Back cover

Closed Loop

Threads from CSGnet

This reproduction of *Closed Loop* was created by Dag Forssell in 2009. Addresses and phone numbers have not been updated. Most are obsolete.

Posted at www.pctresources.com

Spell checked, but not proofread. Please report any errors to webmaster. Email address at website.

Summer 1992

Volume 2

Number 3

Front cover

Closed Loop

Threads from CSGNet

Summer 1992 Volume 2 Number 3

Edited by Greg Williams, 460 Black Lick Rd., Gravel Switch, KY 40328

CONTENTS

Members of the Control Systems Group receive *Closed Loop* quarterly. For membership information, contact Ed Ford, 10209 N. 56th St., Scottsdale, AZ 85253; phone (602)991-4860.

CSGNet, the electronic mail network for individuals interested in control theory as applied to living systems, is a lively forum for sharing ideas, asking questions, and learning more about the theory, its implications, and its problems. The “threads” in each issue of *Closed Loop,* stitched together from some of the Net’s many ongoing conversations, exemplify the rich interchanges among Netters.

There are no sign-up or connect-time charges for participation on CSGNet. The Bitnet address is “CSG-L@UIUCVMD” (use no quotes in this and the following addresses); “CSG-L@VMD.CSO.UIUC.EDU“ is the Internet address. Messages sent to CSGNet via these addresses are forwarded automatically to all participants. Via CompuServe, use the address “> INTERNET :CSG-L@VMD .CSO.UIUC.VMD” to reach the Net. Initially, you should send a note to the network manager, Gary Cziko, at “G-CZIKO@UIUC.EDU“ (Internet) or at “c CZIKO@UIUCVMD” (Bitnet); Gary’s voice phone number is (217)333-4382.

Statistics vs. Generative Models

Bill Powers: Before I spend time trying to explain a phenomenon, I want to know if it’s real or just statistical. I want to know things like how many people show the phenomenon, how you find out that there’s a phenomenon, how many trials show the effect, and how many don’t—all that stuff. Once I’m convinced that there’s a real phenomenon, it’s time to think up explanations.

I’m not interested in 80 per cent correlations. That’s way too low to define a phenomenon.

“Superficial” knowledge is knowledge gained by observing apparent causal or coincident relationships without any generative model of underlying processes. Statistical studies yield superficial knowledge.

I think that all attempts to apply abstract physical principles and advanced mathematical trickery to human behavior are aimed at solving a nonexistent problem. They all seem to be founded on the old idea that behavior is unpredictable, disorderly, mysterious, statistical, and mostly random. That idea has been sold by behavioral scientists to the rest of the scientific community as an excuse for their failure to find an adequate model that explains even the simplest of behaviors. As a result of buying this excuse, other scientists have spent a lot of time looking for generalizations that don’t depend on orderliness in behavior; hence information theory, various other stochastic approaches, applications of thermodynamic principles, and the recent search for chaos and quantum phenomena in the workings of the brain. The general idea is that it is very hard to find any regularity or order in the behavior of organisms, so we must look beyond the obvious and search for hidden patterns and subtle principles.

But behavior *is* orderly, and it is orderly in obvious ways. It is orderly, however, in a way that conventional behavioral scientists have barely noticed. It is not orderly in the sense that the output forces generated by an organism follow regularly from sensory inputs or past experience. It is orderly in the sense that the *consequences* of those output forces are shaped by the organism into highly regular and reliably repeatable states and patterns. The Skinnerians came the closest to seeing this kind of order in their concept of the “operant,” but they failed to see how operant behavior works; they used the wrong model.

Because of a legacy of belief in the variability of behavior, scientists have ignored the obvious and have tried to look beneath the surface irregularities for hidden regularities. But we can’t develop a science of life by ignoring the obvious. The regular phenomena of behavior aren't to be found in subtleties that can be uncovered only by statistical analysis or encompassed only by grand generalizations. The paydirt is right on the surface.

The simplest regularities are visible only if you know something about elementary physics—and apply it. Think of a person standing erect. This looks like "no behavior." But the erect position is an unstable equilibrium, because the whole skeleton is balancing on ball-and-socket joints piled up one above the other. There is a highly regular relationship between deviations from the vertical and the amount of muscle force being applied to the skeleton across each joint. There is nothing statistical, chaotic, or cyclical about the operation of the control systems that keep the body vertical. They simply keep it vertical.

The same is true of every other aspect of posture control and movement control, and all controlled consequences of those kinds of control. Just watch an ice skater going through the school figures in competition. Watch and listen to any instrumentalist or vocalist. Watch a ballet dancer. Watch a stock-car racer. Watch a diver coming off the 30-meter platform. Watch a programmer keying in a program.

It's true that when you see certain kinds of human activity, they seem disorganized. But that is only a matter of how much you know about the outcomes that are under control. The floor of a commodities exchange looks like complete disorder to a casual bystander, but each trader is sending and receiving signals according to well-understood patterns and has a clear objective in mind—buy low, sell high. The confusion is all in the eye of the beholder. The beholder is bewitched by the interactions and fails to see the order in the individual actions. When you understand what the apparently chaotic gestures and shouts *accomplish* for each participant, it all makes sense.

Of course, we don't understand everything we see every person doing. It's easy to understand that a person is standing erect, but *why* is the person standing erect? What does that accomplish other than the result itself? We have to understand higher levels of organization to make sense of when the person stands erect and when the person doesn't. We have to understand this particular person as operating under rules of military etiquette, for example, to know why this person is standing erect and another is sitting in a chair. But once we see that the erectness is being controlled as a means of preserving a higher-level form, also under control, we find order where we had seen something inexplicable. We see that an understanding of social ranking, as perceived by each person present, results in one person standing at attention while another sits at ease. Each person controls one contribution to the pattern that all perceive, in such a way as to preserve the higher-level pattern as each person desires to see it.

It seems reasonable that once we have understood the orderliness of simple acts and their immediate consequences, we should be able to go on and understand more general patterns that are preserved by the variations that remain unexplained. As we are exploring a very large and complex system, we can't expect to arrive at complete understanding just through grasping a few basic principles. We must make and test hypotheses. But if we are convinced that the right hypothesis will reveal a highly ordered system, we will not stop until we have found it. If, on the other hand, we are convinced that such a search is futile, that chaos reigns, we will give up the moment there is the slightest difficulty and turn to statistics.

I claim that human behavior is understandable as the operation of a highly systematic and orderly system—at least up to a point. I say that it is the duty of any life scientist to find that orderliness at all discoverable levels of organization, and to keep looking for it despite all difficulties. We must explore all levels, not just the highest and not just the lowest; what we find at each level makes sense only in the context of the others.

We have a very long way to go in understanding the obvious before it will be appropriate to look for subtleties. I have no doubt that we will come across mysteries eventually, but I'm convinced that unless we first exhaust the possibilities of finding order and predictability in ordinary human behavior, we won't even recognize those mysteries when they stare us in the face. I don't think that anyone is prepared, now, to assimilate the astonishments that are in store for us once we have understood how all of the levels of orderly control work in the human system.

We won't get anywhere by looking for shortcuts to the ultimate illuminations that await. Most of the esoteric phenomena of physics that are taught in school today were occurring in the 19th Century. But who, in that century, would have recognized tunneling, or coherent radiation, or shot noise? If we want to see a Second Foundation of the sciences of life, we have to begin where we are and build carefully for those who will follow us. If we succeed in trying to understand the obvious, the result will be to change what is obvious. As the nature of the obvious changes, so does science progress.

Chuck Tucker: I think that the majority of those who have difficulty accepting our approach simply hold to the assumptions about the world attacked by Dewey in *The Quest for Certainty,* and rejected by us: that the real world will be revealed to you if you just use the "proper" methods and work hard enough. If we tell these people that their approach won't reveal the "true forever world," then they seem to have much less interest in what we have to say. Another feature of many of those who reject our view is that they are not "problem-oriented"—that is, they do not tolerate ambiguity, uncertainty, and problem-solving activity for very long; they want the answer quickly and cheaply (or statistically). But our approach does not offer such a magic solution; just hard, dirty, difficult work, with no absolute assurances that a solution will be fashioned, let alone work. Think about it: would you give up such a pleasant life of certainty and bliss for the one we offer? Probably not.

Bill Powers: David Goldstein and I have been conducting an argument for several years. David tends to win many of the rounds because he is working with clients who have both real and severe problems, and I often have to admit that when you're faced with solving such problems, you have to do what's possible. If a person is so depressed as to be on the verge of suicide, you give the person a pill that takes the edge off, and you're glad that such a pill exists. Afterward, you can think about trying something else. Even control theory can't cure a dead client.

A lot of our arguments are conducted in the context of such practical limitations. But I don't have David's responsibilities, so I can argue against conventional methods even if I don't have an immediately applicable alternative to propose. One of these arguments has to do with the utility of testing, particularly testing that involves questionnaires and other means of self-description such as Q-sorts. Basically, I argue that verbal tests are too imprecise to do much good, and that they inevitably put us in the position of applying statistical methods to individuals. I argue that we should be trying to apply control theory directly, trying to find out what individuals can and can't control, and trying to find out why they are having trouble. This means abandoning old diagnostic categories and old attributions of traits and conditions in the attempt to explain what's wrong. I claim that we must make a conscious effort to break free of cultural assumptions, which always steer us back toward the conventional categories. David doesn't exactly disagree with me, but—well, he can speak for himself.

David has proposed "qualitative modeling," the sound of which I rather like. He says, "Suppose that we plotted the urge to perform action X against time. The lowest point of the curve can be taken to be the reference level for whatever perceptions are being controlled by action X. Suppose that on a scale of 0 to 10, the intensity of perception Y1 = 2 and the intensity of Y2 = 5 at the lowest point. As a person deviates from these values, control theory leads us to expect increasingly stronger urges to perform action X the further we move away from these reference level values. If we do not obtain a U-shaped function around these values, then the particular clinical hypothesis may be rejected.

What do you think?"

I think that the method as stated predetermines too many variables. The first objective should be to see what perceptions are under control. To do that, you have to allow the action-variable to be free. If the perception is "people like me," the action that will contribute to that perception will be different under different circumstances (meaning different disturbances of the sense that people like me).

Under the conventional approach, we would be most concerned with the action, because that is what other people experience. But to understand the acting person, we first have to understand what perceptions are under control. A given perception can be controlled through many different actions, so no one action is significant by itself. Furthermore, we might see both an action and the opposite action being taken as a means of controlling the same perception, depending on whether disturbances are pushing the perception above or below its reference level. The object of control theory can't be to explain one particular action.

So I would propose backing up a step or two, and starting by testing Yl, Y2, Yn to see if they are controlled variables. This is hard to do using a verbal test, first because while taking the test, the person isn't experiencing the perception, but only a description of the perception, and second because the only way to apply disturbances is hypothetically, by describing them and asking how the described disturbance would affect the described perception (and, presumably, what the person would do if the perception changed). I much prefer direct interaction in real situations, with perhaps a discussion afterward if you want to cast the interaction in verbal terms. Maybe role-playing would be a compromise that allows setting up hypothetical situations while still allowing real perceptions and direct interaction with disturbances (supplied by the experimenter).

Gary Cziko: I have read Philip Runkel's book, *Casting Nets and Testing Specimens* (Praeger, 1990), and I believe I understand his arguments about why multiple regression (MR) and other "relative-frequency-based" analyses based on group data cannot tell us much, if anything, about the functioning of organisms. Bill Powers has suggested that MR cannot even be profitably used for predictions about individuals. But everything I've learned about MR tells me that this indeed can be done.

Let's use a medical example. I can draw a random sample from some population of interest. I want to be able to predict blood pressure, so I obtain data on weight, per cent body fat, smoking, dietary habits, and perhaps even have each person fill out some questionnaire relating to stress. I can then do an MR which will provide me with a weighting of independent variables best predicting the dependent variable, blood pressure. If I get a high multiple correlation (r-square), I can then use this regression equation to predict the blood pressure for someone whose blood pressure I have not yet measured, but for whom I know the values of the independent variables. Of course, this person must be a member of the original population. I know that I will not be able to predict his or her blood pressure exactly, but if I do the statistics right, I should be able to attach probabilities to ranges of values, i.e., establish confidence limits for his or her predicted blood pressure.

I realize control theory says that such a study does not necessarily tell me anything about what causes blood pressure to rise or fall in people in general or in any individual (Runkel's book makes this point well). And I realize that it would probably be easier just to measure the blood pressure instead of predicting it (it's a poor example from that viewpoint). But why can't I use this technique for predicting for individuals?

Bill Powers: Gary, I wish Phil Runkel were on the net, but I'll try to defend my statement without an expert's help (with the usual risk of getting it all wrong).

My basic argument is that you could use the MR method to predict the average relationship of various factors to blood pressure in *another group of the same size from the same population,* but you have only a tiny chance of guessing right about any individual from either the old group or the new group. I won't even get into the problem of how you know you're drawing from the same population, a subject on which Phil Runkel has some cutting remarks.

The reason for my opinion is that the "independent variables" (or the factors you get from them) are not known to be physically causative of high or low blood pressure: they are simply associated by experience with blood pressure. When you use multiple tests, the intuitive thought would be that getting at the relationship from many independent angles ought to improve your ability to predict for a single person. I'm quite sure that it doesn't, but let's see if I can work up a coherent justification for saying that.

If you looked at the raw data from the tests, you would find that some people high in each factor had high blood pressure, while others did not. Let's be generous and suppose that 80 per cent of the people in the original group who scored high on each factor actually had high blood pressure.

If that is true, and if 1000 people participated in the study, 800 of them who scored high on the first test had high blood pressure, while 200 of them didn't. We now have 800 people left whose scores on the first test truly indicated high blood pressure, or seemed to. Now we give the second test. After this test, we have 80 percent of 800 or 640 people who indicated high on both measures and did indeed have high blood pressure. After the third test we have 512 people left, after the fourth test, 410 people left, and after the fifth test, 328 left. Therefore, out of the original 1000 people, only 328 who scored high on all five tests proved to have high blood pressure. So if you give all five tests to an individual, and the individual scores high on all five measures, the chances of high blood pressure are about one in three. In other words, you'd be safest in betting that a person who scores high on all five "indicators" does *not* have high blood pressure.

Why this counterintuitive result? I think the reason is that we confuse association with causation. If it were true that, for example, a high load of body fat *physically caused* high blood pressure, then there would be no way for an otherwise normal person to have high body fat and not have high blood pressure. The only room for error would be in measuring body fat or in finding the right curve relating body fat to blood pressure. A deviation would basically be a measurement error, not a matter of chance membership in a population. Body fat would amount then to a measure of blood pressure.

In the same way, each other measure, if it were truly a physically causative factor, would also amount to a way of measuring blood pressure, and you would expect using these multiple measures to reduce the error of measurement. But these measures are *not* measures of blood pressure. They're not "measures" at all. They are simply factors that common sense tells us might have something to do with the matter. That being the case, we are not perturbed by finding that a person who has high body fat happens to have low blood pressure. If there were a physical chain of causation involved, we would be very perturbed indeed to find our measuring instrument suddenly indicating the wrong way. This is the difference between physical or model-based measurements of relationships and statistical inference of relationships. There are no physical principles operative in a statistical inference, and of course the only model is pretty elementary.

This misuse of statistical "facts" is encouraged by the habit into which most empirical scientists fall, which is to say not that "80 per cent of people with high body fat have high blood pressure and 20 per cent don't," but that "high body fat predicts high blood pressure." The customary wording implies that this is *always* true; this makes the factor look like a physical cause. Just look at any summary of findings in a statistical study. Does it tell you the chances that a given person does not show the effect or shows the opposite effect? It does not. It says "A is associated with B." In *everybody.* That is why you expect the result to apply to *anybody.*

In truth, nobody knows why, in some people, the reference level for blood pressure is set to a high value. Nobody knows, because all the big research money goes into statistical studies instead of into developing a competent model of how the human system works. I wouldn't recommend that we just do studies of physical causation, because I don't think that's how you come to understand a system, but I do recommend that we study the ongoing networks of relationships that constitute a functioning body and brain. Until we do that, none of this statistical crap is going to do much good for an individual who has to make decisions based on an N of I and gets only one chance to bet right.

I smoke, eat eggs and bacon, weigh about 30 pounds too much, don't get a lot of exercise, and have, at last measurement, a blood pressure of about 125/80. Just a statistical fluctuation, that's me.

One last consideration. I think that studies involving very large numbers of people, like the cholesterol studies, are probably worse indicators of an individual's characteristics than studies involving only a few subjects. My reasoning is that large studies are necessary only when the effect is very small—when the number of people showing the effect is only slightly larger than the number not showing it. If 80 or 90 per cent of subjects in a pilot study showed the effect, why on earth would anyone then expand the study to huge numbers of people? In a large study we are justified in suspecting that the split is not 80/20, but more like 51/49. The numbers are needed to get statistical significance out of an effect that's just barely there.

In medicine, the practices are even worse than that. I recently saw a glowing report on a drug which statistics proved to help 16 per cent of the people who took it. In other words, 16 per cent got better and 84 per cent didn't. I think that result leaves room for a lot of questions about just why those people actually got better, and what effect the drug had on those who didn't. This sort of mindless application of statistics goes on all the time. Remember that the next time someone tries to get you to pop a wonder pill (unless you have as many chances to try to get well as necessary). Ask for a warranty.

One more last thought: Suppose it happened that all five tests together were a very good predictor of high blood pressure. Is that any reason to think that reducing all five factors would reduce the blood pressure? This is another elementary logical error: thinking that an implication works both ways. Suppose that the blood pressure is high for the same reason that leads to high values of these other factors. Statistics says *nothing* about causation.

See my paper in the *American Behavioral Scientist* issue edited by Rick Marken (September/October 1990) for a demonstration of how a statistical analysis can yield an apparent relationship that actually goes the wrong way.

Gary Cziko: OK, Bill, here's some thought data: 0 indicates low on a factor, 1 indicates high; A through D are independent variables, Y is dependent (blood pressure):

Subject	A	B	C	D	Y
1	0	1	0	0	0
2	0	0	1	0	0
3	0	0	0	0	0
4	0	0	0	1	0
5	1	0	0	0	0
6	0	1	1	1	1
7	1	0	1	1	1
8	1	1	1	0	1
9	1	1	0	1	1
10	1	1	1	1	1

Note that only 80 per cent (4/5) of those scoring high on A have high blood pressure; the same holds for B, C, and D. The one person who is high on all four independent variables has high blood pressure, the one low on all four independent variables does not. In addition, *everyone* scoring high on at least four out of five independent variables has high blood pressure, and no one who scores low on four out of five has high blood pressure. And so perfect prediction is possible with these data. Of course, things might not be so pretty when I get another sample, since this sample is very small. But if with a larger sample I still don't get individuals deviating from this pattern, I would feel pretty confident in predicting an individual's blood pressure based on his or her characteristics as defined by the independent variables.

Looks pretty good to me.

Bill Powers: For those finding my statistics hard to swallow: If you propose that each of five conditions is associated with high blood pressure, but have no model and no knowledge of the physical means by which each condition has its effect, you can only assume that each association is independent of the four others. There is no a priori reason to assume that testing high on one measure predicts testing high on another.

The upshot is that you must assume that, on each test, the distribution of people measuring high on that parameter is independent of the distribution for any other parameter. When you isolate the 80 per cent who scored high on a given measure *and* had high blood pressure, you have not thereby isolated those who will score high on any other test (as Gary's example assumed). You have only eliminated those who tested high on one test but showed low blood pressure. Among those who are left, however, only 80 per cent, again, will score high on another test *and* have high blood pressure. Having high blood pressure is not sufficient to predict how a person will score on a test that seems to predict high blood pressure. It is a common error to suppose that this is true, but it's not. Implications don't work backward, as I said. Getting on a train at the next-to-last station implies—very reliably predicts—getting off at the last station. But if you see a person getting off at the last station, this does not imply that the person got on at the next-to-last station.

Finding, through factor analysis, a factor related to blood pressure *reduces* the credibility of an individual measure having a causal role. The hidden factor correlates better with the dependent variable than do the individual measures, which indicates that the hidden factor might be having a direct effect on the dependent variable and a lesser effect on the initially proposed independent variables. Of course, the hidden factor could itself be a side-effect of an even more important cause that also affects the dependent variable. It's simply a mistake to assume that an association implies a dependent and an independent variable. The fact that it's commonly assumed doesn't make it right.

Suppose that a person were in conflict. This can mean being physiologically prepared to act but not being able to carry out the actions that would normally "use up" the prepared state. One consequence of this state might be an elevation of the reference level for blood pressure. Among other consequences would be the tendency to measure high on stress, to seek comfort in good food or to gobble fast food, to be unable to act vigorously (a direct effect of conflict that equates to "little exercise" and thus being overweight), and so on. So it is not at all farfetched to propose a common reason for the high blood pressure and for the high scores. When that is the case, lowering the test scores will have no effect at all on the blood pressure.

Phil Runkel has laid out the circumstances in which statistical studies are appropriate and meaningful. These do not include the prediction of individual behavior or the exploration of natural laws. You learn through statistics what masses of people actually do, but you learn *nothing* about the underlying processes that lead to individual behavior. Statistics, when applied to individuals, is not science. It is organized superstition and systematized prejudice. It gives the illusion of knowledge, which is probably worse than ignorance.

Gary Cziko: Bill, please note that I have read (several times!) Runkel's book and find his arguments quite convincing that group statistics do not necessarily tell you anything about how individuals function. I do not, however, understand the part of Chapter 8 on regression, and that is perhaps what started all this. While statistics might not tell you much of anything about how people function, I still suspect that they *can* help in certain types of predictions about individuals.

You say: "If you propose that each of five conditions is associated with high blood pressure, but have no model and no knowledge of the physical means by which each condition has its effect, you can only assume that each association is independent of the four others. There is no a priori reason to assume that testing high on one measure predicts testing high on another."

But if one has no model, why does that force one to assume independence among the four independent variables? In fact, we know in the behavioral sciences that everything often seems to be at least a little related to everything else, so why assume independence? Your "upshot" is suspect if the assumptions are suspect.

Regardless of train riding practices, correlations, as I understand them, work both ways. If there is a 0.7 correlation between percent body fat and blood pressure, then there is a 0.7 correlation between blood pressure and body fat. Now, the regression line (and equation) will be different depending on which way you go, but that is only because the variances of the two variables are not likely to be equal.

Bill, you talk about causality; I'm only talking prediction. Why do we need causality for prediction? There is probably a positive correlation between shoe size and reading ability among elementary school children. This doesn't mean that kids use their feet to read; the causal factor is more likely to be something like age (but even this alone will not cause better reading skills). But as long as there is a nonzero correlation between shoe size and reading ability, I can use shoe size to make a prediction about reading ability that is better than a prediction made without knowledge of shoe size. Being ignorant of shoe size, I can only predict the mean of the group with a standard error of estimate equal to the standard deviation of the reading scores. With shoe size, I can reduce this error of prediction so that it is *less* than the standard deviation of the reading scores. And if I have a perfect correlation, there is no error at all. Why I do I need to find causal factors to make predictions? The daffodils coming out of the ground do not cause Easter. And yet when I see them growing, I can predict that Easter is not far away.

You also say that "through statistics... you learn *nothing* about the underlying processes that lead to individual behavior." I agree, but that still doesn't make it clear to me that statistics is useless for predicting aspects of individuals. Insurance companies would all probably go broke if they didn't use statistics for these purposes.

Let's try to keep away from the "understanding specimens" argument. Runkel does this well, and anybody can read his book. However, if we can effectively dismantle the individual prediction rationale for statistics, this will really pull the rug out from under the social (including medical) sciences, and this would indeed be great fun. I'm really on your side (I think), but I'm not yet convinced. Please be patient.

Mark Olson: Bill, like Gary, I understand that we want to keep away from an "understanding specimens" argument, and that the idea in question is whether statistics has any predictive value. Gary's argument makes complete sense to me, so I am anxiously awaiting your rebuttal, and like Gary, I hope you are right. If I may make a trivial request, could you stick with the shoe size and reading ability example—this is the example I use in my educational psychology class to teach the concept of correlation—the train example confuses things. Thanks.

Chuck Tucker: The important point Runkel makes that can get lost in these discussions is not that statistics is bad or dumb or worthless, but that it is a tool that can be used for some specific purposes but not for others. Statistics is a very weak tool to make sense out of what people do—some statistics make sense or are useful, but others are not as useful. It is like using a hammer to put a screw into wood—you can do it, but it will mess up the screw head and the wood and probably won't hold very well. This is the case with most of statistics *if* you are concerned with how the human being works; its use is very limited and might in fact be harmful to your understanding. The argument is pragmatic in the best sense of the word.

Rick Marken: Bill says: "You learn through statistics what masses of people actually do, but you learn *nothing* about the underlying processes that lead to individual behavior." Gary replies: "1 agree, but that still doesn't make it clear to me that statistics is useless for predicting aspects of individuals. Insurance companies would all probably go broke if they didn't use statistics for these purposes."

I think we are getting philosophical here—so I'll jump in blindly. I think there is nothing harder for people to understand than the point you guys are trying to make. People make individual decisions based on mass data all the time, and they consider it very reasonable. In other words, they are predicting aspects of individuals (themselves) based on statistical data. Lots of behavior is done solely because the statistics imply that you, as an individual, are more likely to be X rather than Y if you do Z. Even a somewhat rational person like me bases some individual decisions on what the statistics say.

Gary is right about prediction and statistics—my prediction that a person will have value X on a particular dimension is better (smaller RMS error over predictions) if I know some predictor variables and the equation relating them to values on the dimension of concern. But Bill is right because this kind of prediction is of no use for an individual. Accuracy is defined over prediction occasions, and an individual is just one occasion. So it is perfectly reasonable, I think, for an insurance company to charge me more for life insurance if I smoke. But it is silly for me not to smoke based on statistical data. I am not a likelihood. I'm just me, once. I can only base my attempts to control things (and that is what you are trying to do when you base life decisions on statistical data) on what is happening now, not on what might happen on repeated samples of my life. I can control my insurance premium, my attractiveness to those I care for, and other things by not smoking. But I have no way of controlling how long I live or whether I get cancer. Those things only happen once, and there is no evidence that they can be reliably controlled by individuals' variations in their smoking behavior (individually—I know that, statistically, non-smokers do better on these things, but this is irrelevant to me individually).

Maybe control is the operative concept here (not statistical control, but perceptual control). Statistical evidence gives no evidence of an individual's ability to control variables. Statistics on smoking tell me nothing about how I, individually, can control cancer in myself. People often point out the individual irrelevance of smoking statistics by pointing to folks like George Burns. This irrelevance does not mean that smoking might not be bad for many people—eating candy is bad for some people, too. Also, there are probably perceptual consequences of smoking that can be controlled by cutting down or stopping. If people want to control these consequences, then controlling their smoking might be tried. But trying to control variables by basing individual actions on statistical data is just silly. People can only control perception; controlling imagination doesn't help anything. In fact, spending a lot of effort controlling imagination is called neurosis, isn't it? The applicability of statistical data to any particular individual is imaginary, so controlling individual behavior based on its imagined statistical consequences seems to me like neurosis.

Joel Judd: I got the impression from Gary's last comments that he was looking for some logico-mathematical reasoning for arguing against inferential statistics, instead of the "specimens" argument. But it seems that all one needs to do when contemplating the use of a tool—e.g., statistics—is ask, "What do I want to use this tool for?" One doesn't have to delve into the physics and whatnot of screws and screwdrivers and hammers to figure out that a hammer doesn't put in screws well (Chuck's example). Every statistical tool has some mathematical assumption(s) underlying it, delimiting its use. What else should one have to say when defending a perspective such as Runkel's? I want to know *why* someone does X. Group statistics can't tell me.

Mark Olson: Rick, I think I follow your smoking/cancer example. But I first need a distinction to be made before I feel I truly understand. We say that smoking and cancer are correlated. We also say that children's feet size and reading ability are correlated. Yet I see these as being correlated for very different reasons. In the former example, smoking "could" cause cancer, while in the latter example, size and ability cannot be causally related. It seems that this difference should have some importance in this whole issue, and I can't quite seem to articulate what that might be any insights?

Bill Powers: Rick says: "Lots of behavior is done solely because the statistics imply that you, as an individual, are more likely to be X rather than Y if you do Z. Even a somewhat rational person like me bases some individual decisions on what the statistics say."

Statisticians like to point out that people who use informal statistical analysis as a basis for choosing behavior don't do very well at it. I bought two lottery tickets because the pot was $60 million on Wednesday. A rational analysis shows that if I had bought *all* of the tickets, I would have been *certain* to lose something like $20 million (or some big number). So the optimum number to buy, considering that the $2 could have been spent on a hamburger which would certainly do me some good, was zero.

But Rick's point is well taken. It reminds us of what statistics is all about: trying to make predictions about what will happen on the basis of what has happened. This is all people could do prior to science: they didn't know how to figure out the underlying processes so they could predict what is going to happen without having to remember and analyze what has happened. Once you have a workable idea of the inner organization of any system, you can predict what it will do even under circumstances that have never happened before. Of course, you have to study what happens in the world in order to find a good model. But once you have the model, you predict from it, not from average past behavior. The record of physics and chemistry shows that this approach is far superior to merely watching behavior and assuming that the future will be like the past.

When your motorcycle starts making a funny tapping sound, there are two ways to fix it. One is to try to remember what the mechanic found the last time that sound happened and replace the same part. The other is to understand how the engine works, inside, and figure out that *this* time it's the tappets. What was wrong the *last* time is then irrelevant. Of course, if the previous trouble was also the tappet adjustments, then this time you should *not* merely adjust the tappets. First, you should figure out why the setting isn't holding. You have a different problem, and the tappet maladjustment is only a symptom of it.

Tom Bourbon: Concerning the recent discussion about statistical predictions, there was an observation that there is a difference between

correlations such as the one between smoking and lung cancer, and the one between shoe size and reading skill. That is true. A correlation between two sets of numbers means nothing more than that the positions of individual cases on one measurement scale resemble their positions on another scale. The equations used to calculate the degree of correlation could care less where the numbers came from or what they mean. That is as it should be, and that is one reason statistical analyses alone cannot reveal information about individuals.

However, when used in the context of research driven by a theory that makes bold predictions (i.e., specific, quantitative, falsifiable predictions), correlations can provide strong evidence about causal relationships. In the case of correlations found in control behavior, however, the correlations go counter to what most behavioral scientists have come to expect. For example, if a person is controlling a variable that is subject to independent disturbances, the actions of the person will be essentially *uncorrelated* with the value of the variable the person is controlling, but will be *highly negatively* correlated with the net disturbances acting on the controlled variable. To an uninformed observer, the person's actions will appear random, and the person's control over the perhaps unchanging controlled variable will go unnoticed.

In tracking studies such as those used by some of us who do control-theory modeling, the correlations between 1800 pairs of values of positions of a control handle and of values of the net disturbance on a controlled cursor are as high as -0.998. Of course, with n = 1800, no test of statistical significance is needed to know that the person moved the handle to negate the effect of the net disturbance. To do a statistical test of significance on data such as those would be utterly ridiculous.

In tracking data, the correlation between positions of the cursor and of the handle varies around 0.0, but it can be as much as +0.2 or -0.2. With n = 1800, those correlations are highly statistically significant; but of course they are totally meaningless.

In more traditional psychological research, correlations can provide some grounds for prediction, but only if the assumptions and requirements of the statistical procedures are met. That was one of Phil Runkel's major points in his book. Phil did not reject the "method of relative frequencies," *as* he identified traditional research designs and statistical analyses. But he did rightfully and masterfully show that those methods cannot work if one uses them to gather information that lets one make firm statements about individuals.

An excellent example of the problems encountered when people try to use statistical evidence to make statements about individuals can be found in R. M. Dar, D. Faust, and P. E. Meehl, "Clinical vs. Actuarial Judgment," *Science 243,* 1989, 1668-1674. The authors summarize the now sizeable literature which reveals that nearly any simple-minded actuarial procedure can out-diagnose nearly any practitioner who relies on "clinical judgment." Those results are telling. But the authors make another major point: even the best actuarial procedures are not very good. The actuarial procedures produce validity coefficients a few per cent higher than those produced by clinicians acting on professional judgment alone. The correlations between diagnoses and confirmed "pathology" are in the 0.20-0.50 range, which is the range one typically sees in the literature for the behavioral sciences. It appears that the clinical psychologists, burdened as they are with the "scientist-practitioner" model under which they train, do about as well as the behavioral scientists when it comes to identifying relationships—and neither group does very well.

Dar, Faust, and Meehl also draw a distinction between the state of affairs in clinical diagnostics and that in science, where access to a strong, corroborated model gives the edge to the scientist over actuarial procedures. The reason, of course, is that the scientist has an understanding of *causes*. Those who rely on actuarial procedures labor under the handicap of ignorance about causes—or else they act as though they understand causes, as when they assume causal relationships among the variables that enter into a multiple regression equation.

Gary Cziko: Reading some of Tom's comments, I get the feeling that the issue we are discussing here all reduces to the notion of individual differences in reference levels (internal standards). If everyone in *a* population had the same reference level for some perception, then we would get nice group correlations between disturbances (which would look like stimuli) and behavior which (it seems to me) *would* tell us something about the workings of individuals. However, individual differences cloud this relationship, so the only way to get at it is to examine individuals separately and then see what the invariances are at a more abstract level.

As far as I know, all strips of copper or containers of oxygen are basically alike. We can push and pull on them and send electrical currents through them and see how they behave without worrying about differing internal standards. And this is what traditional psychological methods do with people. Maybe psychology has forgotten why people in experiments were originally (and are still today) called "subjects." For the type of research usually done in the behavioral/social sciences, aren't they really treated as objects?

Tom Bourbon Gary Cziko has remarked that the behavioral and social sciences treat people like objects. That is true, not just of their treatment of people, but of living things in general. It is as though behavioral and social scientists expect living mice to "obey" the same causal

laws as the obliging "creatures" whose tails plug into computers, and who jump at our merest touch.

Nestled among the ever-increasing contents of my CST bookshelf is Lewis Carroll's *Alice's Adventures in Wonderland & Through the Looking-Glass*. Carroll understood the distinction and expressed it eloquently in the chapter on "The Queen's Croquet-Ground." I believe Carroll's message is one every control theorist understands—one every behavioral and life scientist should learn:

> Alice thought she had never seen such a curious croquet-ground in her life: it was all ridges and furrows; the croquet balls were live hedgehogs, and the mallets live flamingoes, and the soldiers had to double themselves up and stand on their hands and feet, to make the arches.
>
> The chief difficulty Alice found at first was in managing her flamingo: she succeeded in getting its body tucked away, comfortably enough, under her arm, with its legs hanging down, but generally, just as she had got its neck nicely straightened out, and was going to give the hedgehog a blow with its head, it *would* twist itself round and look up into her face, with such a puzzled expression that she could not help bursting out laughing; and, when she had got its head down, and was going to begin again, it was very provoking to find that the hedgehog had unrolled itself, and was in the act of crawling away: besides *all* this, there was generally a ridge or a furrow in the way wherever she wanted to send the hedgehog to, and, as the doubled-up soldiers were always getting up and walking off to other parts of the ground, Alice soon came to the conclusion that it was a very difficult game indeed.

That's life!

Mark Olson: Tom said that it is true that we can't compare correlations of smoking and cancer to correlations of feet size and reading ability. But this didn't answer my question about what *is* that difference between these two examples. What Tom wrote was helpful, but it didn't answer my question (at least not directly). Any comments?

Tim Cutmore: Would we say that smoking causes cancer if it were found that all (or perhaps just "most" would do) people who smoke also were exposed to Z-rays when children, and the Z-ray exposure induced the degree of desire to smoke? *And* it was also noted that Z-rays have a dose-related latent effect in causing cancer (amounting to accounting for 99 per cent of the variance in lung cancer!)?

In this case, we would have a superordinate variable which caused both smoking and cancer (vis-a-vis age reading experience -4 reading ability and age → foot size; age is the superordinate variable). The difference in what we believe appears to depend on perceiving the relations of the dependent variable (reading ability or cancer) to a superordinate variable (or not).

Izhak Bar-Kana: As the name says, a correlation only shows that some relation apparently exists between two different things, for example when one is large, the other is mostly large, etc. It doesn't say if one is the cause of the other, if one precedes the other, or not. The difference between the smoking and cancer vs. feet size and reading ability examples is only in the *additional* knowledge or assumptions involved. People have assumed for a long time that smoking might lead to cancer, and the correlation shows that, statistically, there might be something here. If the correlation is all you have, you might assume that cancer *is* the cause of smoking, or that both have some common cause.

In the second case, one only starts measuring and finds some statistical relationship between feet size and reading, and then tries to make something out of it. But one then needs more: assumptions, revelations, or some discovery that would prove/disprove that the statistical result is relevant.

Tom Bourbon: Mark has convinced me that I did not make my point clearly. One may assert that *any* two (or more) sets of correlations are comparable. Nothing in the procedures for calculating correlations rules out any use to which a person might put the results of the calculations. As I understand it—and I am not a skilled mathematician—computational procedures of all kinds are blind as to the origins of, and the meanings of, the numbers that are fed into them. And they are equally blind to the meaning of the results. Meaning and significance are in the eyes of those who behold the results, not in the results.

That is why Tim is free to tell us that his hypothetical Z-rays really do explain the variance in occurrence of lung cancer, and that the putative association with smoking should be put aside. For some reason, I doubt that Tim would do that, not because of anything in the rules by which one plays the correlation game, but because such an argument would not sound plausible to the professional community. Too many other things people believe they already know would be in jeopardy—and I do not mean that in a trivial sense. The assertion of as-yet-unrecorded rays that can play a major role in a prevalent medical problem would stretch at the boundaries of science. (Goodness knows, the boundaries need stretching from time to time—ask any control theorist who tries to publish!) Unless Tim could offer clear evidence that passed the scrutiny of scientists, and, more importantly, of good professional magicians, his

assertion would sound too much like the N-rays that Blondlett and his associates could see in France, early in the century. (Heard much about N-rays, lately?)

Which is merely another way of saying what I did in my last post: the smoking-cancer association *seems* more plausible than the shoe size-reading ability one. It is all in the sense of how the assertions fair with (fit with, form a nice figure with) the other things we know. And that has nothing to do with the numbers, per se.

Wayne Hershberger: Tom, your reference to the article by Dar, Faust, and Meehl reminds me that Meehl published an article within the last three years—in one of the APA journals, I think—comparing the methodologies of the hard and the life sciences. His arguments are consistent with, if not identical to, Bill's emphasis on "model building" and Phil's concern with "testing specimens."

Bill Powers: It seems to me that there are three topics concerning statistics needing separate discussion here. One is the question of causality; another is the question of applying a statistically obtained regression line to individuals; the third is the quality of the data on which the analysis is based.

On causality: I think we are all agreed that correlations do not reveal causation. Causation could run backward to the intuitively assumed direction (incipient cancer causes a desire to smoke), could result from a superordinate cause (Z-rays cause both a desire to smoke and cancer), or could be symptoms of some other process (smoking is a normally successful attempt by the system to counteract the onset of cancer—what percentage of smokers don't get cancer?). No information about these possibilities or any other comes out of a statistical study.

On the application of statistical relationships to individuals: Large studies involving many individuals yield a scatter of data. The common assumption is that this scatter is due to uncontrolled environmental variables. But an even stronger assumption is that measuring many individuals under varying conditions is the same as measuring *one* individual under varying conditions: in other words, all individuals in the population are alike and interchangeable.

Even granting an underlying justification for associating a statistical relationship with a causal relationship (for example, having a model whose properties agree with the statistical results), the statistical relationship (regression line) for a population might have nothing to do with the quantitative relationships inside each individual that link individual behavior to the independent variable(s). I showed in my *American Behavioral Scientist* paper that individual differences can account for the slope of a population regression line, while inside each individual the relation of behavior to the independent variables has a slope opposite to that of the population.

Also, confidence levels do not apply to individual measures. If p is less than 0.05, this means only that there is less than one chance in 20 that the correlation observed in the aggregate data is due to a chance fluctuation in variables that are actually unrelated. If the entire study were repeated 20 times, only once would the correlation measure zero. Is there any way to calculate the chance that an individual deviation from the mean is due to random departure from the population mean effect rather than a random departure from the condition of no relationship? It seems to me that this would be like the effect of an individual not actually being from the same population (where a population is defined as people with identical properties). What is the chance that an individual is not a member of the assumed population? Isn't it the product of the probabilities that the person will test positive on each indicator of population membership?

On the quality of the data: I've said that a correlation of 0.8 looks terrible on a scatter plot. By this, I mean that if you take the regression equation y = ax + b as a prediction of the value of the dependent variable y from a known value of x, the mean error seems to be very large in relation to the range of predicted values of y. Can someone who is fluent with statistical calculations figure out the general relationship here? Given such-and-such correlation and a Gaussian distribution of errors, what is the RMS error of prediction of a single measure from a regression line?

There's another way to view data: in terms of signal-to-noise ratio. This is the ratio of peak-to-peak fluctuations of a signal to RMS noise, where signal and noise are defined in different frequency bands. For ordinary purposes of transmitting quantitative analogue data such as an audio waveform, a signal-to-noise ratio of 6 to 1 is barely tolerable; for high-fidelity purposes, it should be at least 80 decibels, which is a ratio of 10,000 to 1 in amplitude terms. Ordinary meter readings useful for diagnosing electrical system problems need a signal-to-noise ratio of 30:1 or greater (3 per cent accuracy). This latter signal-to-noise ratio is about what we get in tracking experiments for the prediction error using a control-system model. The corresponding correlations are around -0.995. So a correlation of -0.995 implies the lower limit of acceptable noise in a physical measurement or prediction.

Of course, we sometimes have to accept worse signal-to-noise ratios, but the worse the ratio, the less believable is any statement that the theoretical model "predicts" the data. The question is, how bad a fit are we willing to accept while still claiming that the theory has any scientific usefulness?

I think that to claim scientific respectability, we have to insist on very

good fits of theory to data. The reason isn't aesthetics, but the need to be able to make deductions from multiple premises. When a scientific deduction depends on the truth-value of each of several premises that all have to be true for the conclusion to be true, the truth-value of the conclusion is the product of the truth-values of the premises. Four premises *and*ed together to create a conclusion, each premise having an 80 per cent chance of being true, result in a conclusion that has a probability of truth of 0.41. Sad but true.

Any science is built on a foundation of premises that have individually been checked experimentally and found to be acceptably true. A grown-up science is a large structure of logically related statements describing facts of nature. But what kind of science can you have when you can't string together four premises and come up with a conclusion that is probably true? The answer is: a very fragmentary one. You end up with isolated observations that have some small chance of being true in a narrow range of circumstances, but which have to remain isolated because the quality of the data is too low to permit building anything like a complex structure of knowledge.

My chief objection to the way data are analyzed and used in many of the life sciences is that observations of very low precision and repeatability are used just as if they were as precise and repeatable as those of physics. Deductions from premises are made just as if each premise had a truth-value of 1.0. There is an enormous gulf between the achievements of the physical sciences and those of the behavioral sciences. It directly reflects, I think, the difference between a model-based approach to nature, in which very high standards are set, and a statistical approach that provides an excuse for setting very low standards concerning what will be accepted as a true statement.

I have a feeling that we're starting to preach to the converted about statistics. Maybe there is some further point in doing this, and if so, why not? But I'm starting to get the itch to see control theory applied to some real problems some more. There are probably lots of people out there who are searching for applications pertinent to their interests, and who didn't intend to do statistical studies anyway. Of course a lot of participants on this net are in the position of having to develop an interface between control theory and conventional approaches, so maybe that's really what we're doing right now. As we're rejecting 90 per cent of the work being done by hundreds of thousands of well-funded investigators with loads of clout, however, it might be optimistic to think that these arguments are going to sway anyone who doesn't already accept them. There are limits to the vaunted open-mindedness of scientists, no matter what Carl Sagan says in *Parade.* We'll probably get furthest in the end by keeping our noses to our own grindstone as we've been doing for lo, these many years, welcoming those who are interested in joining forces with us, and otherwise ignoring the stuff we no longer believe.

Here is something I worked out, with the help of a mathematics manual, right after I wrote that I was tired of statistics.

Let X be the independent variable (for example, a disturbance acting on a controlled variable). Let Y be the dependent variable (a measure of the action that opposes the disturbance). Let r be the correlation coefficient calculated from N samples of X and Y. The regression equation is then Y = r * (sigy/sigx) * (X - Xbar) + Ybar, where sigx and sigy are the standard deviations of X and Y, and Xbar and Ybar are the average values of X and Y.

The ratio of standard deviations, output/input, is sigy/sigx. This is the scaling factor that represents the average amplification factor applied to the input to produce the output. That ratio takes care of any overall scaling needed to convert X into Y. The correlation coefficient can then range from -1 to 1, indicating the match in waveforms of X and Y (considering them to be time functions).

The standard error of an estimate of Y from X, according to my manual, is given by Sy = sigy sqrt(1 - r^2), or Sy/sigy = sqrt(1 - r^2).

The ratio Sy/sigy is the RMS discrepancy between the predicted and actual values of Y divided by the RMS variation in Y. Because we have pre-scaled the predicted value according to the ratio of sigy/sigx, a complete failure of prediction would make the standard error of the estimate equal to the RMS variations in Y: in other words, Sy/sigy = 1 means complete failure. A perfect prediction would give Sy/sigy = 0. I thus call this measure the "coefficient of failure."

We can now construct a table showing the relationship between the measured correlation of X and Y and the coefficient of failure defined as Sy/sigy.

Per Cent Prediction Failure	|Correlation Coefficient|
0	1.0
3	0.9995
5	0.9987
10	0.995
30	0.954
44	0.900
50	0.86
60	0.80
70	0.71
80	0.60
90	0.43
98	0.20
100	0.0

This is not like an error bar, because the average ratio of Y to X (RMS) has been removed in the calculation of r. A prediction error of 100 per cent is the maximum possible error, representing complete failure. At the low end, the prediction error is approximately the normal proportional error of prediction.

We can see that very high correlations, indeed, are needed to achieve prediction errors of only a few per cent. The error rises drastically as the correlation coefficient falls from 1.0 to 0.8. At a correlation of 0.6, there is an 80 per cent failure of prediction, and at 0.2, a 98 per cent failure (almost total failure).

The "failure of prediction" here is precisely the failure to predict the value of a single point using the regression equation obtained from all of the data points: in other words, the error in predicting individual behavior from the behavior of the aggregate. The significance of the larger errors must be judged not as if on a linear scale, but with the realization that a failure coefficient of 100 per cent means the ultimate degree of failure.

I think that this vindicates my informal estimate that correlations below 0.95 (failure coefficient 0.30) indicate that the model is too far off the mark to use in predicting individual behavior. An individual could actually show the opposite effect at this level of failure, over a significant range of values of the independent variable, with a probability of 50 per cent.

A more sophisticated treatment than I can produce would be needed to show the relationship between the failure coefficient and probabilities of various predictions. But I think the general picture is clear enough.

David Goldstein, I believe, told me that thinking of a regression line as a predictive model is not the normal way to use statistical results. But when mass statistics is used to predict individual behavior, that is exactly how the regression equation is being used. Isn't it?

Gary Cziko: Bill, you provided a very interesting table relating correlation coefficients to your "coefficient of failure." I've never seen this coefficient used before to give an idea of the error involved in predicting individuals based a group correlation coefficient; it would have been an ideal companion to Jimmy Carter's "misery index."

This coefficient is simply the ratio of the standard error of estimate (i.e., the typical amount of error for an individual prediction) compared to how much you would be off just using the mean value of the predicted variable in the sample. Simple enough. But to make sure you weren't pulling a fast one, I worked out a concrete example to convince myself. Perhaps others will find this useful as well, but it is really quite mundane stuff, and those of you who are wise about statistics should probably stop here.

To give a concrete example, I often get a correlation of about 0.60 between height and weight for the ca. 60 students in my (you guessed it) introductory statistics class. Imagine that the mean weight (X) of the class is 60 kg (132 lb) with a standard deviation (SD) of 5 kg, and the mean height (Y) is 160 cm (5′, 3″), with an SD of 10 cm. These figures, along with the correlation coefficient of 0.6, give a regression equation of height = 1.2 * (weight) + 88, so that someone weighing 60 kg would be predicted to be 160 cm tall (that makes sense—someone of average weight is predicted to be of average height).

Now, you say using this regression equation will give a whopping 80% error. Let's see how. Recall that the SD of height is 10 cm. Using the formula for the standard error of estimate (Sy), we get 10 * sqrt(1 - r^2), which, with r = 0.6, gives us Sy = 8 cm. This means that by using this regression, we will typically be off by 8 cm in making our predictions. Not using the regression equation at all, i.e., just using our knowledge of the group mean height (with no knowledge of weight), will give us an error of 10 cm (which is the SD of height). So it looks like you're right in that our typical error in using the regression equation is 80 per cent of what it would be if it were not used at all. Or, we could say that a correlation coefficient of 0.6 reduces error by only 20 per cent (should this be called the "coefficient of success"?).

Now, this example is a bit silly, because if I have both the height and weight of my students, and I want to know their height, I will not use a regression equation to predict their height—I will just look at the height I have already measured. If I were to be brave and predict the heights of my *next* class based on just their weights, my predictions would most likely be significantly worse than the original 80 per cent error, even if they were from the same population, whatever that means. Hmm.

Only two problems remain. First, why is it that statisticians always talk about r-square, the misnamed "coefficient of determination"? They would take my r = 0.6, square it to get 0.36, and then say that variation in weight explains 36 per cent of the variation in height. This 36 per cent is not great, but it does look better than a coefficient of failure of 80 per cent or coefficient of success of 20 per cent. I've yet to figure out how r-square relates to these two new quite pessimistic indices of the predictive power of regression equations.

Second, you have been arguing that adding in more predictors makes the error even worse. But typically, adding more predictors does increase the absolute value of the correlation coefficient (multiple r), which, by your own table, *reduces* the coefficient of failure. I can't see how your argument holds, unless you get into problems of sampling and cross-sample validation.

Mark Olson: I just wanted to thank those of you who explained the difference between the smoking/cancer and reading/feet situations. I think the statement that "there is no difference between the two except the assumptions one brings to each" is what "enlightened" me.

Gary Cziko: As a follow-up to my last post, I just discovered that Bill Powers' "coefficient of failure" does appear in one of my statistics books, where it is called the "coefficient of alienation" and is calculated as k = sqrt(1 - r^2). It would be interesting to see how many statistics books even mention this coefficient.

I would prefer to call it the coefficient of "uselessness," since it tells how useless a predictor (or group of predictors in multiple regression) is in predicting the Y of an individual.

I recently had a colleague give a presentation showing how, using all sorts of measures in the right combination, he can obtain a multiple r of 0.5 in predicting children's adjustment/happiness in school. He justified this by saying that this is about the best you can get in the social sciences. I wish I had been able to tell him that his findings were 86 per cent useless in predicting the adjustment/happiness of individual children.

Finally, it occurs to me that r-square looks better than k because the former does not depend upon making predictions for individuals but uses the rather more abstract concept of "shared" or "explained" variance.

Bill Powers: Gary, if I understand Phil Runkel's argument, what you gain by adding more predictors is more than offset by the smaller N in each group. If you had started with only one predictor (weight predicts height) in your class of 60, the N is 60. If you now add, say, grip strength as a second indicator of height, you now have at least four combinations of independent variables instead of one: high-high, high-low, low-low, and low-high. Each subgroup now has only 15 students in it. One-fourth the N means twice the standard error. Now, in order to fit the prediction, a person not only has to be heavier than average and taller than average, but also stronger than average. All you've done is to eliminate some of the heavier people who are taller. Even if the N in the high-high group is larger than in the other three groups, I think you always lose some predictivity. If you don't add any new people to increase N, it seems to me that you've just cut down the number of people who fit all the criteria: instead of just heavier and taller, they have to be heaver, stronger, and taller. I think that this is what Phil Runkel calls fine-slicing.

I don't know how to work this out mathematically. Can you do something analogous to what I did with the one-dimensional case?

My hunch is that the higher correlations found in multiple regressions are offset by the increased standard error, or more than offset. Higher correlation, but higher uselessness index—maybe.

As to "explained variance," individual measures don't have any variance, do they?

Gary Cziko: This continues the discussion about how group statistics are not very useful for making decisions about individuals.

Effect sizes have become a commonly used metric in educational research to describe the difference between an experimental group (e.g., new way of teaching math) and a control group (e.g., old way of teaching math). The effect size is the difference in means divided by the standard deviation. So if the standard deviation of the math test is 10, and the experimental group mean after treatment is 55 compared to the control group at 50, there is a 0.5 effect size.

For some reason, an effect size of at least 0.5 has become accepted as indicating that there is a practically significant difference between the two groups, hence the new method is better than the old. I wouldn't be surprised if a similar standard has become adopted in other areas, for example in medical research. One positive consequence of using effect sizes is that it gets around the problem of tiny differences being "highly statistically significant" simply because one has used large samples.

But let's see just how exciting an effect size of 0.5 really is. With two normal distributions whose means are separated by 0.5 standard deviation, we find that 31 per cent (almost one-third) of the individuals in the low group are actually higher than the mean of the high group. Also, an additional 38 per cent of low-group individuals will not be more than one standard deviation below the mean of the high group. This gives us a total of 69 per cent of low-group individuals which are either higher than the mean of the high group or not more than one standard deviation below the high mean. The same, of course, could be said conversely of the high-group individuals (69 per cent are lower or not more than one standard deviation above the mean of the low group).

An effect size of 0.5 does not seem very impressive in making predictions about individuals.

Chuck Tucker: The discussion on statistics is wonderful. I hope that all of you who teach statistics will incorporate these ideas in your courses and make it a point to catch those who claim they are not interested in individuals (that is the retort in my sociology department) when they try to use statistics to talk about them.

Bill Powers: Gary, I hadn't heard about "effect sizes." Half a standard deviation? Surely you jest. Do people ever actually replicate studies of this sort? I approve of getting rid of statistical significance that's based mainly on large N, but is it an improvement to accept smaller N and also relax the meaning of significance even further ("practical significance")?

You say: "One positive consequence of using effect sizes is that it gets around the problem of tiny differences being 'highly statistically significant' simply because one has used large samples." Now you can get significance with tiny differences, even without using a large sample. It seems to me that someone is trying to recycle the garbage. How to do a bad experiment and still get it published?

Rick Marken: I want to just say "bravo" to all those involved in the statistics discussion. I don't think any conventional psychologists will be converted from the statistical to the modeling game, but it's nice to point out the problems for posterity, and for the unconverted who could contribute to the development of a science of life.

Martin Taylor: Gary defines "effect size" as the difference between the means of two distributions measured in units of the standard deviation. In psychophysics, this measure is called d′ ("d-prime"), and a d′ of 1 is taken as roughly what people mean when they say that there is a "threshold" effect. A subject will usually not claim to have detected an individual signal at a level giving a d′ much less than unity, but will usually claim to have detected an individual signal at a level giving a d′ appreciably greater than unity. Gary says that in educational research, an effect size of 0.5 is taken as practically significant, and he thinks the same is true of other areas. In psychophysics, the usual equivalent is an effect size of unity, which seems appropriate, given that the subjects in an experiment *are* working with individuals, and unity is roughly the d′ that separates conscious detection from non-detection.

Gary Cziko: Martin, could you provide a bit more information about what the psychophysical "effect size" d′ is as used in psychophysics? You say: "A subject will usually not claim to have detected an individual signal at a level giving a d′ much less than unity, but will usually claim to have detected an individual signal at a level giving a d′ appreciably greater than unity." Are you referring to a type of signal-to-noise ratio here? If this is analogous to the effect size in educational research, what are your two means, and what is your standard deviation based on? I suppose a simple example would help us non-psychophysicists to understand this.

I would guess that psychophysics should be of some interest to control theorists, since, as I understand it, it uses the method of specimens (one individual at a time to find invariant laws) in much the same way that control theory does.

As a follow-up to my previous post, I have constructed a table to show how various effect sizes can be used to make predictions about individuals in low” and “high” groups. The table assumes Normal distributions. I wouldn’t be surprised if I made some typos or calculation errors here, but the numbers all go in the right direction, so there are no obvious errors.

In the definitions below, the words “low,” “lower,” and ˉbelow” can be interchanged with “high,” “higher,” and “above,” respectively.

A = Effect size, (Xbar-Ybar)/SD

B = proportion of low scores higher than mean of high group (“surprises”)

C = proportion of low group no more than 1 SD lower than mean of high group (low group scores as close to high mean as typical high group score is to high mean)

D = total of B and C (total proportion of low group scores easily construed as being part of high group)

A	B	C	D
0.50	0.3085	0.3830	0.6915
0.75	0.2266	0.3721	0.5519
1.00	0.1587	0.3413	0.5000
1.25	0.1056	0.2954	0.4010
1.50	0.0668	0.2417	0.3085
1.75	0.0401	0.1865	0.2266
2.00	0.0228	0.1359	0.1587
2.25	0.0122	0.0934	0.1054
2.50	0.0062	0.0606	0.0668
2.75	0.0030	0.0371	0.0401
3.00	0.0013	0.0215	0.0228

Column D is most informative (and most damaging) because it gives the total proportion of individuals in the low group who would not be out of place in the high group (or vice versa).

Note that at the “practically significant” (in educational research, anyway) ES of 0.5, more than two-thirds of the low group fit nicely into the high group (and vice versa). Even at a “whopping” ES of 1.00 (equivalent to a difference in mean IQ of 16 points, for example), this is still the case for half the individuals in each group. It is only when we reach a “mammoth” ES of close to 1.75 that this proportion drops to less than 0.25. An ES of 2.75 is nice, since then the proportion is less than 0.05. Has anybody ever seen one this big in the social sciences? Perhaps the difference in height between Pygmies and Dinkas in Africa.

Of course, all this looks even worse when we try to use findings like these to make predictions about *new* individuals who were not part of the original data, and who might or might not be considered part of the same population (whatever that means).

Bill Powers: Gary, if you think about publishing this sort of analysis, I hope you’ll make the paper a comparison of what’s good for education as opposed to what’s good for the student. What’s good for education is, of course, a good track record. What’s good for each student is to be evaluated accurately, to be treated appropriately, and to learn successfully. What we’ve been doing in these posts is developing a way to show that the goals of educators can be met, while, in significant numbers of cases, those of students are not. It’s no good to point out, as defenders of the present methods will do, that substantial numbers of students are treated properly. We have to focus on those who are misjudged by the statistics. Even with two standard deviations between group means, one student in six will be treated as if he or she belongs in the wrong group, according to your chart. In a class of 30, that’s five people about whom the teacher will get the wrong idea. I don’t think that this kind of misevaluation is harmless. It ought to be actionable on the basis of an implied warranty.

All this would be more convincing if we could come up with a way to apply control theory in teaching or testing that would work better than the present methods. Let’s talk about it.

What I am hostile to is the misuse of group statistics. If you want to compare two methods or two tests to see which is “better” with respect to producing or measuring some aggregate phenomenon, statistics works fine. Just don’t make the mistake of using the methods or the tests to evaluate individuals. Not unless your correlations are running 0.99 or better.

Gary Cziko: Bill says: “What I am hostile to is the misuse of group statistics. If you want to compare two methods or two tests to see which is ‘better‘ with respect to producing or measuring some aggregate phenomenon, statistics works fine.” But even this idea seems based on a linear, one-way view of causality which does not seem compatible with control theory. Much (if not most) of quantitative educational research is determined to show that certain combinations of inputs (“independent” variables) will give you certain outputs (“dependent” variables), and of course group statistics is used to try to do just this. Results have been rather dismal so far, but that just means that not enough variables were taken into account, or the measures were not reliable/valid enough, or the statistical analyses were not abstruse enough (structural equation modeling using a program called LISREL is the latest trend in statistical analysis). This is done, of course, in the hope that once the input-to-output links are known, teachers and administrators can better control the behavior (i.e., success, achievement, drop-out rate, motivation, etc.) of their students. It seems that even your statement seems to imply an input-to-output view.

Group statistics seem to be used in at least four ways in educational research:

(1) to tell us about the psychological processes/functioning of students;

(2) to make predictions about individuals;

(3) to find out what combinations of input variables (e.g., teaching method) cause certain patterns of output variables (e.g., mathematics achievement); and

(4) for polling (survey) research.

Runkel's book and your *American Behavioral Scientist* article do what I feel is a convincing job to debunk the first. Our recent discussion about individual predictions using correlations and effect sizes addresses what appear to be serious problems with the second. We are discussing the third now. It might be that only the fourth is a legitimate use (if we can figure out what a random sample is and don't worry too much about the problems that the Bayesians point out).

Fred Davidson: In response to the recent discussion of statistics, effect sizes, and what's-good-for-the-student (Cziko, Powers, and others), I recommend J. R. Frederiksen and A. Collins, "A Systems Approach to Educational Testing," *Educational Researcher 18(9),* 1989,27-32. There are many in educational testing who would love to see the downfall of norm-referenced epistemologies. Frederiksen and Collins propose an elegant new "validity" (= truth) of measurement: "systemic validity." They say: "Evidence for systemic validity would be an improvement in those skills [which the test claims to measure] after the test has been in place within the educational system for a long time." (p. 27)

In language testing, we call this "backwash"—the effect of testing on instruction. We backwashers believe that testing is the servant of successful learning. That's a concept that the quasi-scientific, clinical, detached, norm-referenced-measurement establishment seems to have forgotten. I like "systemic validity" better than "backwash," since the former elevates the concept to the level of a "validity"; there are about four validities taught in educational measurement courses: face, content, criterion (predictive and concurrent), and construct. Politically, that is a good idea.

Now to control theory: I suspect that control theory offers a way to further justify systemic validity/backwash. Isn't successful learning also a well-functioning control system?

Bill Powers: Gary, I said that group statistics can be used to compare methods or tests. You said: "But even this idea seems based on a linear, one-way view of causality which does not seem compatible with control theory. Much (if not most) of quantitative educational research is determined to show that certain combinations of inputs ('independent' variables) will give you certain outputs ('dependent' variables), and of course group statistics is used to try to do just this." We have to be careful about treating control theory as a dogma with which we must keep faith. If a lineal cause-effect model could predict individual behavior accurately, we would have to accept it as a contender against control theory. We don't really need to consider control theory when evaluating a cause-effect explanation of behavior. If we reject a cause-effect explanation, we should do so on the basis that it predicts poorly, not because it violates the precepts of control theory or because there's something that says cause-effect systems can't exist. This means we judge against standards of prediction. So where are we to set those standards? Is a measure that has a uselessness index of 60 per cent OK? Are we willing to accept the many wrong predictions that result from such a low standard? If so, then, as Rick Marken would say, go for it. It would certainly make life easy for those who need to publish regularly. But this isn't how you achieve real knowledge about nature.

What it all comes down to is a system concept. What kind of science do you want to mean when you call yourself a scientist?

Of course, I agree with you about the cause-effect approach. It isn't really even a model, because it tries to explain the output on the basis of the input without any idea at all of what goes on between them. That's truly just floundering around in the dark. You don't even know if the change of behavior isn't produced to counteract the effect of the input!

But I don't think that we've effectively debunked anything yet. How many conventional educators have called you up all weepy and apologetic and promised that they'll stop doing those bad things? I think we have to concentrate on finding something that works better, so it can be taught and used. That's the only thing we can offer that will change anyone's mind. Nobody will prefer a method that works worse over one that works better. Not for long.

Gary Cziko: Bill says: "If we reject a cause-effect explanation, we should do so on the basis that it predicts poorly, not because it violates the precepts of control theory or because there's something that says cause-effect systems can't exist." Yes, I basically agree with this, although I wonder what your reaction would be to someone who wants to show you a perpetual-motion machine (perhaps even one that can do work). I suppose you should ask to see if it works, although most of us wouldn't waste our time, since all we know about physics says such machines can't work. But, yes, control theory has nowhere near the status of the laws of thermodynamics, so we need to keep our eyes open to see what works.

Now, here's a concrete problem. I've been showing the "random" program which you describe in your article in *American Behavioral Scientist,* September/October 1990. One reaction I get is that a multiple regression (MR) could make good sense of these data if you included the reference level, cost, and wage variables. Something tells me that this is *not* the case, since this would still be an analysis of relative frequencies, not a test of individuals.

What I'd really like to do is to get the program to generate some data which I could try to analyze using MR (or better yet, give it to one of the many MR-whizzes around here) and see what could be done. So my two questions are:

1. Would it be possible and worthwhile to get a data matrix from this program for such an analysis?
2. Do you have any ideas about what MR analysis could reveal about such data? Could it find that reward is under fairly tight control and that costs and wages are disturbances?

I hope that those who are familiar with this article and know something about MR analysis will join in here.

Bill Powers: [In reply to a post by Peter Parzer, in the Department of Psychology at the University of Vienna.] It's now beginning to look as though we have been using the concept of correlation incorrectly in talking about our tracking experiments. When we speak of using a model to predict behavior, the independent variable used both for the model and for the real person is predetermined and exactly known (i.e., not a random variable). This implies that we shouldn't be talking about the "correlation" of the independent variable with the dependent one. Intuitively, we have realized that when you get correlations of 0.99 and up, correlation ceases to be a very useful measure and starts becoming a tool for making an impression on someone. The more useful measure is just the RMS error of prediction in proportion to the range of the expected value, which I have already referred to as the signal-to-noise ratio.

I'm not sure of this conclusion, however. Perhaps if I describe a basic experiment, Peter can tell us the right measure to use.

The task is for a person to use a control handle to keep a movable object on the screen aligned between two "target" marks. The position of the movable object (the "cursor") is determined by the sum of two numbers: one represents handle position relative to the midpoint, and the other is a time-varying disturbance generated by smoothing and

scaling a table of random numbers. When the target marks are stationary (the simplest case, "compensatory tracking"), accomplishing the task perfectly implies moving the handle in exact opposition to the disturbance, so the net effect on the cursor remains zero (which is the position between the target marks). The disturbance thus becomes an independent variable that predicts handle position.

The disturbing function itself is invisible, being applied inside the computer that runs the experiment. Stabilization of the cursor is not, of course, perfect; the cursor wobbles slightly up and down during a typical one-minute run. Its wobbles do not resemble the variations in the disturbance. The data consist of 1800 samples of cursor and handle position (the disturbance waveform is stored beforehand), or one set of samples every 1/30 second (more or less, depending on which computer is used).

The model used is that of a control system, which for this case is indistinguishable from a stimulus-response system except for the fact that the most obvious "stimulus," the cursor position, is continuously dependent on the "response," the handle position, as well as on the "independent variable," the disturbance waveform. In addition, all variables are continuous, instead of discrete as is usually assumed in stimulus-response analyses. The control-system model that we use most commonly also puts one time-integration into the output of the system. The output is a constant times the time integral of the deviation of the cursor from the target marks. For slow variations of the disturbance, this integrating model works only slightly better than a pure proportional model.

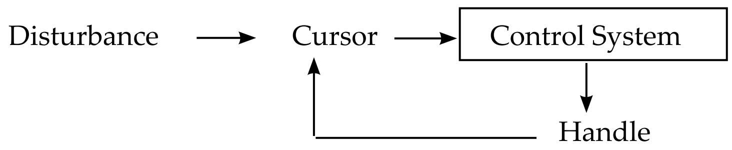
The subject and the model are both run with the same disturbing waveform. This enables us to find the value of the integrating constant (or gain of the control system for the proportional case) that makes the model fit the data best. Typical errors of fit are about three per cent RMS of the peak-to-peak excursions of the handle. Next, a new disturbing waveform is generated by the computer and the model is run using the parameters already obtained. This result is now a prediction of the way the subject will move the handle when the same new disturbance is applied during a "live" run. The errors of prediction are typically three per cent to five per cent of the handle excursion.

Predictions of the *cursor* position are not so accurate, because the cursor position represents the difference between the handle position and the optimal position called for by the magnitude of the disturbance at any given instant. For very slow disturbances, the cursor prediction error can be quite large—100 per cent RMS or more. But the more difficult the disturbance (so that stabilization errors become larger), the better the prediction, the RMS error dropping sometimes to 10 per cent of the cursor excursion.

Correlations of cursor position against handle position are probably meaningful, because unsystematic tracking errors are seen; these correlations are typically 0.2 or less (positive or negative), becoming smaller as the task gets easier.

We have also been calculating correlations between the momentary handle positions and the momentary magnitudes of disturbance. The disturbance variations, however, are accurately known, so this "independent variable" is not really random, although it is derived from a table of random numbers. In principle, because of the smoothing used to limit the speed of variation of the disturbance, some short-term prediction of the independent variable is possible (for this reason, some workers have proposed that control systems must contain predictors). Our model, however, does no predicting, and it works well enough that I don't think we need to add such a feature to the model. But the question still remains as to whether the disturbance should be considered a random variable or a given variable. That's what I'm asking Peter to think about, if this explanation of the experiments has given enough information to allow making a judgment.

[Following another post from Peter Panzer.] Before proceeding, I'd like to clear up the nature of the control-system model, as well as our way of using it for predictions. Let's see if I can construct a diagram that will make the relations clearer:



The effect of the disturbance on the cursor position occurs inside the computer; the disturbance itself cannot be seen by the subject except through its effects on the cursor. The handle also affects cursor position at the same time. So the input to the control system (visible cursor position) is not independent of the output (handle position). The true independent variable is the disturbance, a slowly and continuously varying waveform. The disturbance and the handle position affect the cursor at the same time, so cursor position depends jointly on the disturbance magnitude and the handle position. The behavior of the cursor does not reveal what either the disturbance alone or the handle alone is doing.

We can measure the cursor position within about one part in 350 to 480 of the maximum possible excursion on the computer screen (depending on display resolution). We can measure the handle position (in my equipment) to one part in 4096 of the maximum possible handle excursion, give or take a per cent of nonlinearity. The disturbance values are known exactly. So we really aren't talking about errors in measuring the input or the output, are we? We know what the input and output are with relatively high precision. The problem is to guess how the control system in the box is organized such that it produces the observed relationship.

If t represents the stationary target position (zero by definition), c represents the cursor position, and h represents the handle position, the simplest model that seems to predict well has the form: h′ = k * integral(c′ - t) * dt, where dt = about 1/30 second. The experimental apparatus is set up so that (exactly) c = h + d, where d is the current magnitude of the disturbance, and h is the current measured position of the handle.

The model is run by solving these two equations simultaneously via simulation, since d is not an analytical function of time. The variables c and h are given initial values, and then the disturbance is run through all its values while the values of c′ and h′ are computed over and over, yielding tables showing positions as a function of time. The subject is run by being put in the same relationship to the apparatus as the box labeled Control System, above.

The primes in the expressions designate the *predicted* values of c and h. Let *c* and h (without primes) represent the observed values (from a run with a real subject). We are then interested in the departure of c from *c′* and of h from h′. Generally, the RMS departures are enough larger than the errors of measurement of c and h that we can ignore those errors of measurement.

We can measure both the model's and the subject's handle positions with an accuracy of much less than one per cent. We take the subject's handle position as the definition of zero error, and evaluate the model's error of prediction by comparing its simulated handle positions with those of the subject over the course of the experimental run. It seems to me that this definition of prediction error is not arbitrary or model-dependent [as suggested by Peter].

What is arbitrary, of course, is the form of the model in the box labeled Control System. There is actually more in that box than is discussed here, because we have to be able to account for other cases—for example, the case in which the subject holds the cursor some fixed distance *away* from the target marks. We have picked the simplest model that accounts adequately for the data. More complex models can slightly improve the results. For example, by putting a time-delay of about 0.15 second into the model, we can halve the RMS prediction error. But it's always possible that Mother Nature has put something else into the Control System box. All we can do is make our best guess and hope that more detailed data about the neuromuscular systems will help us to find a still better model. But as the simplest model leaves only about three to five per cent difference between model and reality, we aren't going to gain much more accuracy.

There are two steps in making a prediction. First we match the model to the behavior as well as possible by adjusting k in the equation above. Then we generate a new waveform for the disturbance (when we're fussy we require that it correlate less than 0.2 with the former one) and use that to make a predicted run, with the previously found value of k (the only adjustable parameter). The predicted handle waveform will be different from before because the disturbance waveform is different. Finally, the (same) subject's behavior with the new disturbance waveform is recorded and compared with the prediction. This latter step, in which the model is used first under new conditions, is what we call a true prediction. The RMS difference between model and real handle positions in the second step is typically three to five per cent. Tom Bourbon has shown that this same accuracy of prediction is found even with a lapse of *one* year between the prediction and the real run. The property represented by k thus appears quite stable over time, although it differs markedly (2:1) between individuals.

We have not said where the random errors come from in our model, but clearly they have to be coming from inside the subject, because our knowledge of d, c, and h is relatively exact.

I wonder if it still seems to Peter that there is no difference between the statistical and the model-based approaches (at least ours)? I have a suspicion that the way we are using the term "model" isn't quite the same as the way Peter is using it.

Gary, here is the part of the "random" program that generates the data:

```
for i := 0 to maxdata do begin
b := 1.5 + 3.5 * random; {for Hercules and EGA}
  k := 5.0;
  d := -random(40);
  r0 := 100 + random(200);
  effort := k * (r0 - d)/ (1.0 + k * b);
  reward := (b * k * r0 + d) / (1.0 + k * b);
  v2[i] := round(effort); v1[i] := round(reward);
ref[i] := r0;
end;
```

I set maxdata to 4000, but there's no need to go that far. The error sensitivity is fixed at 5.0 (k). The "cost" is d; the "wages" are b. The resulting effort and reward figures for each person are stored in two arrays: v1 (effort) and v2 (reward). The reference signals (amount of reward desired) are stored in the array ref. The entries in the reward and effort arrays amount to a single determination for each person.

In the article, I pointed out that in order to measure the reference signal for each person, it would be necessary to do a control-system type of experiment with every individual. You would have to vary the disturbance to find out what level of reward leads to zero effort *in each individual* (the definition of a measured reference level of a controlled quantity). As presented, the data do not show this: we know the internal reference setting for each person only because we know the correct model for each person. For an experimenter who does not know about reference signals, there is nothing to indicate their settings. The only externally observable variables are effort and reward.

I doubt that MR analysis would reveal the reference levels for each person. The concept of a reference level, a preferred level of input, is model-dependent, and here the model is that of a control system, not an input-output system. Similarly for the idea of error sensitivity (k). You can't measure k for an individual from a *single* observation. The loop gain of the system can't be seen unless you vary the disturbance and observe how much the disturbed variable, the reward in this case, changes. The loop gain would be the ratio of the disturbance magnitude to the change in reward relative to the no-disturbance value, minus 1. We know the external part of the loop gain (the wage) but must deduce the internal part, the error sensitivity k. I don't think any of these concepts are part of the model assumed under MR analysis.

The above program would be easy to implement in BASIC or any other language, or even on a spreadsheet. Rick Marken has done control systems on spreadsheets. Most statistics packages, I believe, can import data from spreadsheets.

You also said: "... I wonder what your reaction would be to someone who wants to show you a perpetual-motion machine (perhaps even one than can do work)." After all my experiences with control theory, I wouldn't reject a working perpetual motion machine on principle. But I would like to be alone with it for half an hour, with a few hand-tools.

Tom Bourbon: To Peter Parzer: I have enjoyed watching the dialogue between Bill Powers and you. You have certainly raised some important points concerning the nature of modeling. The most significant reminder you made for me is that the selection of variables and metrics is always in the hands of the modeler and can be done in various ways that can enhance the apparent success of the modeling enterprise.

As for the reliance on correlations in presentations of the results of modeling by control theorists, that selection was driven in part by a desire to have at least the index of performance be familiar to psychologists and other behavioral scientists, the majority of whom never work with continuous variables, and who never use other indices, such as RMS error.

Bill Powers: (In reply to a post by Peter Parzer.] It seems to me that the simplest comparison between the simulated handle position and the observed handle position would be a plot of the differences between them for the 1800 data points in a tracking run. We want to do this so that we can compare different models and see which predicts the results the best. We could simply look at two plots of prediction error against time and say, "Ah, the first one stays closer to zero over most of the points." Or, more likely, we would look for some measure that would be more reproducible over observers, such as the RMS error calculated for all the data points. As you imply, there isn't any "objectively right" way to measure overall error. But there are ways that are useful, simple, and reproducible.

Whether absolute or relative errors are used depends on the application. If you're talking about arithmetical calculation errors, absolute error is all that makes sense—after all, the relative error is always zero, in comparison with the range of values that numbers can take on (infinity). On the other hand, if you're judging how well a person steers a car, relative error makes sense, because what matters is how much the car wanders in relation to the width of its lane. I agree that there is a choice, but usually there's a pretty good reason for the choice. There's no one measure of error that suits all occasions.

In a tracking experiment, we have a record of 1800 positions of the handle. The model reproduces these positions with some error. But why should we assume that the errors we see are due to a random variable in the subject? Why shouldn't we assume that the model still does not capture all the properties of the real system correctly and that the remaining errors are systematic? Indeed, we find that when we refine the tracking model—for example, by putting in that time-lag I mentioned—the prediction errors become significantly smaller. In one experiment, the RMS errors of prediction dropped from 3 per cent to 1.5 per cent (noise-to-signal ratio). That tells us that at least half the error we obtained before was not random. Why should we assume that all of the remaining error is random? Of course, at some point we will run into what looks like a basic noise level, but the errors are already so small that they're approaching those of a physical measurement. When you speak of an "adequate" model, you have to ask "adequate for what purpose"? I think that in terms of predicting simple behavioral phenomena, the control-system model is adequately precise for any purpose we can now imagine. Our biggest problems now are in modeling more complex behavior.

The difference between models and statistical analysis really comes down to a difference in basic assumptions. I assume that prediction errors occur because although the person's behavior is completely systematic, the model is not yet exactly correct. It might not have been apparent, come to think of it, that when we speak of predicting handle movements in the tracking task, we mean predicting all details of movement with quantitative accuracy, not just comparing mean slopes or other average measures. The tracking model generates a trace of simulated handle movements that can be laid right over the trace of the real handle movements. It's hard to realize that the two simple equations I presented can do this, but they really can.

The other assumption would be that the model must be correct (for some philosophical reason), so the prediction errors are the organism's fault. Psychologists decided long ago that the variability of behavior was caused not by an inadequacy of their lineal cause-effect model, but by some inherent randomness of behavior. I have always felt that they gave up about 150 years too soon. We will surely have to give up trying to improve our models some day, but I would rather see that day come when "random" errors of prediction are in the 1 per cent range rather than the 100 per cent to 1000 per cent range.

In the models we use, not only the variables have empirical meaning, but the individual relationships between them have empirical meaning, or at least a proposed empirical meaning. We propose, for example, that an error signal results from neurally subtracting a perceptual signal from a reference signal. The subtraction process is part of the physical model. In the tracking experiment, d, c, and h have empirical meaning, but so does the relationship c = h + d. If we gave the handle twice as much effect on the cursor, the relationship would be c = 2 * h + d. This part of the model embodies known physical relationships. The other equation proposes physical relationships inside the control system. The behavior of the system grows out of the interaction of these two aspects of the model.

We use "generative" models. That is, they do not directly represent behavior, but propose an underlying physical organization that creates behavior because of its inputs and the way it treats signals internally. Such models predict not only the specific input-output relations observed in a single experiment, but a whole family of relations that can be seen under many different experimental conditions. The model I described for the tracking experiment, for example, predicts just as accurately when we make the target position a function of time, without any change in the parameter k (still applying a disturbance directly to the cursor as before), and when we halve or double the effect of a given handle movement on the cursor. Most experimental psychologists who actually try these experiments find the generality and accuracy of the models to be little short of uncanny—especially in comparison with what they're used to.

This is why I can't get too excited over just how we measure prediction errors. We're talking about errors an order of magnitude smaller than those that are usually seen in behavioral experiments (outside psychophysics).

Martin Taylor: Gary, one could indeed say d′ is a measure of signal-to-noise ratio in some abstract sense. Given an ideal observer under specified constraints on information gathering, one can determine the SNR that gives a specific d′. (Actually, it is signal energy rather than power that usually determines the d′, but the details always depend on the observing constraints). One asserts that there exists some perturbation of the observation (noise) that can move a non-signal observation to a more signal-like state, or a signal observation to a more noise-alone state. If the signal is weak enough, the distributions induced by the perturbations can overlap. One asserts furthermore that there is some criterion on which the observer makes a judgment as to whether a signal was present, and that "signal" is more likely the greater the value of the observation on this criterion. If the criterion axis can be transformed (squashed) so that the perturbation-induced distributions take on a Normal form, and particularly if the Normal distribution has the same variance whether or not a signal was present, then d′ is the distance between the means of the distributions in units of their common standard deviation. In more complex situations, the definition is different, but related. With common Normal distributions, it is exactly your "effect size," and unity is often taken to be the dividing line between "perceptually nonexistent" and "perceptually valid," though the subject sees each individual signal presentation as there or not, regardless of d′. The problem for the subject is that the signal might be perceptually there when none was presented, or not there when one was presented.

Perception *is* a problem of statistics, and treating it (properly, in my view) as a control problem will not make that go away.

Bill Powers: Martin, I agree that statistics can enter into perception, but I doubt that a properly designed "test for the controlled variable" (which identifies controlled perceptions, as nearly as we can) will leave us worrying about effect sizes and standard deviations in the way you suggest. When you've identified a controlled variable using control theory, it's pretty unequivocal.

In control theory, we seldom do experiments with perceptions at their lower limits of detection. The normal case, which I think represents the overwhelming majority of real cases, involves perceptual variables that are far above their thresholds of detection or discrimination, and neural signal frequencies that are comfortably above the levels where individual impulses have any appreciable effects. After we have models that function well in this middle range, we might want to explore behavior and perception near the limits of operation where noise becomes a significant consideration. But I don't think we've reached that point yet.

Rick Marken: Here is another thought I had about statistics—just to see if it can stir up some comment. The previous statistics discussion has dealt mainly with the problem of using group-level statistics to form conclusions about individual processes. This was approached in several ways—in particular, showing that even relatively high group-level correlations imply substantial error in individual prediction (the coefficient of failure).

But group-level statistics do work on groups. Lowering my cholesterol intake might not help me personally (indeed, it might kill me), but that does not diminish the fact that, at the group level, there is evidence of lowered heart disease with lowered cholesterol intake. This is "true" at the population level. On PBS last night, they reported that a government program to reduce dietary fat in Finland has led to a 30% decrease in heart disease. Ignoring the problems of attributing all of that 30% to the dietary change, this is evidence of a group-level change having a group-level effect. The same thing happens with seat belts. Death rates, at the group level, do (I believe) decrease substantially with mandatory seat belt laws—even though this is not necessarily the case individually. In fact, many people who might have survived an accident (like a burning car) were probably killed because they were wearing their seat belt. But overall, the death rate does go down.

That's the basis of my question. What do you folks think of this problem? Apparently, we can have some control over group data by doing things individually which might not be in our best interests. Apparently, we can influence the group-level rate of heart disease by collectively (but as individuals) reducing fat intake. We can do this even though some of us, individually, might actually be worse off as the result of taking that action (though we can't know that, of course, because we only have the poorly predictive group-level data to go on). This seems like a crazy paradox; and it seems to occur a lot in society. "Should I ignore the potential group-level good and continue to do what I want based on the extremely good argument that it is meaningless to base my individual actions on group-level data? Or should I cooperate with the statisticians in order to produce a beneficial group-level result by taking action that could possibly have negative individual consequences?"

If the data say "80% of people who take X get cancer," and (1) I like X, but (2) I don't want to get cancer, isn't it a good bet for me to avoid X? (Assume that I like X *far less* than I dislike cancer).

Gary Cziko: The answer to Rick's last question depends on how much he likes X and how much he dislikes cancer. This is the stuff of classical decision theory. A nice introduction to this kind of thinking can be found in Ronald Giere's *Explaining Science.* (But Bill Powers would probably add to this that it also depends on how similar you think you are to the 80% of people who get cancer doing X.)

Here are two quotes from J. G. Taylor, "Experimental Design: A Cloak for Intellectual Sterility," *British Journal of Psychology* 49, 1958, 106-116.

> If Newton had had at his disposal not a vast amount of detailed information about a single solar system but a much smaller number of facts about each of a thousand solar systems, collected by a thousand observatories, he might conceivably have developed statistical methods for organizing this material. He might have found correlations between such variables as the number of planets in the system, the average number of satellites per planet, the average distance of the planets from the sun, and the like. He would, by this means, have learned a good deal about solar systems in general, but he could not have calculated the time and place of the next eclipse of the sun, and he could not have arrived at an understanding of the laws of planetary motion. He would have learned a lot about the ways in which solar systems differ from one another, but nothing about the ways in which any one of them works. For this it was necessary to know as much as possible about one system. Fortunately Newton had no alternative, and the result of his labours was the construction of a theory that survived until the advent of Einstein's theory of relativity. (p. 109)

> Suppose that an investigator, knowing nothing about the construction of a motor car, decided to choose as his area of research the behaviour of the speedometer needle, and to this end took a series of readings in each of a hundred different models. Just to make the problem more like a real one we shall suppose that the speedometer dials are not provided with scales, but that the investigator can measure the angular deviation of the needle. Among the variables he might be expected to record are the distances of the accelerator and brake pedals from the floor, the position of the gear lever, the gradient of the road, the direction and velocity of the wind, and, of course, the speedometer reading. He takes a succession of simultaneous readings of all those variables in each car, and then proceeds to examine his data in the hope of solving the riddle of the speedometer needle. At first the material looks completely chaotic. There is no single independent variable that is functionally related to the dependent variable, and he decides to have recourse to statistical analysis. He finds negative correlations between the speedometer reading and (a) the distance of the accelerator pedal from the floor, and (b) the gradient of the road; and positive correlations with (c) the position of the gear lever, and (d) the distance of the brake pedal from the floor. He finds significant differences between the speedometer readings when the gear lever is in first, second, third, and fourth positions, but the distributions overlap extensively. He now decides to record additional data, such as the weight of the car and its consumption of petrol, but the riddle remains unsolved. Of course we know the answer. If our investigator will only take independent measurements of the speed of the car he will find that in each system (car) the speedometer reading is a function of speed, but not necessarily the same function in all systems. He will find, moreover, that he can now dispense with statistical methods and can examine each system, considered as a matrix of pointer readings representing the several recorded variables, to determine how it hangs together. He will discover that what he at first took to be evidence of arbitrariness or caprice in his data was actually an artifact arising from the simultaneous examination of pointer readings taken from a hundred different systems. He will find that the same general principles apply to all the systems, but each of them has its own specific set of parameters, with the result that, in Ashby's (1952) terminology, the lines of behaviour of all the systems are different. Continuing to use Ashby's terms, each system is regular and absolute. It is regular because whenever it starts from a given state and a primary operation is applied to it, such as an increase in the gradient of the road or a specific depression of the accelerator pedal, the system will change to another state, and always to the same state. It is absolute because this is true no matter how the given initial state was arrived at." (pp. 110-111)

I'm not sure that even Bill Powers or Phil Runkel could say it better than this.

Rick Marken: Thanks to Gary Cziko for his response to my little statistical question. I'll tell you why I asked. I had a discussion with my wife and daughter about the value of using statistical information for individual decisions. I took my typically extreme position, claiming it was useless. I, of course, was creamed in this discussion, not only because both of my opponents are orders of magnitude smarter than I am, but also because they made it personal. They asked if I would feel any different if my daughter were walking around at night in a statistically dangerous as opposed to a statistically safe neighborhood. Well, I'd rather she weren't walking around alone at night, period—but the fact is, I would rather she avoid the dangerous neighborhoods. We do base personal decisions on statistical data (in a decision-theoretic sort of way, as Gary pointed out). I suppose that we do so mostly when we can imagine a plausible causal relationship between what we do and the possible results. That's also why we don't stop listening to Bing Crosby when we find out that Bing listeners don't live as long as others; there is no plausible causal link that we can imagine doing anything about.

What I was looking for was a nice, clear, simple, and compelling way to justify ignoring group statistics if they really are irrelevant to individuals, and to show why and when this is the case. I think this is relatively important, because this is how medicine, social science, and most of the other life sciences work right now—they present group data as something that should be used as guidance for individual behavior. If this is a bad idea (and I feel somewhat that it is), then we should have a clear, crisp explanation of why this is so. I have been unable to clearly articulate that explanation.

I don't think it's often a problem, but I think many people actually do have serious conflicts (and control theorists should be interested in them) resulting from the fact that they are given group data suggesting that they should change their wants. In this sense, group statistics, which suggest ways to get "group-level improvements," can create individual conflicts.

Bill Powers: Rick, regarding your statistical question, if the indications are that 80 per cent of people like you are put at risk by taking X, you will only take X if you like it at least five times as much as you dislike getting cancer. But do you think that the numbers for any of these highly publicized risks are anything like 80 per cent? Consider this statement: "Among all people with clinically high cholesterol, p per cent of them die from heart attacks." Can anybody supply an actual number for p? Then consider this statement: "Among those who undergo a program designed to reduce their blood cholesterol, q per cent die of heart attacks." Again, can anybody tell us what q is?

With knowledge of p and q, you could then get a realistic picture of how worthwhile it is to try to reduce your blood cholesterol. My hunch

is that p is going to be a small number, and q is going to be only slightly smaller. The data for risks like these are never presented honestly; they're hyped up to create the most alarming numbers possible. They say, "People with high blood cholesterol are five (or whatever) times as susceptible to heart attacks as people with normal cholesterol." They don't tell you what the actual odds are, or how effective cholesterol-reduction programs are, because those numbers would be much less scarey or promising. In his book *Heart Failure,* Tom Moore pointed out that with the stroke of a pen, the Surgeon General declared 25 per cent of the population of the U.S. to have a medical condition (high cholesterol) demanding the immediate care of a physician. Drumming up business, that's what it was.

Gary, those quotations from J. G. Taylor show that, whether he intended it or not, he was helping to lay the foundations for a change to the method of modeling and the abandonment of statistics as a way of understanding human organization. Three cheers for Taylor. I'll even forgive him for citing Ashby and for overlooking invisible disturbances.

Joel Judd: Rick says: "I don't think it's often a problem, but I think many people actually do have serious conflicts (and control theorists should be interested in them) resulting from the fact that they are given group data suggesting that they should change their wants. In this sense, group statistics, which suggest ways to get 'group-level improvements,' can create individual conflicts." This strikes me as relating to cultural anthropology. Hunters and gatherers (to make a sweeping generalization) didn't have the *New England Journal of Medicine* giving them statistical data on what was safe to consume, etc. It's not simply a question of making decisions alone—we make them with regard to culture/society. We do not function in isolation. We do, however, make our own decisions. Hence the conflicts which often arise between what *we* want and what we *should* (is that a good way to put it?) want according to cultural institutions such as medicine, government, etc. Perhaps this gets back to the insidiousness of behaviorism—the propensity for those institutions that wield so much influence in our world to use behavioristic modes of thought to make decisions about what is right/ wrong, good /bad, healthy/unhealthy—whether or not they do it explicitly. And so we are faced with dilemmas in making our decisions.

Rick Marken: Thanks to those who helped with my question about taking group statistics into consideration when making individual decisions. The solution seems simple: just ask how good the group statistics actually are (i.e., do 80% of people like me show the result, and do only 10% who are not like me not show it?); then, based on those data, decide if the result of changing to be not like yourself is worth it to you. It seems that, in most cases, the group results are so weak that it really isn't worth it at the individual level.

Tom Bourbon: In the many discussions about statistics, one issue we have neglected is that of the rates of occurrence of various conditions in the general population. An analysis of this issue goes to the heart of some of the more ridiculous abuses of statistics, and of the people to whom they are applied. This is a problem that even Phil Runkel misses in his delightful and devastating book.

An elegant recent example of how far thoughts can stray when scientists ignore base rates might be pertinent to Rick Marken's defeat in the conversation with his daughter and wife about crime, criminals, and "statistically crime-infested" neighborhoods. And this case shows how even the most sophisticated experimental procedures and analyses cannot save those who ignore base rates.

The study is A. Raines, P. H. Venables, and M. Williams, "Relationships between N1, P300, and Contingent Negative Variation Recorded at Age 15 and Criminal Behavior at Age 24," *Psychophysiology 27,* 1990, 567-574. (With a title like that, you know something good is in store! "Sliced and diced," a la Runkel's analysis.) N1, P300, and contingent negative variation (CNV) are measures of brain activity—in this case, electrical activity recorded from the scalp.

The study is predicated on previously published data showing that 16.2 per cent of boys who are not criminals at age 15 become criminals by age 24. The authors report the results of their work in which they recorded brain responses (ERPs), elicited by brief stimuli, from the scalps of 15-year-olds. They administered a variety of "psychological instruments" to the boys. At age 24, they determined how many of the 101 boys were criminals. Then they looked back at the ERP data and the psychological assessments and determined which of the many possible features of the ERPs correlate significantly with anything—test scores, criminal record, one another, etc. The results convince the authors that certain "cognitive components" of the ERPs predict criminality.

For example, there is a "highly significant" correlation between amplitude of N1 at 15 and "psychopathy" at 24. (They report r = 0.73, which means p(failure) = 0.68.) Another "highly significant" (r = 0.65, p(failure) = 0.76) correlation occurs between amplitude of CNV at 15 and "psychopathy" at 24. Now those results really tell me a lot about criminality! For Rick, I guess it means you might want to set up an evoked potential system by the front door, for testing your daughter's dates!

The reason for that is that of the 101 boys, 17 became criminals by age 24. (That means 84 did not.) And a discriminant function analysis using N1 amplitude and P300 latency (why *that* particular combination?!) at 15 as "predictors" of criminality status at 24 correctly identified 75 per cent of the budding crooks! That means ERPs correctly predicted 13 of the 17 who became criminals. Impressive, isn't it? It isn't! The same "predictors" incorrectly tapped 26 per cent of the innocent boys as future felons. That means 21 boys.

The authors attend to the *percentages,* within a limited sample; by doing that, they see that the ERPs correctly identify nearly three times as many criminals as they misidentify (75 per cent vs. 26 per cent). But if you look at the *numbers* of boys, nearly twice as many innocent boys are pegged as future criminals as are guilty ones.

Oblivious to that fact, the authors go on to talk about the use of ERP data as possibly playing a role in identifying potential criminals. What if they were to succeed in that goal? Imagine a major program designed to spot the little buggers and nip them in the bud. If they tested 1,000,000 15-year-old boys, and if everything worked as they report in their research, 162,000 boys would be criminals by age 24, and the ERPs would have spotted 121,500 of them. Now *that* is war on crime! But they would have misidentified 217,880 innocent boys.

Imagine what kind of world this would be if people really *believed* the stuff that comes out of behavioral research! Wouldn't it be nice if each editor of a journal in the behavioral sciences required that authors report the results of an analysis of base rates—the actual numbers of people *in* the population—who would be correctly and incorrectly identified by the procedures described by the authors? That policy, along with a requirement that no correlations be published below r = 0.87 (the 50-50 point for being right in a prediction), would reduce the literature to about one slim volume a year. A person could read it in an evening and could have faith that at least part of the material was worth even one evening.

Bill Powers: Base rates! I knew there must be a term for it. Thanks. Tom, why don't you work up all this material for a letter to *Science?* No doubt we would be dismissed by professional statistical types as amateurish, but if you could get a letter published, at least a discussion might be started, and we would be trying to do something about these atrocities. Maybe we could at least get p(failure) accepted as a necessary part of any report on statistical data.

Statistics is an excellent tool for evaluating data and even for seeing whether there is something to a new hypothesis. You can't (apparently) get along without it in quantum mechanics. We use statistical measures even in tracking experiments. And Rick Marken has used a statistical method for identifying controlled variables in situations where the reference level for the controlled variable is continually being changed by the subject. I envision many applications for statistical analysis in the control-theory approach to behavior.

What I insist on, however, is the proper use of statistics. A statistical measure should be used only for the population from which it came. Mass measures should *never* be used to evaluate individuals if the odds of a misevaluation are significant *in terms of the payoff for the individual.* There are legitimate uses for mass measures, but the most common uses do not properly take into account the potential (and very often actual) unfairness to individuals that results from mechanical applications of statistical facts. Too often, statistics is used as an easy way to get a publishable result, with (as Tom indicated in his post) a consequence of flooding the literature with meaningless garbage (not that I'm in favor of publishing meaningful garbage, either).

Statistics is really not a tool for prediction, because all predictions imply that we want to know the value of a variable at a particular time and under particular circumstances, whereas the statistical analysis is derived from many variables evaluated at many times under variable circumstances. If we understood the underlying principles that make one variable dependent on others, we would not have to use statistics except to judge the uncertainties of measurement. More importantly, the principles that relate variables in actual behavior can hardly ever be boiled down to a simple cause-effect relationship, nor should they be. Even when we know that a person reacts with fear to dogs 80 per cent of the time, we do not know why the person reacts to any one dog with fear. Reducing that person's fear-reactions to 10 per cent might do the person a terrible disservice, if there are pit bulls and attack-trained Dobermans in the environment. Knowing the particulars is always better than knowing generalities.

And never forget that *real* statistical results seldom give us probabilities anywhere near 80 per cent.

Cross-correlation is a valid statistical method, in fact the first method I used some 15 years ago to try to detect a transport lag. I based my initial opinion about the lack of *a* transport lag on the fact that a cross-correlation measure had a peak at zero delay. But it was also true that the cross-correlation function did not show a clear peak; it was very broad, too broad to discriminate well. I think I now understand the reason. The cross-correlation method deals only with the intact closed loop of control processes, so the variables (cursor position and handle position) are not really independent. Cursor movements are dependent on handle movements, as well as on the independent disturbance. I did not find any effect of a transport lag until I put it into a working model in the forward part of the loop (the person) and by trial and error found the value that minimized the RMS error between the model's handle behavior and that of the real person. The minimum in the prediction error is still very broad, but it occurs quite reliably at the same value, trial after trial, and that value is not zero.

Control theorists are often criticized for using single-subject data. But if I had tested this model for transport lag in the usual way, proposing a one-size-fits-all model and fitting it to pooled data from many subjects, I doubt that there would have been a significant result. The model parameters differ from person to person (although the best transport lag differs less than the other main parameter, integration factor). The use of a model applied to individual data is essential here; without it, the statistical results would mean very little.

So 1 believe in the use of statistics, but only when it is properly applied and subordinated to a model. Predictions should be made from a model tailored to the particular system being observed, not from statistical measures alone (which rest on too simple a model). There is no way to avoid studying individuals if you want to understand individual behavior. I believe that current attempts to understand mass behavior are mostly ineffective. I believe that once we have a decent model for individual behavior, we will be able to synthesize predictions of mass behavior that work far better. If we see any point in doing so.

Also, Tom, from your numbers, I take it that a total of 13 + 21 boys, or 34, were predicted to become criminals. Of the 17 who became criminals, four were predicted innocent, while among those who were innocent, 21 were predicted guilty. This means that 73 per cent of the predictions of criminality were wrong, doesn't it? The "coefficient of failure" is 0.68, so it's an underestimate in this case.

You mentioned two criteria: N1 and CNV both correlated with criminality. How many subjects showed *both* N1 and CNV, and what was the criminality rate for those showing both? This is pertinent to the discussion that Gary raised (which got us into all this) about using multiple criteria for evaluating risk. My contention was that multiple criteria would do even worse than any single one.

Tom Bourbon: Bill, I am working on a letter, or a short report, on this topic. If I include a few of the many other examples from different types of journals and on a selected range of topics (to show that no major area of the behavioral-social-life sciences is clean), it might be a bit long for *Science.* Another possibility is *American Psychologist.*

And yes, the multiple criteria did have a higher likelihood of being wrong! Another thing about that multiple-variable, discriminant function analysis is that the variables entered into it are not the same ones used to report on significant single-variable correlations with "psychopathy." For the simple correlations, the authors used "amplitude of N1" vs. "psychopathy" and "amplitude of contingent negative variation" vs. "psychopathy." (By the way, the "instruments" used to "assess" "psychopathy" are yet another grisly issue!) For the discriminant function analysis, the amplitude of N1 is still in, but CNV is replaced by the latency of P300. Now, why was that done? Of course, I do not have the details, and I do not wish to impute dishonorable motives to the authors. However, brain response data offer a wealth of conceivable "measures" to enter into analyses: the amplitudes and latencies of every distinguishable "event" in the data record, the ratios of any conceivable combination of measures of "events," and so on. The list is immense. So why do any two, or more, of those measures happen to "predict" in one study, but some other combination or combinations work in another? The answer is that none of the combinations predict, except in the trivial sense of meeting a criterion of statistical significance. And the many discussions, post hoc, of why that particular combination worked in an earlier study, but this combination worked this time, lead nowhere.

Joel Judd: Tom, do I detect a note of *cynicism?* Just to keep you a little wider awake at night, the "study" you mentioned reminds me of a CIA contract the psychophysiological lab here on campus was trying to get a couple of years back when I was attending lab meetings. The "shop" was dangling fat grants to labs which could produce a sure-fire ERP lie-detector test. Fortunately, I don't believe anything ever came of it, at least not here.

Tom Bourbon: Joel, you seem to share my concerns over the misapplication of "objective" physiological measures which correlate, however pitifully but significantly, with important behavioral and psychological processes. In the late '60s, I was asked by a company in the region to look at a proposal submitted to them by a neuroscientist-psychologist. He wanted the company to put up venture capital for the manufacture and distribution of his device for measuring the latency of one "component" of human auditory evoked potentials (EPs).

He claimed, in his proposal, in several publications, and in the reports submitted to federal funding agencies, that the latency of that one component correlated significantly with various full-scale and subscale measures of "intelligence" (with n = 566 children, he had r's from -0.04 to -0.35 between latency and various IQ scales and subscales; and, as he reported, with n = 566, Pearson r's of 0.16 are significant at p less than .0001).

The scientist went on to say that his "findings" (why does that word always remind me of "leavings"?) could have "considerable educational significance," principally via the use of the EPs for "objective, culturally independent biological assessment of mental potential useful in exploring possible racial differences in intelligence." And he

went on to suggest that EPs recorded from fetuses might weigh heavily in decisions about whether a pregnancy should go to term or be aborted. All of that from correlations the best of which would lead to incorrect predictions at least 94 per cent of the time.

My report to the company was not received kindly. And the "real scientist" (who was I to question him?) took umbrage. By that time, his research was featured in various educational journals and magazines, and in offerings to school districts, which could purchase the system or the services of professionals who would administer the assessments.

This abomination vanished soon after. I like to think that my report helped it on its way. The episode marked my awakening from graduate training in which I had to virtually swear a solemn oath that the answers to psychological questions were to be found in physiological research.

The assumptions one makes about the causes of behavior and the data one accepts as supporting those assumptions are not matters of idle sport and speculation. When they work their way into decisions about policies that affect the lives of innocent people, the scientists who offer them ought to be held strictly accountable and responsible. All the more reason for us to insist on models that work at least in simple instances of behavior and on data that predict what actually happens, at least half of the time!

Gary Cziko: Tom has been providing some fascinating accounts of the misuse of statistics in predicting individuals. But I am having some difficulty understanding the way he is conveying information about correlation coefficients.

For example, he says: "... there is a 'highly significant' correlation between amplitude of N1 at 15 and 'psychopathy' at 24. (They report r = 0.73, which means p(failure) = 0.68.) Another 'highly significant' (r = 0.65, p(failure) = 0.76) correlation occurs between amplitude of CNV at 15 and 'psychopathy' at 24." And he also says: "All of that from correlations the best of which would lead to incorrect predictions at least 94 per cent of the time."

It seems in the first quote that Tom is saying a correlation of r = 0.73 gives a p(failure) of failure of 0.68. I don't think this is quite the way to put it, since, to me at least, p normally indicates a probability, which this isn't.

If we take 0.73, square it, subtract the squared value from one, and then take the square root of the difference, we will indeed have a value of 0.68, which I have seen referred to in at least one statistics text as k, the coefficient of alienation. That is, $k = \text{sqrt}(1 - r^2)$. But k is no probability, it is rather the ratio of the standard error of estimate of using one variable to predict the other to the standard deviation of the criterion variable. So if 0.73 is the correlation between years of education and income, using education to predict income will give us 68 per cent (about two-thirds) of the error (difference between predicted and actual income) that we would get if we knew nothing about anyone's education and just used *the* mean income of the group to predict each individual's income. Or, subtracting 0.68 from one, we find that the correlation of 0.73 gives a 32 per cent improvement in predicting Y based on X over not knowing anything at all about X.

So it seems to me that the p(failure) notation is misleading if Tom is using p for probability. In fact, the probability of predicting someone's score exactly right on a continuous variable measured with infinite precision is actually zero (which is why statisticians don't like point estimates and use interval estimates instead).

Also note that correlations start to look better when you are trying to simply predict whether someone will be higher or lower than some predetermined criterion. If I simply want to know whether someone has an above average or below average IQ based on some predictor (e.g., some brain-wave measure), then the probability of correct predictions rises dramatically (I can give some tables if this is of interest). But then the question arises as to what average IQ is, how it is determined, and how just being above or below average correlates with some other variable of real interest (such as whether someone finishes high school or not). So I doubt that the predictive value is really much better even in this dichotomous case. (It might be better if the criterion variable were something clear-cut like sex, but there are probably easier ways to predict sex than by using brainwaves.)

Maybe the best way to talk about this new index we like so much is to subtract it from one, multiply the difference by 100, i.e., 100 * (1 - k), and call it something like "per cent improvement" (PCI). So in the above case of *r* = 0.73, PCI = 32 per cent, meaning that errors of prediction using the predictor variable are on average 32 per cent better (i.e., less) than just using the mean of the group to predict each individual's score in the group.

This is what Tom's interesting statement would look like using PCI: There is a "highly significant" correlation between amplitude of N1 at 15 and "psychopathy" at 24. (They report r = 0.73, which means PCI = 32 per cent.) Another "highly significant" (r = 0.65, PCI = 24 per cent) correlation occurs between amplitude of CNV at 15 and "psychopathy" at 24.

Hmm. After looking at this, I think I prefer the "uselessness" approach after all. Just like above, but don't subtract from one. That gives the "per cent uselessness" (PU; it even sounds right). Now the statement looks like this: There is a "highly significant" correlation between amplitude of N1 at 15 and "psychopathy' at 24. (They report r = 0.73, PU = 68 per cent.) Another "highly significant" (r = 0.65, PU = 76 per cent) correlation occurs between amplitude of CNV at 15 and "psychopath)," at 24.

Yes, I like PU much better, since most of the correlations we find in social sciences research really do stink. Suggestions welcome. Vote for PCI or PU.

Tom Bourbon: Gary properly chastised me for saying that k might represent the probability of failure in predicting Y from X, given a correlation r. My initial interpretations of Bill Powers' remarks on k were to blame—the fault is mine, not Bill's.

My utter lack of familiarity with this index puzzled me: the coefficient comes directly from the calculations for Pearson's r, so why is it not discussed in statistics books with which I am familiar?

I did find one fleeting paragraph in a text from my student days, but it is in a section marked "not assigned." I just located a rather thorough discussion in a text from 1956 (before my university days): J.P. Guilford, *Fundamental Statistics in Education and Psychology,* McGraw-Hill, New York. On pages 375-379, he discusses "the correlation coefficient and accuracy of prediction." Guilford characterizes the relationship between r and k as follows: "Whereas r indicates the strength of relationship, ... k indicates the degree of *lack* of relationship.... If r is 0.50, k is not also 0.50 but 0.886. Where r is 050, then, the degree of relationship is less than the degree of lack of relationship. It is when r = 0.7071 that the relationship and *lack* of relationship are equal."

And, as Gary suggests, multiply k by 100 and: "Our margin of error in predicting *Y with* knowledge of X scores is (k * 100) per cent as great as the margin of error we should make *without* knowledge of X scores."

Guilford goes on to describe 100 * (1-k) as the "percentage reduction in error of prediction," also known (then) as the "index of forecasting efficiency, E." I wonder why all of this dropped out of the statistics texts?

I vote for PU, of course!

Chuck Tucker: I think that the comments on statistics on the net are clear, concise, well documented, and will disturb the social and behavioral scientists (sic) to no end. These comments question the "articles of faith" that support the social sciences. They should be published in some form, if nothing more than being sent in outline form to every electronic network in the country with members who are social scientists. I only have a few comments by way of refinement.

(1) We should not make the error that everyone else makes when using the word "group." A group is a set of people who at least interact

with each other. My criticism of sociologists is that they define their discipline as the "study of groups," but they only study individual characteristics—not the individual as a person, or even a personality. So the statistics we are talking about are numbers generated (how?) from individual characteristics and put in categories or other aggregate forms through various means of classification—we don't have group statistics. The closest we come to group statistics (which sociologists have completely ignored in their work) is to be found on the sports page of the newspaper and, to some extent, the business section. Most of the statistics that we are told about and find in our journals are *not from* groups.

(2) We should note very clearly that these statistical presentations have serious effects: many people, especially government officials, "control for" such numbers. There is very good evidence (yes, numbers) that most journal editors (Clark McPhail has a series of papers on this issue) and readers will not consider a paper suitable for publication without statistics. We have developed a nation of quantofanatics!

(3) I wonder if those who are critical of the use of aggregate statistical analyses being applied to individuals and also believe that extreme competition leads to many of the problems we have among people have abandoned the use of "curving" or distributions for deciding what grades students receive. I believe that one of the most serious problems of our public education system is the use of "curves" to determine a student's grade, rather than the use of a standard set by the instructor/teacher and understood by the student. When the "standard" is merely doing better in a statistical distribution than some others, students only have a minimal notion of what is "excellent work." When we have raised a generation of parents and teachers who have experienced such procedures and continue to pass them on, then we should expect a continual lowering of the statistical standards (by the way, this would be an excellent experiment to be done by those in education—does it lower standards?). The point: to be consistent with control theory, a teacher should set a standard, encourage students to use that standard, and judge students' performance by the standard set, without regard to any statistical distribution of an aggregate (a college class is not a group, either!). When this is done, all can get high, medium, or low grades. Students can study together; there is less conflict among them and between students and teachers (although I do get complaints when they don't get high grades—but I am the only teacher at my institution who approaches grading in this way, and a less-than-high grade is a disturbance).

Martin Taylor: I am just starting to read Bill Powers' 1973 book for the first time, and in talking about time-scales of response (page 54), I

come across the following quote: "Psychologists who believe that intermittent reinforcement is more effective than continuous reinforcement should give this whole speed-of-reaction problem serious thought—for a long enough time." I realize that this was written a long time ago, and might have been amended later in the book, but it does resonate with some of the threads that have been weaving through the net—statistics, in particular. So although it might be unfair, I will comment.

Intermittent reinforcement is not usually seen as "more effective," but as more resistant to extinction. And a statistical reason is not hard to find. In the laboratory, the animal is confronted with a situation in which it is sometimes rewarded for behavior A, but never for behavior B (or less often, perhaps). Now, if the experimenter decides to stop rewarding behavior A, how can the animal know that the world has changed its rules? Previously, failure of reward for A has been followed by further reward on a later occasion. It cannot know that this will no longer be true. Only by implicitly evaluating the statistics of the reinforcing event can it determine after a while that a long period of non-reinforcement would have been unlikely under the regime to which it had become accustomed. If you like, there is a "continuous" higher-order event—a statistical event—which occurs on a time-scale much longer than that of the single reinforcement.

In such an experiment, the experimenter tries to make sure that the animal has no access to information that might let it know which rule is in effect. Many experiments have been found to give results that depend on the animal hearing a click or something that the experimenter had not noticed, but that occurred only when reinforcement was going to be provided. The animal then has a context that turns the statistical event into a predictable event. It can know that the world has changed if it no longer hears the click.

It should be much easier to learn a behavior that has a perfectly predictable consequence, but normally we do not have access to all factors that influence the consequences of our behavior, and so we have to resort to statistics to determine how our behavior is influencing our perception. The control system can be fully determined in its behavior, but if we cannot tell the difference between a context in which behavior A leads to result P and one in which it leads to result Q, then all we can do is to take advantage of the best information we have; that is, for example, that A then P has happened 75 per cent of the time we did A, and A then Q has happened 25 per cent of the time. If we want P to happen, and it is not too bad if Q happens instead, then we would do A. But if Q would on this occasion be disastrous, we might try another way of getting P to happen rather than risking behavior A.

Life, even in a control-system view, is a statistical game.

Sorry if that's all too obvious to have been mentioned, but I have read so much trashing of statistics on the net that it seemed rather to be so obvious as to have been overlooked.

Bill Powers: Martin, you say: “Intermittent reinforcement is not usually seen as ‘more effective,’ but as more resistant to extinction. And a statistical reason is not hard to find.” I agree in both regards. I was thinking in terms of “habit strength” and Skinner’s “shaping” experiments when I said “more effective.” Both are related to extinction. (Skinner found that by changing the schedule so as to deliver fewer reinforcements for the same behavior, he could *increase* the rate of responding. He cited this as an instance of the power of intermittent reinforcement, never realizing that this relationship is the opposite of the one he always assumed to hold between reinforcement rate and behavior rate.)

As to the statistical reason, there are many cases in which a statistical analysis comes out with the same results as a modeling analysis without statistics. Suppose that an animal has learned to perceive the rate at which some almost-rhythmic stimulus appears. Representation of this rate as a neural signal (by analogue means) would require a smoothed frequency detector. The smoothing is required to eliminate the individual instances of an input and produce a signal whose magnitude is proportional to the rate of appearance. The amount of smoothing used determines the range of input frequencies over which the signal magnitude usefully indicates input frequency (too long a smoothing time yields a maximum signal for all rates above a certain limit). Within the range of operation, the signal magnitude corresponds roughly to the probability that an input will occur within a given time interval, related to the smoothing time. So the analogue perceptual function can accomplish the same end as a probability calculation, but in a quite simple way. If we were choosing on the basis of simplicity of circuitry, I would pick the analogue method. Of course, we must ultimately pick the method that the nervous system actually uses.

Given the smoothing time, it will take a certain number of input events to bring the perceptual signal to a constant level, and this will determine about how fast the related control system can act. When the input events stop occurring, the perceptual signal will take the same length of time to decay, so the system will go on attempting to control the signal after the input events have actually stopped (the extinction curve). This is in fact how it works: if learning takes a long time, so does extinction, at least in certain learning experiments.

I believe that this analogue model gives about the same results as a statistical-perception model does. The analogue model works with inputs that have an average frequency with random variations. It does *not* work properly (and neither does the statistical model) when the input frequency is perfectly regular. We notice the first tick of the clock that is missing or comes too soon or too late. So that sort of situation requires not an average rate detector, but a synchronized detector (I think I would put it at my “event” level of perception, whereas the other kind of rate detection would go one level lower, at the “transition” level).

Generally, I think that your analysis of intermittent reinforcement is correct. I’m only proposing an analogue method that does, in effect, the same computations but without requiring statistical calculations.

I’m not against statistics in general, or even against statistical explanations of neural functioning (at the appropriate level). When we consider noise in control systems, statistical methods help us appreciate its effects. What I “bash” with enthusiasm is the misapplication of statistical facts to individual occurrences. I’ve tried to make my criticisms specific to that case. That would seem to be a subject different from the one you are talking about.

I don’t think we often get into situations where the environment is ambiguous or unpredictable. When you look around, you see a pretty noise-free visual field, with clear demarcations between objects, colors, sensations, relationships, and so on. When uncertainties do arise, we might sometimes use statistical methods to deal with them, by which I mean literally computing chances, but I think in many cases we simply smooth out our perceptions and operate on the basis of the artificially unambiguous result—often wrongly. Anyway, people don’t seem to compute their behavior on very good statistical grounds, do they?

Just for fun, a poem by Maurice G. Kendall, originally published in *American Statistician* 13(5),1959, 23-24:

Hiawatha Designs an Experiment

1. Hiawatha, mighty hunter
He could shoot ten arrows upwards
Shoot them with such strength and swiftness
That the last had left the bowstring
Ere the first to earth descended.
This was commonly regarded
As a feat of skill and cunning.

2. One or two sarcastic spirits
Pointed out to him, however,
That it might be much more useful
If he sometimes hit the target.
Why not shoot a little straighter
And employ a smaller sample?

3. Hiawatha, who at college
Majored in applied statistics
Consequently felt entitled
To instruct his fellow men on
Any subject whatsoever,
Waxed exceedingly indignant
Talked about the law of error,
Talked about truncated normals,
Talked of loss of information,
Talked about his lack of bias
Pointed out that in the long run
Independent observations
Even though they missed the target
Had an average point of impact
Very near the spot he aimed at
(With the possible exception
Of a set of measure zero).

4. This, they said, was rather doubtful.
Anyway, it didn't matter
What resulted in the long run;
Either he must hit the target
Much more often than at present
Or himself would have to pay for
All the arrows that he wasted.

5. Hiawatha, in a temper
Quoted parts of R. A. Fisher
Quoted Yates and quoted Finney
Quoted yards of Oscar Kempthorne
Quoted reams of Cox and Cochran
Quoted Anderson and Bancroft
Practically in extenso
Trying to impress upon them
That what actually mattered
Was to estimate the error.

6. One or two of them admitted
Such a thing might have its uses
Still, they said, he might do better
If he shot a little straighter.

7. Hiawatha, to convince them
Organized a shooting contest

Laid out in the proper manner
Of designs experimental
Recommended in the textbooks
(Mainly used for tasting tea, but
Sometimes used in other cases)
Randomized his shooting order
In factorial arrangements
Used in the theory of Galois
Fields of ideal polynomials
Got a nicely balanced layout
And successfully confounded
Second-order interactions.

8. All the other tribal marksmen
Ignorant, benighted creatures,
Of experimental set-ups
Spent their time of preparation
Putting in a lot of practice
Merely shooting at a target.

9. Thus it happened in the contest
That their scores were most impressive
With one solitary exception
This (I hate to have to say it)
Was the score of Hiawatha,
Who, as usual, shot his arrows
Shot them with great strength and swiftness
Managing to be unbiased
Not, however, with his salvo
Managing to hit the target.

10. There, they said to Hiawatha,
This is what we all expected.

11. Hiawatha, nothing daunted,
Called for pen and called for paper
Did analyses of variance
Finally produced the figures
Showing beyond peradventure
Everybody else was biased
And the variance components
Did not differ from each other
Or from Hiawatha's
(This last point, one should acknowledge

Might have been much more convincing
If he hadn't been compelled to
Estimate his own component
From experimental plots in
Which the values all were missing.
Still, they didn't understand it
So they couldn't raise objections
This is what so often happens
With analyses of variance).

12. All the same, his fellow tribesmen
Ignorant, benighted heathens,
Took away his bow and arrows,
Said that though my Hiawatha
Was a brilliant statistician
He was useless as a bowman,
As for variance components
Several of the more outspoken
Made primeval observations
Hurtful of the finer feelings
Even of a statistician.

13. In a corner of the forest
Dwells alone my Hiawatha
Permanently cogitating
On the normal law of error
Wondering in idle moments
Whether an increased precision
Might perhaps be rather better
Even at the risk of bias
If thereby one, now and then, could
Register upon the target.

Tom Bourbon: Several of my colleagues are somewhat tolerant of me and of students who turn on to PCT, but others are not so open or supportive. The person who asked seniors in a statistics course to present a talk on some controversial topic concerning uses of statistics in psychology was not prepared to have one student give a reasoned discussion of the "coefficient of failure," as discussed on the net. Nor was he ready for another student who, by all accounts, gave an elegant review of Phil Runkel's critique of abuses of the method of relative frequencies.

My students are told by some people that they don't care what kind of evidence he (I) might present, PCT isn't right, and it isn't psychology (I believe that!). During a discussion with several students who invoked PCT as part of a challenge to his pet theories, a faculty member blurted out, "What does he [II do to you people, brainwash you?"

Bruce Nevin: I think there is a confusion of statistical prediction with prediction for an individual control system.

One can predict that most middle-class children will get an education; one cannot predict that a particular one will, unless that one is controlling for getting an education (for whatever reason). Likewise for learning their native language (exceptions may be autistic, severely retarded, kept locked in a closet, etc.). One could predict that Bill would marry an intelligent person because that was what he was controlling for (among other things), but not because most engineers in the field of astronomy who are former psychology students marry intelligent people.

Joel Judd: Bruce, isn't this why, in a certain sense, prediction becomes trivial in PCT? The trick is to find out what someone is controlling for. Also of interest is what the person does to reduce error. This might also be why, historically, so many psychological and educational researchers haven't told us much about process and mechanism, so concerned are they with predicting the right damn outcome. Why the outcome occurs and *how* it occurs must be explained by that black box up there.

Bill Powers: Perceptual control theory is fundamentally a theory of individual organization. You get to statistical predictions for populations in a different way. First you study enough individuals to find how their control parameters are distributed. Knowing that, you can predict how a population of "similar" (oops) individuals will do the same sort of control task. You will also know better than to speak of the "average way of controlling in this task." Nobody controls that way.

If you have ways of measuring individuals' control parameters, wouldn't it usually be unnecessary to go through the population-study route? When you study populations, you get characteristics of the population, but you don't learn anything about an individual, except perhaps the outer limits of variation within which this person might be found—unless the person happens to be from a different population and your criteria for population membership just didn't happen to pick that up.

One point of using control theory is to get away from statistical studies in which experimenters are jubilant (typically) over correlations as low as 0.8. Facts that are determined statistically are true only of a population and are next to useless for predicting the performance of an individual. There is a tendency to elevate findings that are true only of a majority of a population (say, 60 per cent of subjects) so that they are assumed true of the whole population.

There are two ways to understand natural phenomena. One is like trying to figure out a system for winning at roulette. You observe and observe, and finally you get an idea: every time two blacks and a red show up in that sequence, an odd number between 11 and 27 will win, but if the sequence is black, odd, black, red, the best bet is a number ending in 5. This is "looking for rules." It is also the basis for statistics, because when you're testing a rule like that, you have to keep track of how often it worked. If it doesn't work often enough to be useful (i.e., to keep you from going broke), you go back to searching for more rules.

The problem, of course, is that even if a rule appears to work, you have to consider how many chances you had to find it, how many times it might have failed before you noticed it, and how often it will fail in the future. Even if the rule appears to work in all your tests, it might still have nothing to do with anything. Even if the rule works 20 times in a row, there is always the chance that it is irrelevant or will become irrelevant without advance warning.

In fact, all you need is one exception to show that the rule is irrelevant. If you can have one exception, then you can have two in a row, 10 in a row, 100 in a row, and go broke.

Of course, there's always the chance that the rule you found actually has some explanation; it might be a reflection of a real regularity in nature, so that the rule really has to work (even though you don't happen to know why) or sometimes has to fail (depending on occasional underlying circumstances you haven't discovered). This, of course, is what we hope for when we try to guess at the rules. This is the mode of research that I call "trying to get lucky." Getting lucky means stumbling across one rule among all the others that is an expression of an underlying mechanism.

If you get into the gambling hall after hours, you can look under the roulette table. When you see a little button where the croupier stands, you can immediately deduce a rule for betting that has some reason for working: bet (small) against the biggest betters. The game is rigged.

So this leads to the other way of understanding nature: look for the way in which the game is rigged. Don't waste too much time trying to guess at the rules just by watching phenomena. The only rules that actually work are those that work for an underlying reason. All the rest are illusions. If you just look for rules, you can't tell the illusory rules from the real ones. And the real rules don't work just because they work: they work because they have to work. The game is rigged that way. The system is organized that way.

Modeling is an attempt to see under the roulette table.

Rick Marken: The only time I have encountered anything approaching hostility to control theory is when the listener figures out that control theory is completely inconsistent with the whole experimental/statistical framework on which psychology is based. Most psychologists really believe in this model. They spend years learning statistics and experimental design. It is the core of the discipline: the basic foundation on which the search for psychological truth has been built. Control theory says: forget it. When you say that to the people who wrote the texts, taught the courses, labored in the statistics classes, and paid their dues running hundreds of subjects in complex factorial experiments, you don't get big cheers. Even if you carefully show why conventional statistics/experimental design seems to work but really reveals little if anything about the internal organization of living systems.

So my experience is that control theory has the biggest problems when it comes face to face with faith in the *scientific method* as articulated in the pages of the exalted textbooks of statistics and methodology that are the bedrock of *all* (cognitive, behavioral, ecological, etc.) psychological science.

Bill Powers: [Replying to a researcher in cognitive science.] Experimentation under the control-system model is aimed at the characterization of individual behavior. The only reason for using multiple subjects in a single experiment, other than checking for flukes, is to see how variable the individual measures are over a population. We would never average such measures together! Question: What is the average damping coefficient of the human arm control system? Answer: That's not a meaningful question, because the damping coefficient must be appropriate to the build and organization of each control system, if it's stable. Details of organization vary greatly from one person to another.

I don't think that statistical studies can hack it in the long run. They have their uses, but once you've seen how control-theoretic experiments go, you'll be spoiled for statistical work. I say that with fingers crossed, because actually nobody is doing systematic research on PCT at the cognitive levels where you work—this is by way of inviting you to learn the basic principles of PCT and be a pioneer. Doing so will earn you the distrust of your colleagues, difficulties in publishing, and experiments with clear-cut results that you know are right. And friends like us who give you a hard time. You have to weigh the costs and benefits yourself.

The question we always ask people who report statistical results is "How many subjects *didn't* show the effect, and how does your hypothesis explain *their* behavior?" I claim that if you have to use multivariate analysis to show that there is an effect, you haven't got an effect. Real effects stand out like sore thumbs. They aren't the results of causes, but of organization.

My biggest objection to most statistical analyses (I don't know about your analyses) is that almost uniformly they employ a cause-effect model of behavior. We can *prove* that's the wrong model. Organisms produce consistent outcomes by variable means. It's easy to demonstrate this principle in almost any context, at any level. Most experimenters carefully avoid disturbances that might interfere with output, not realizing that the same outcome would happen anyway. Of course, if they did introduce disturbances, and the outcome did repeat, this would completely screw up their experimental paradigms. Maybe that's why they don't do it.

[In reply to a post from Eileen Prince asking what PCT has to say about autism.] Control theory isn't like most other theories: it doesn't say that if X happens to people, Y will be the resulting effect on their behavior. It's about the way behavior works; it describes relationships of a very general nature between perception and action. At the same time, it is a theory of individual behavior: in order to apply it to an individual, one must determine what variables that individual is controlling, and with respect to what internally specified states, and the quality of that control. The hierarchical model suggests a nested stack of types of controlled variables that people seem to be able to control when all is well—but the particular examples of these types that an individual controls can be discovered only by studying that individual.

Control theory doesn't use categories such as "autism" to explain behavior. To say that a person is autistic is only to say that certain externally visible patterns of action have struck people as similar enough (and unusual enough) to be lumped into a "disease entity." This does not mean that the same defect exists in all autistic people, or that the symptoms arose from some common history, or that the same treatment will succeed with (and not harm) everyone included in this category. The conventional empirical approach to treating problems as "diseases" is simply to try something on people in a given category and see if it helps a statistically significant number of them. There is no attempt to analyze what has actually gone wrong—what the person can still do normally, and what the person can't do. There is no attempt to relate deficits to a model of internal functioning. I suppose the idea is that if you accumulate enough experience with treating people in arbitrary categories, you will eventually be able to look up the symptoms in a big book and read off the treatment that has been effective most often in the past. In my view, this approach is an ill-advised attempt to bypass understanding of the human system and find solutions by relying on guesswork and luck. Before the advent of science, it was all we had. Sometimes it works. But there has to be a better way.

Rick Marken: Here's my hypothesis about what variable conventional psychologists (of virtually all stripes) are trying to control: the perception that they are able to have relatively (statistically) predictable effects on what other organisms do. Not surprisingly, the behavior of other organisms is, from the point of view of a psychologist, a controlled (or potentially controllable) variable. This holds even in cognitive psychology, I think. I used to do some research on visual search. Nearly all of this work is aimed at trying to find factors that affect the rate of search -such as similarity of target to background, statistical properties of the background, and so on. You can find things that have pretty strong *statistical* effects on search rate. So you can control search rate (or at least the average rate) by messing around with the background. To the extent that you get the effects you want in your study (effects that match your reference) then you are happy. The experimenter is typically more concerned with his own ability to control what happens than in the organism's ability to do so.

Bruce Nevin: As regards control of perceptions relative to internal reference values, statistical measures are of little use. As regards the processes by which people set internal reference values of the "social convention" sort, measures are in order that correspond to the way individuals generalize across the outputs of other members of their population. This way of formulating the problem might suggest more apt ways of formulating statistical analysis, ways that can be modeled in control-theory terms.

Chuck Tucker: Here is a recent version of the statistics so frequently used in social science.

*Relationships Among Several Descriptive Statistics**

r	r2	k2	k	E (%)
1.00	1.00	0.00	0.00	100
0.9995	0.999	0.001	0.032	97
0.9987	0.997	0.003	0.054	95
0.995	0.99	0.01	0.099	90
0.954	0.91	0.09	0.299	70
0.90	0.81	0.19	0.435	56
0.87	0.756	0.244	0.493	51
0.865	0.748	0.252	0.50	50
0.80	0.64	0.36	0.60	40
0.71	0.50	0.50	0.70	30
0.60	0.36	0.64	0.80	20
0.50	0.25	0.75	0.87	13

r	r2	k2	k	E (%)
0.40	0.16	0.84	0.92	8
0.31	0.10	0.90	0.95	5
0.20	0.04	0.96	0.98	2
0.10	0.01	0.99	0.995	0
0.00	0.00	1.00	1.00	0

*Compiled by Chuck Tucker, with the encouragement and assistance of members of the CSG (especially Gary Cziko) and Jimy Sanders.

Definitions and Interpretations of the Above Statistics

All of these measures describe two variables (X and Y) within a particular sample. It should be stressed that these descriptions and interpretations, especially those involving "predictions," are limited to a particular sample; if another sample is not a random sample from the same population, then predictions about Y will be unpredictably worse.

r is a correlation (or coefficient of correlation) which describes the linear association of one variable with another. It can also be characterized as "... a relative measure of the degree of association between two series..." of values for two variables. It varies between 1 (perfect positive correlation) and -1 (perfect negative correlation). The closer this measure is to a perfect correlation, the more confidence one has in "predicting" the values of one variable from another variable.

r2 is a measure of "explained" variance (or coefficient of determination) which describes "shared" variation, or the amount of variance of one variable "explained" by the other variable, or the proportion of the sum of y2 that is dependent on the regression of Y on X. The larger the numerical value of this measure, the more confidence one has in "predicting" the values of one variable from another.

k2 is a measure of "unexplained" variance (or coefficient of nondetermination) which describes "unshared" variation, or the amount of variance of one variable *not* "explained" by the other variable, or the proportion of the sum of y2 that is independent of the regression of Y on X. The smaller the numerical value of this measure, the more confidence one has in "predicting" the values of one variable from another.

k is a measure (called coefficient of alienation) which describes the lack of linear association of one variable with another, or the ratio of the standard error of the estimate to the standard deviation of the variable. The smaller the numerical value of this measure, the more confidence one has in "predicting" the values of one variable from another.

E is computed as 100 * (1 - k) and is called the "index of forecasting efficiency" (Downie and Heath, 1965, p. 226). It indicates the "improvement" for a prediction by knowing the coefficient of correlation (r) for two variables, as contrasted with knowing nothing about the linear association of the two variables. For example, with a coefficient of correlation of 0.71, one can "predict" the values of one variable from another about 30 per cent better (on average) than one could "predict" those values *without* any knowledge of the relationship between the two variables; *or* one has decreased the size of the "error of prediction" by 30 per cent (on average) by knowing that the correlation of the two variables is 0.71.

References

Herbert Arkin and Raymond R. Colton, *Statistical Methods,* 4th edition, College Outline Series, 1956.

N. M. Downie and R. W. Heath, *Basic Statistical Methods,* 2nd edition, Harper and Row, New York, 1965.

Bill Powers: In the social sciences, the word "theory" is used to describe a proposed statement of relationship: people who have characteristic X exhibit a tendency toward behavior Y. I would call this a proposed fact: either X's show Y or they don't. If they do, we now have an observed relationship (never mind how reliable it is) that demands theoretical treatment. The corresponding theoretical statement would tack on "because..." to the observation, and propose a mechanism that accounts for the observed dependency.

Another way in which description is confused with explanation is through the manipulation of categories. A specific instance of behavior by a specific person (Joe opens a door and walks out of the room) is converted to an instance of a class of behaviors of a class of persons (a male college sophomore exits from an enclosed space). The specific antecedent conditions are also converted to a category: "the room contains 400 people" converts to "the population density in the enclosed space is more than two persons per square yard." Now the happening becomes: "A white male sophomore exits from an enclosed space when the population density exceeds two persons per square yard." This now looks like a more general statement that will apply to more people than just Joe, and more instances of crowding in larger and smaller rooms. In many branches of the social sciences, this is considered to be an explanation.

Of course, the statistical approach and the generalization approach are used together.

The theoretician has to take the point of view of the behaving system. When you imagine being a particular control system, you realize that the actual environment is almost irrelevant: all you can know about it is contained in the perceptual signal, and the relationship of the perceptual signal to external processes and entities depends entirely on how the input function is organized. So the control system can control only its perception; the effects it has on the external world while doing so are unknown to it.

The key is not so much being able to prove that the model is right, but simply understanding how to propose processes in such a way that they *could* be right. This amounts to appreciating what sort of thing has to be accomplished by the system in order for its externally observed behavior to be as it is. We might not know how to build a general configuration-perceiver, but at least we know that the input has to be a set of sensations, and the output has to be a signal that covaries with our own sense of configuration. If *we* can think of one mechanism capable of doing this in one instance, that is better than not knowing of any mechanism. And when we have one mechanism that works, we can try to find another one that works better, seeing how the first one fails. And so it goes until we have a good model.

But we can never know that we have accomplished something in the same way that an organism accomplishes it, in every detail. For that matter, we have no reason to think that every organism of a given species accomplishes its functions in the same way as other organisms of the same species. Judging from the very large differences in brain anatomy that exist from one person to another, in fact, it's unlikely that all people are internally organized in the same way even if they behave in roughly the same way. The brain is plastic, and its organization is influenced by the experiences of a single lifetime. Our modeling is fundamentally limited by this fact: no one model can ever reproduce to the last detail the inner functioning of all examples of any kind of higher organism, because the originals are not all designed in exactly the same way. We will always be limited to modeling the "general idea" behind an organism, because that is the limit of consistency in the originals. The method of modeling is primarily a method of understanding individuals, and only secondarily a way of saying general things about all individuals. Models must always contain parameters that can be adjusted to fit the "general idea" to a specific organism.

This, naturally, has some serious implications concerning the nature of scientific research into human nature. It's usually assumed that one is dealing with a standard instance of *Homo sapiens*—the very idea of assigning such a term to the whole human race is to assert that fundamentally we are all the same. In the psychology lab, great attention has been paid to using a standard animal model—the Sprague-Dawley rat, during my formative years. If you have a standard rat or a standard person, you should get standard responses to standard stimuli. If any human being is as good an example of *Homo sapiens* as any other, you can study groups of people as interchangeable units, drawing generalizations from the data which you assume to be measures of common underlying properties fuzzed out by uncontrolled stimuli.

But what if, below some level of observation, there are no common underlying properties? Then the whole rationale of statistical studies of populations collapses.

A generative model is one that will reproduce the phenomenon of interest by operating strictly from the interplay of its own properties. A generative model of control behavior is a control system with an input function, a comparator, and an output function, in an environment that links output to input in a specific way. There is no component in a control-system model that "controls." Control is the result of operation of a system with these functions in it, connected as specified by the control-system model, and operating as dictated by the input-output properties of each component.

So, given inputs, constraints, and parameters, a generative model must always produce some kind of behavior. We can't necessarily anticipate what such a model will do, but whatever it does is rigidly set by the properties we have given it, and by the surroundings with which it interacts. We hope that the behavior of the model will resemble the phenomenon we're trying to explain. If it doesn't (and few models do, the first time they are set in motion), we have to modify the model. That's how models grow and improve.

Greg Williams: Many physicists make a living *describing* certain phenomena, just as many psychologists are experimentalists. But modern theoretical physicists eschew the "hypotheses non fingo" stuff. And make *extrapolative, explanatory* models. Unfortunately, the bulk (well, there really aren't all that many) of theoretical psychologists still persist in making *descriptive, nonexplanatory* models solely at the level of the phenomena—rather than *generative* models of *underlying* mechanisms. "If you do basically the same procedures again, the organism will do basically the same thing." The weasel word is "basically," because these folks cannot circumscribe its bounds. So, the turn toward statistics.

I claim that the only reasonable answer to Hume's inductive skepticism (i.e., why should the sun rise tomorrow?) is making generative models which "hang together." Hypotheses non fingo leaves open the possibility that matter might disappear at any moment, since it can't predict that it *will* disappear at a *particular* moment. Contemporary generative modeling in physics says there's no "disappearing at such-and-such-a-time" relation within its (modeled) structure, so give us a break from your concocted philosophical "possibility" tales, Hume!

Descriptions at the behavioral level don't explain behavior, and descriptions at the sociological level don't explain sociological observations. A description in the Skinnerian vein would be that people show certain behaviors which are correlated with certain outcomes which can be lumped into a class termed "rewards" (of course, this begs the question of why some outcomes end up in the class and others don't, which can only be answered by invoking structural constraints embodied in organismic physiology). However, such description at the individual behavioral level is, I claim, what counts as an explanation of sociological phenomena. It appears to me that people generally accept accounts such as the following as explanation: How come the voting turnout rates of the poor are much lower than those of the rich? Continuing in the Skinnerian vein, for argument's sake, it is because some individuals receive few rewards from voting (and reduce their voting rates), while other individuals receive many rewards from voting (and keep voting). That is pure description at the individual behavioral level. But it isn't an explanation of the sociological phenomenon, just yet.

What must be added to the description at the individual behavioral level, as given above, is description at the sociological level, to wit: most individuals belonging to the class "poor" *actually are* in the first (few rewards from voting) group described above, and most individuals belonging to the class "rich" are in the second (many rewards from voting) group. Now we can *deduce* that the poor will come to vote less frequently than the rich. We have a generative model at the individual behavioral level which, coupled with a description of certain conditions observed at the sociological level *(but not the phenomenon to be explained at that level),* results in an explanation of the sociological phenomenon in question.

In this example, pure faith (precisely as criticized by Hume!) is the *only* basis for believing that *tomorrow* the poor will continue to vote less frequently than the rich, since there is no basis except a belief that "what was, will be" for extending the "functional relationship" (Skinner's term) from past correlations between voting and reward to future frequencies of voting. Without limits on the generalizability of such relationships, which I claim can only be placed by generative models at the level below individual behavior, you're in free fall. One might call it the free fall of statistics—comfortable, until you meet a boundary. Then, *splat!*

Bill Powers: I think that Popper's idea of "falsification" is predicated on the prevailing view of theories as being primarily statistical. Statistical theories don't propose any models, so there is no positive way to verify a theoretical statement. All that significance does for us is to assure us that the experimental results probably didn't happen by chance. There is no a priori or logical argument against the result being a chance occurrence; it is reasonable to admit the possibility that chance played a part. This negative conclusion doesn't tell us that the hypothesis is reasonable, connected to a systematic world, or useful in any context other than the original experimental conditions.

Models, on the other hand, are tested by changing the conditions and verifying that the model still behaves as the real system does under the new conditions. The model provides an a priori systematic reason for the system to behave in some new way under new conditions, and commits us to specifying exactly what that new way will be. When the real system does behave that way, this is a positive indication of the model's worth. Of course, one could argue that there is still a possibility that the real system behaved in the new way by chance, but if the standards for acceptance are set as high as they are in the physical sciences, this possibility goes beyond the bounds of reason: there's a qualitative difference between p less than 0.05 and p less than 0.0000000005. More likely is the possibility that the real system behaved in the new way for a reason other than the reason for which the model behaved that way. This does not involve chance; it says merely that the model needs to be modified, and that sooner or later circumstances will reveal the needed change. The modeling approach is fundamentally systematic, not statistical. Modelers assume that the underlying processes, whether we have correctly identified them or not, are systematic.

Thus, I would say that I use the criterion of *testability,* not falsifiability. Falsifiability is a subset of testability that considers only the possibility of rejection. Testability also demands that hypotheses that are not rejected be accompanied by quantitatively correct predictions of new behaviors in new circumstances. The kinds of theories Popper was thinking about never went that far.

A true science needs continuous measurement scales so that theories about the forms of relationships can be tested. This means that correlations have to be somewhere in the high 90s. True measurements, with normal measurement errors, require correlations of 0.99 upward. If this were universally understood among scientists, two things would happen. The first is that most statistical studies would end up in the wastebasket. The second is that the good studies would be done again and again, with successive refinements to reduce the scatter, until something of actual importance and usefulness was found.

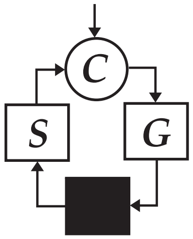
One of my objections to the statistical approach to understanding behavior is that after the first significant statistical measure is found, the experimenter quits the investigation and publishes. If you get a correlation of 0.8, p less than 0.05, your next question should be, "Where is all that variance coming from?" If you set your sights on 0.95, p less than 0.0000001, you won’t quit after the preliminary study, but will refine the hypothesis until you get real data.

Quantitative methods in conventional psychology don’t deal with quantitative data, despite the tremendous sophistication of statistical techniques. When you consider the models used in physics, where the systems are claimed by some to be “simple” relative to organisms, you find extremely complex structures in these models, extending from simple algebra, through systems of hundreds of differential equations, to tensor calculus. When you look at the models used in psychology, you find basically y ax + b. Of course, in order to see whether this model represents any regularity in a data set, you might have to apply very complex techniques for extracting signal from noise, but the basic model being tested is elementary, if that. So if the subject matter of psychology is so complex, why do psychologists try to handle it with such simple models?

The place where psychology is the least quantitative is in the data-taking stage. Most data exist in the form of simple and artificial events, which either occur (1) or don’t occur (0). The behaviors investigated are characterized in only the crudest qualitative ways; quantitative continuous measures of behavior almost never occur except in psychophysics.

When I read the psychology literature, I see almost nothing being investigated that strikes me as a real phenomenon. Even when something real-looking is investigated, I see no quantitative measurements being made. The *only* quantitative analysis that shows up in most articles is the statistics, which takes for granted that the data are about something and offers no explanations at all.

I think that the control-systems approach, which is fundamentally quantitative, offers the promise of handling even complex behavior in a way that is as clean as the methods of physics. I don’t buy the idea that psychologists have the problems they have because of the complexity of the subject matter. I think their problems come from a primarily non-quantitative, idiosyncratic, and disorganized approach to observing human behavior, and the acceptance of very low standards for what will be considered a fact of nature. The latter bothers me the most. You can’t base a science on facts that have only a 0.8 or 0.9 probability of being true. Such low-grade facts can’t be put together into any kind of extended argument that requires half a dozen facts to be true at once. You need facts with probabilities of 0.9999 or better—if you want to build an intellectual structure that will hang together. I don’t think that psychology has come anywhere near meeting that requirement, individual cases aside. I would argue that we do not yet have any *science* of psychology.



The Control Systems Group is a membership organization which supports the understanding of cybernetic control systems in organisms and their environments: *living control systems.* Academicians, clinicians, and other professionals in several disciplines, including biology, psychology, social work, economics, education, engineering, and philosophy, are members of the Group. Annual meetings have been held since 1985. The CSG Business Office is located at 73 Ridge Pl., CR 510, Durango, CO 81301; phone (303) 247-7986.

The CSG logo shows the generic structure of cybernetic control systems. A Comparator (C) computes the difference between a reference signal (represented by the arrow coming from above) and the output signal from Sensory (S) computation. The resulting difference signal is the input to the Gain generator (G). Disturbances (represented by the black box) alter the Gain generator output on the way to Sensory computation, where the negative-feedback loop is closed.