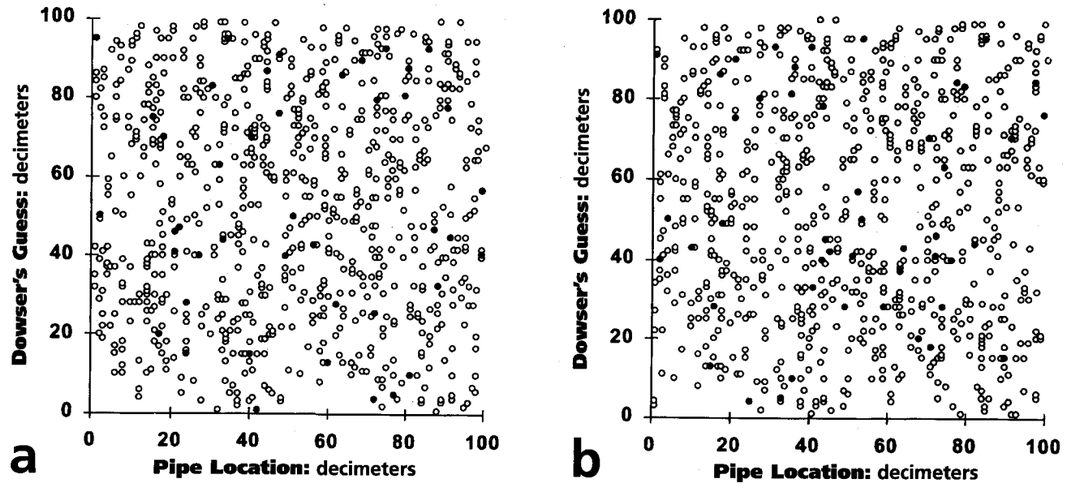

Figure 4: A and b: Two sets of results from complete randomizations. Dowsers' guesses were paired randomly with pipe locations from other test series. Filled symbols represent two data points at identical coordinates.

three of those test series are presented in Figure 6a. Furthermore, there were four other test series (from three other dowsers) that were considered by the researchers also to indicate lesser but nonetheless remarkable skill; those results are summarized in Figure 6b. These two graphs of the very "best" test series present the outcome of the entire research program in its most favorable light. Despite many errors, there are indeed an impressive number of guesses that were not far from the pipe's actual location.

Do those results justify the assertion of the final report that "some few dowsers" were remarkably successful? Decidedly not! Each of the six dowsers who contributed to the data shown in Figures 6a and 6b participated in other test series, and the outcomes of those replicated series by those same dowsers (Figures 6c and 6d) seem to be quite unimpressive: just as scattered as

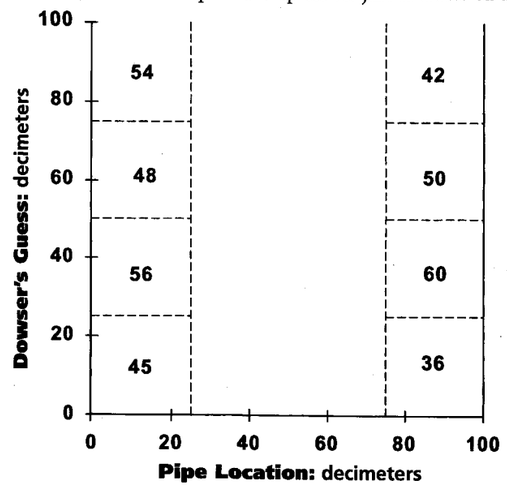

Figure 5: Numbers of dowser guesses in selected quadrats of Figure 3, based on subdivision of the coordinate system into squares, each 250 centimeters on a side.

the overall outcome (Figure 3). What it amounts to, then, is that among the 104 test series, there were several that seemed somewhat interesting, but those dowsers responsible could not reproduce that kind of result in other comparable tests. And seven series out of 104 (each with a probability, as evaluated by the peculiar statistical test of the researchers, of less than 0.05) is not appreciably different from what might be expected due to chance alone from ordinary statistical testing. So dowsing "skill" in the Munich experiments proved to be unreproducible across a spectrum of 500 candidates, as well as within a group of forty-three individuals selected because they initially seemed to be particularly talented (Figure 3); nor was it reproducible even by those six special individuals who on one occasion or another seemed to have guessed relatively well.

A Simple Alternative Strategy

There is another way of evaluating the results from those dowsers who produced the "best" test series of Figure 6. Suppose that they had always left their dowsing equipment at home in the closet, and had simply, in each and every test, just guessed that the pipe was located exactly at the middle of the test line. As shown in Figure 7b, all six of the "best" dowsers would have done better on average by making mid-line guesses than achieved by actual dowsing, in terms of the root-mean-square error, a commonly used index of reliability (similar to the standard deviation).

The root-mean-square error puts particular emphasis on gross mistakes, but these results can also be evaluated in terms of a different criterion that does not have that property: average absolute values of the errors. (Absolute values are preferable to simple averaging of errors; simple averaging would mean that an error of two meters to the left and another two meters to the right of the pipe might be regarded, on average, as perfect performance.) On the basis of this absolute-value criterion, as

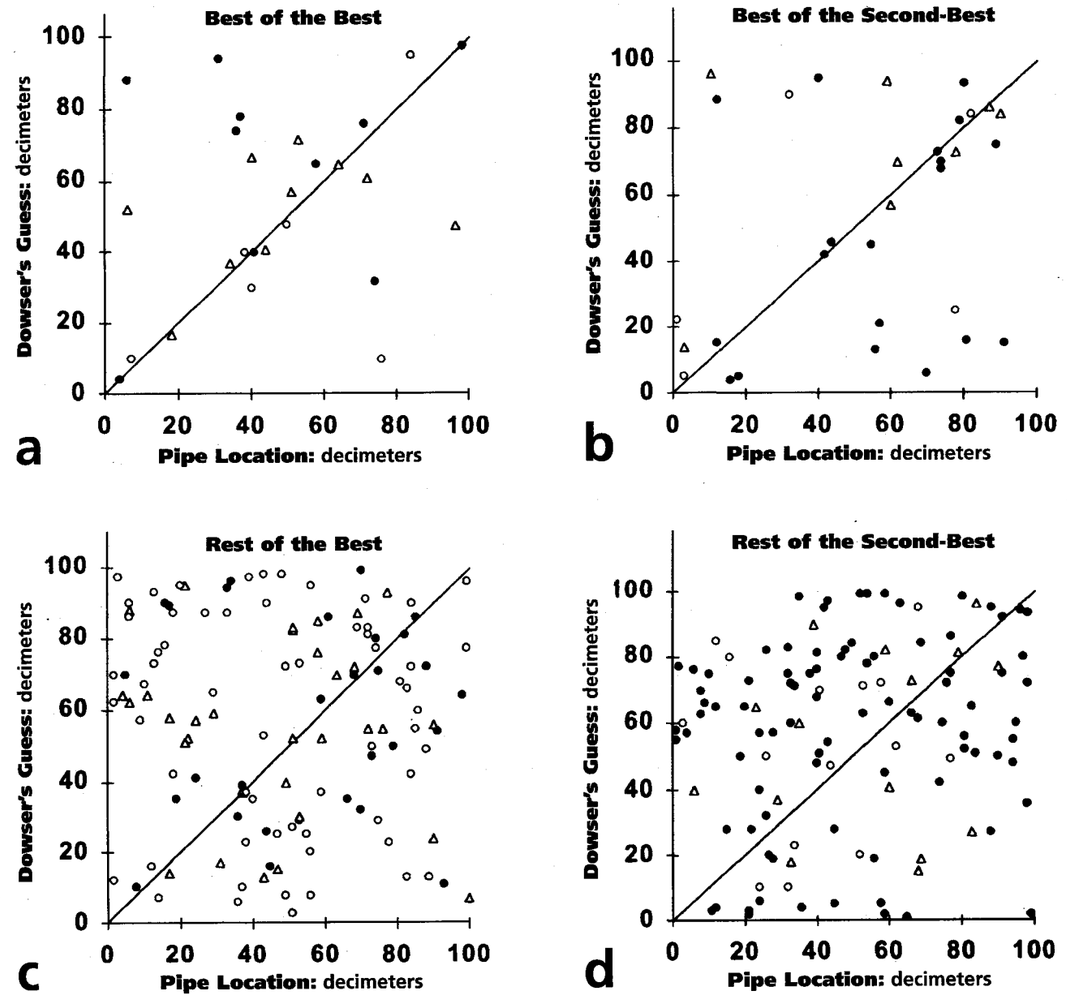

Figure 6: Selected examples of dowser performances. A: results from the three data sets that the Munich researchers considered "best" of all 104 series; b: results from the four test series, by three dowsers, that the researchers evaluated as also being exceptionally "successful"; c: Results from the other test series in which the dowsers of part a participated; d: Results from the other test series in which the dowsers of part b participated. In parts a and c, filled circles represent Dowser #99 (whose results are also presented in Figure 2), open circles represent Dowser #18, and open triangles represent Dowser #108. In parts b and d, filled circles represent Dowser #23, open circles represent Dowser # 110, and open triangles represent Dowser #89. Separate graphs of the results of these six individual dowsers are presented as Figures 2 and 3 in Enright, 1995. There is no hint in those plots that any one of the six was appreciably more or less successful than the others.

shown in Figure 7a, five of the six "best" dowsers would have made smaller errors relative to the pipe by using the mid-line strategy than they actually made with their dowsing tools. And what about the sixth, whose dowsing was somewhat better than middle-of-the-line guesses (#89)? Those results from actual dowsing were on average 4 millimeters better than mid-line guesses would have been. An average improvement of 4 millimeters (0.16 inch) by one dowser out of six (or out of forty-three, or out of 500) along a 10-meter test line scarcely seems worth the time and effort that the researchers and the dowsers invested in this project, nor worth the 400,000 marks that the German taxpayers invested in the study.

On the basis of these results (Figures 3, 5, 6 and 7), then, the Munich experiments constitute as decisive and complete a failure as can be imagined of dowsers to do what they claim they can.

A Sad and Sorry Postscript

Professor Betz (the primary spokesman for the Munich study) and his colleagues have published a response (Betz, H.-D., König, H. L., Kulzer, R., Trischler, R. and J. Wagner 1996) to

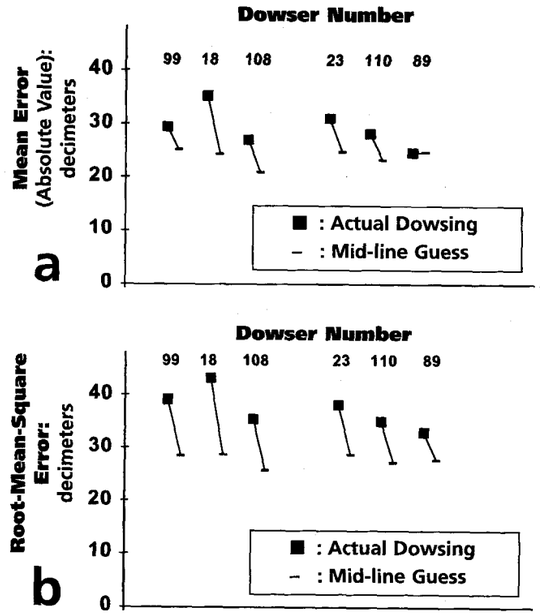

Figure 7: A: Dowser errors (averages of absolute values, indicated by filled squares) compared with errors that would have been made by guessing that in each test, the pipe was located exactly at the middle of the test line (also average absolute values, indicated by short horizontal bars). B: Root-mean-square errors made by the dowsers compared with R.M.S. errors that would have resulted from guessing in each test that the pipe was located at the midline, with symbols as in part a.

my critique of their results (Enright 1995). That defense is a relatively feeble one (Enright 1996), but such exchanges are a normal part of scientific controversies. Subsequently, however, Professor Betz (1997) published a paper in a fringe journal that crossed the ethical boundaries that usually characterize the scientific enterprise. In that article, he asserted that as a result of extensive scientific correspondence, I had conceded the validity of his own analyses and interpretations of the Munich dowsing data.

That statement is absolutely and categorically false. My only correspondence with Professor Betz (or anyone in his laboratory) since the publication of my original critique (Enright 1995) consists of an e-mail message sent him in July 1997, which dealt only with my request for documentation of an apparently implausible assertion about statistical procedures that had been attributed to him. He did not respond to that message, and so I re-sent the same message in August 1997, and again he did not respond. Two unanswered e-mail messages from me clearly do not constitute an "extensive scientific correspondence." And I have never, in publication, in correspondence, or in casual conversation even hinted that I accept Betz's analyses and interpretations of the Munich dowsing data. The results presented here as well as in the formal scientific literature (Enright 1995, 1996) provide such a clear demonstration *against* real dowsing skill that to assert that I had retracted my critique is both a false and an insulting assertion.

Conclusion

The Munich dowsing experiments represent the most extensive test ever conducted of the hypothesis that a genuine mysterious ability permits dowsers to detect hidden water sources. The research was conducted in a sympathetic atmosphere, on a highly selected group of candidates, with careful control of many relevant variables. The researchers themselves concluded that the outcome unquestionably demonstrated successful dowsing abilities, but a thoughtful re-examination of the data indicates that such an interpretation can only be regarded as the result of wishful thinking. In fact, it is difficult to imagine a set of experimental results that would represent a more persuasive *disproof* of the ability of dowsers to do what they claim. The experiments thus can and should be considered a decisive failure by the dowsers.

It seems very unlikely that any future careful experimental study of dowsing will produce results more favorable for the practitioners than the Munich experiments. An atmosphere more sympathetic to the dowsers, with so many concessions to their whims, seems hard to imagine. In view of the outcome of those experiments, it is very unlikely that any sponsor would ever provide funds for an even larger-scale study, such that very weak skills (which might conceivably have vanished into the statistical noise here) could be uncovered. (It is noteworthy that the U.S. Geological Survey concluded much earlier [Ellis 1917] that further testing of dowsing " . . .would be a misuse of public funds.") It seems appropriate, then, to reiterate here the general conclusion originally drawn from these analyses (Enright 1995):

> (These) . . . experiments are not only the most extensive and careful scientific study of the dowsing problem ever attempted, but—if reason prevails—they probably also represent the last major study of this sort that will ever be undertaken. (Enright 1995, 369).

Because of the vigor, however, with which Professor Betz and colleagues defended their positive conclusions (Betz et al. 1996), and in view of the discouraging history of other claims about the occult, one may have residual doubts, as do I, about whether reason will prevail in this arena (Enright 1996).

Acknowledgments

This research was supported by the National Science Foundation under Grant BNS 93-13038. The substance of this article was presented in a lecture at the Second World Skeptics Congress, in Heidelberg, Germany, in July 1998.

References

Betz, H.-D. 1997. Neue Ergebnisse der Rutengängerforschung. *Wetter-Boden-Mensch (Zeitschrift für Geobiologie)* 5: 55–59.
Betz, H.-D., H. L. König, R. Kulzer, R. Trischler, and J. Wagner. 1996. Dowsing reviewed—the effect persists. *Naturwissenschaften* 83: 272–275.
Ellis, A. J. 1917. Water-supply Paper 416, Department of the Interior, U.S. Geological Survey, Washington, D. C.: Government Printing Office.
Enright, J. T. 1995. Water dowsing: The Scheunen experiments. *Naturwissenschaften* 82: 360–369.
Enright, J. T. 1996. Dowsers lost in a barn. *Naturwissenschaften* 83: 275–277.
Wagner, H., H.-D. Betz, and H. L. König, 1990. Schlußbericht 01 KB8602, Bundesministerium für Forschung und Technologie.