SYSTEMATIC ERRORS in PHYSICAL

The author is a consultant to the National Bureau of Standards on the statistical and mathematical design of experiments in physics, chemistry, and engineering. Dr. Youden joined NBS in 1948.

By W. J. Youden

PHYSICISTS today make very little use of statistical techniques. There was good reason for the minor role so long accorded the statistical evaluation of the errors in physical constants. When two laboratories make independent determinations, each may attach to its "best" value a $\pm$ sign followed by an estimate s of the error. This estimate of the error is often based upon a series of observations made under carefully controlled conditions. Experimenters soon discovered that if laboratories A and B reported values C_A and C_B for the same constant, the difference Δ between C_A and C_B was almost always a large multiple of the estimated error s_a (or s_b). Obviously these calculated errors had no more to do with the real errors than the neatness of the laboratory or the promptness with which the investigator answered his mail.

Statisticians in turn sensed that all the observations made in one laboratory, with one piece of equipment, were afflicted with some fairly constant and unknown increment that was a resultant of biases associated with the method of measurement, with the particular assembly of apparatus, and perhaps with some more or less persistent characteristics of the environment. The statistician saw no way either to detect or to assess these "constant" errors. Consequently, statisticians concentrated on other activities where random errors were all that really mattered. The comparison of the yields obtained from two or more varieties of wheat involves only comparisons. Similarly the chemist, seeking to find for an industrial process a set of operating conditions that will give maximum yield, or maximum profit, can compare runs and not worry much that all the results may be half a percent high. That may be discovered later, when the annual inventory is taken.

Both physicists and statisticians apparently agreed to part company. There remained the custom of calculating and reporting the precision of the measurements, partly to establish that very precise habits of work were maintained, and partly in the hope that more weight would be given to a determination if a very small precision error was attached to the result. All recognized that a small precision error was necessary but gave no guarantee that the reported average was close to the truth.

For decades there has been but little contact between experimental physics and statistics, and I think that both parties have been the losers for giving up so easily. Statisticians were not aware that many of the physical measurements either approximate, almost exactly, certain ideal statistical models or else suggest the invention of statistical models that would extend statistical theory. The physicist, in turn, relying on his experimental skill, continued to track down the sources of his errors by traditional methods and overlooked certain advantageous ways of combining his observations.

This paper discusses three main topics. First some remarks will be made regarding the statistical confidence limits that apply to two or three independent determinations of a constant. The major section deals with what appears to be a plausible explanation for the unexpectedly large differences between the values obtained in different laboratories. The last portion presents some statistical aids for tracking down the causes for disagreement among laboratories.

Independent Determinations of a Constant

SUPPOSE laboratories A and B report the values C_a and C_b for the same constant. Precision estimates s_a and s_b may or may not be given. Perhaps the investigators have searched their souls and ventured to indicate the likely maximum errors in the values reported. These estimated errors generally do not determine the opinions of other laboratories regarding these two results. Depending on the laboratory visited, you may encounter one of four possible opinions:

1. The laboratory favors C_a and discounts C_b
2. The laboratory favors C_b and discounts C_a
3. The laboratory believes C_a and C_b are of about equal merit
4. The laboratory is a sceptic and believes both C_a and C_b unreliable.

If the laboratories are approximately split between the first two opinions and one of the determinations is close to correct, then the obvious statistical conclusion can be drawn that about half of the laboratories will eventually be disappointed. Perhaps all will be disappointed if neither determination is near the correct value.

If most of the laboratories are of the fourth opinion, clearly there is no statistical problem. But if a majority

ꞮONSTANTS

of the laboratories feel that both results are worthy and that there is little to choose between the two determinations, then some statistical remarks may be made. We are going to suppose that the method of measurement is a new one and consequently there is no other information available than that contained in the two results already in hand. That there will be some difference between the two results is to be expected. Examination of the difference between the two results tells us little because we have no way of knowing whether this difference is smaller or larger than usual. Statistical tables show that if the average difference between duplicates is ten units, then individual differences of from one to thirty units are not uncommon. So a single difference may be very misleading.

Suppose a third laboratory is about to make a report. If we assume that the three results are independent of the order in which they were obtained, some simple logic suggests that there is a one-third chance that the last result reported will be intermediate in magnitude between the first two results reported. Denote the smallest, middle, and largest results by *s*, *m*, and *l*. These three letters can be arranged in six orders: *sml*, *slm*, *msl*, *mls*, *lsm*, and *lms*. For two of these six sequences the middle result *m* is the last in the sequence. Consequently, without ever knowing the first two results, it is a fair gamble to bet one to two that the third result will lie between the first two results.

Notice, too, that this logic holds quite apart from any knowledge as to how closely the first two agree. Of course, if by chance the first two values are identical or nearly so, one might argue that it would be less likely to get a third result between them than if the first two did not agree closely. But just what other standards can one produce to say, in any particular case, what would be close agreement, or what would constitute poor agreement, if these two results constitute all the information available?

A closely similar question, given two equally esteemed results, is: What is the chance that the two values C_a and C_b bracket the correct value? The answer is one half. After all, the correct result does not go gallivanting around the way a third independent result might and contributes no error. It is quite remarkable that this conclusion rests on a very modest assumption about the underlying distribution, of which these two constitute our sole information. We have only to concede that if a goodly number of qualified laboratories undertook to make determinations, that about half the determinations would be smaller and the remainder larger than the correct value. Symmetry is all that is required. So it is a coin-tossing problem with two coins where heads refer to plus deviations and tails to minus deviations. A quarter of the time we get two heads (both results high), a quarter of the time two tails (both results low), and half the time a head and a tail, or deviations of unlike sign which means that the results bracket the correct value.

I hasten to admit it is conceivable, through some defect in theory, that all the results are afflicted with a component error of the same sign and this will spoil our coin-tossing game. But this is speculation and tantamount to saying that it is useless *ever* to venture an opinion about the confidence to be placed in the determinations. It does seem appropriate to be aware of the probabilities that I have given even if one cautiously states the assumptions upon which the probabilities are calculated.

Now a probability of one half is not a very comforting figure and it is a natural thing to wonder how we might extend our thinking to limits outside the two reported values in order to attain a greater confidence that the correct value lies within these limits. Let $C_a - C_b = \Delta$, where $C_a > C_b$, and suppose we consider limits of the following kind:

$$\text{Upper limit} = C_a + k\Delta, \quad \text{Lower limit} = C_b - k\Delta.$$

It now becomes necessary to examine how sensitive our confidence is to the kind of distribution that would fit a collection of such determinations. Suppose we assume first the traditional normal distribution. Then for k equal to one, the probability is about 0.8; that is, adding the difference between two results to the larger one, and subtracting it from the smaller, gives limits that four times out of five should bracket the correct value. If instead of the normal distribution, we imagine that a determination is equally likely to fall anywhere within some finite, but unknown, interval centered on the unknown correct value, the probability drops from 0.80 to 0.75. And there is a vast difference between the bell-shaped normal distribution and the "rectangular" distribution of equal probability for all values over a finite range.

Table 1 shows, for the normal distribution, how the probability of bracketing the correct value between $C_a + k\Delta$ and $C_b - k\Delta$ increases with k. Remember that Δ is the difference between two determinations that are accorded equal weight.

Table 1. Probability, P, that $C_a+k\Delta$ and $C_b-k\Delta$ bracket the correct value. Normal distribution assumed. Equal weight accorded C_a and C_b; $\Delta=C_a-C_b$.

k	0	1	2	3	4	5	6	7
P	0.5000	0.795	0.874	0.910	0.930	0.942	0.951	0.958

Sad to say, it takes an over-all spread between the upper and lower limits of 13 times the difference between the two determinations to attain the traditional 95-percent confidence limits. You may reply: "Nonsense. Things are not that bad." But you should be prepared to justify your comment. After all, in the light of the peripatetic wanderings of the "accepted" value of some of our constants, how can you, from just two determinations, form a better judgment about the correct value?

The real explanation of the wide limits required in Table 1 is the small amount of information we have on Δ. One pair may give a Δ considerably smaller or considerably larger than the average Δ if many such pairs were available. The way to improve matters is to get additional, truly independent determinations.

The gain in assurance that comes from a third independent determination at first seems disproportionately large. The narrowing of the confidence limits comes not so much from being able to average three rather than two, but from having a firmer grip on the extent of agreement that may be expected among independent determinations. The chance that the three results bracket the correct value rises from one half to three quarters. That is, the chance that both a tail and a head will be obtained when three coins are tossed is six out of eight. If the difference between the largest and smallest of the three values is added to the largest value and subtracted from the smallest value, we obtain confidence limits that have slightly better than a twenty to one chance of bracketing the correct value. Of course the same assumptions discussed above for two determinations are made here, too. Even if there are grave reservations about these assumptions, one can say that the chances are no better than those indicated. A bound has been set to our optimism.

In spite of the difficulties that arise in estimating the error in a constant most scientists agree that the effort should be made. Professor Bridgeman in his talk at the 1960 Gordon Conference on "Information Processing for Critical Tables of Scientific Data",[1] emphasized that critical tables should endeavor to present the "best" value and to make some estimate of the "probable" error of the value selected.

Composite Character of Systematic Errors

SOMETIMES successive measurements may be made in a time interval so short that it is reasonable to regard the measurements as being made with no changes in environment, apparatus, or any other condition that might affect the measurement. Given adequate precision a reasonable number of measurements serve to establish an average that very closely characterizes the measuring system during this interval. This average will differ more or less from the correct value. This departure from the correct value is plainly the algebraic sum of several small effects. For example, the diameter of a diaphram, the resistance of a coil, the temperature and volume of a chamber, and similar quantities will all be assigned values that depart in some degree from the actual values that existed while the measurements were made. These deviations from the actual values influence the outcome—and each may either add or subtract some small increment to the measurements. The experimenter has surely tried to keep these various increments somewhat the same in size and usually he would say it was a toss up as to the sign of each increment.

As an example, a recent determination of *g* presents two sets of 32 measurements—one set made with one rule, the other set with a second rule.[2] Fig. 1, taken from this paper, shows the distribution of the measurements for each set. No elaborate statistical test is required to make convincing the reality of the difference between the means of the two sets. Doubtless there were other components or conditions that had similar increments.

Imagine ten such increments of about equal magnitude but unpredictable in sign. Now the experimenter is surely at the mercy of the laws of chance. There are six different algebraic sums (each either plus or minus) depending on how fate has grouped the signs of these increments.

Division of the signs	*Algebraic sum*	*Frequency*	
5 and 5	0	252	912
4 and 6	± 2	420	
3 and 7	± 4	240	
2 and 8	± 6	90	112
1 and 9	± 8	20	
0 and 10	± 10	2	
		1024	

The foregoing tabulation shows that about once out of nine times the increments gang up on the helpless experimenter and introduce a composite systematic error at least six times as large as the small "uncertainty" he has achieved in his values for the components in his apparatus. There is a chance in three of a net sum of four or more increments. If the experiment is repeated in another laboratory, the same situation holds and half the time the two composite net sums will be of opposite sign. We now see how the difference between the results from the two laboratories may be an order of magnitude greater than the standard of accuracy set for the individual components.

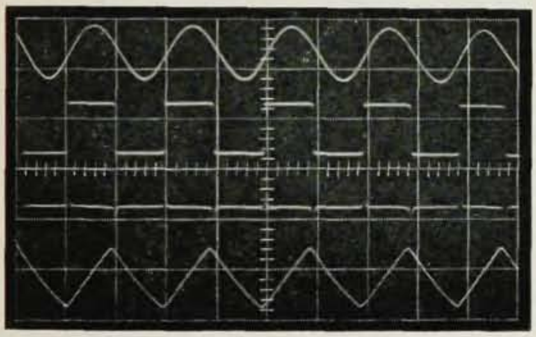
The individual increments are taken as equal in size to simplify the presentation. If the increments vary from small to large, the effect is very nearly the same if their average magnitude equals the "standard" increment used above. While there is a certain amount of cancellation because there may be both plus and minus increments, it is the net *sum* that matters. There is no averaging out here. So the distribution of these "sums" depends on the average size and the number of contributing increments. Experimenters properly enough

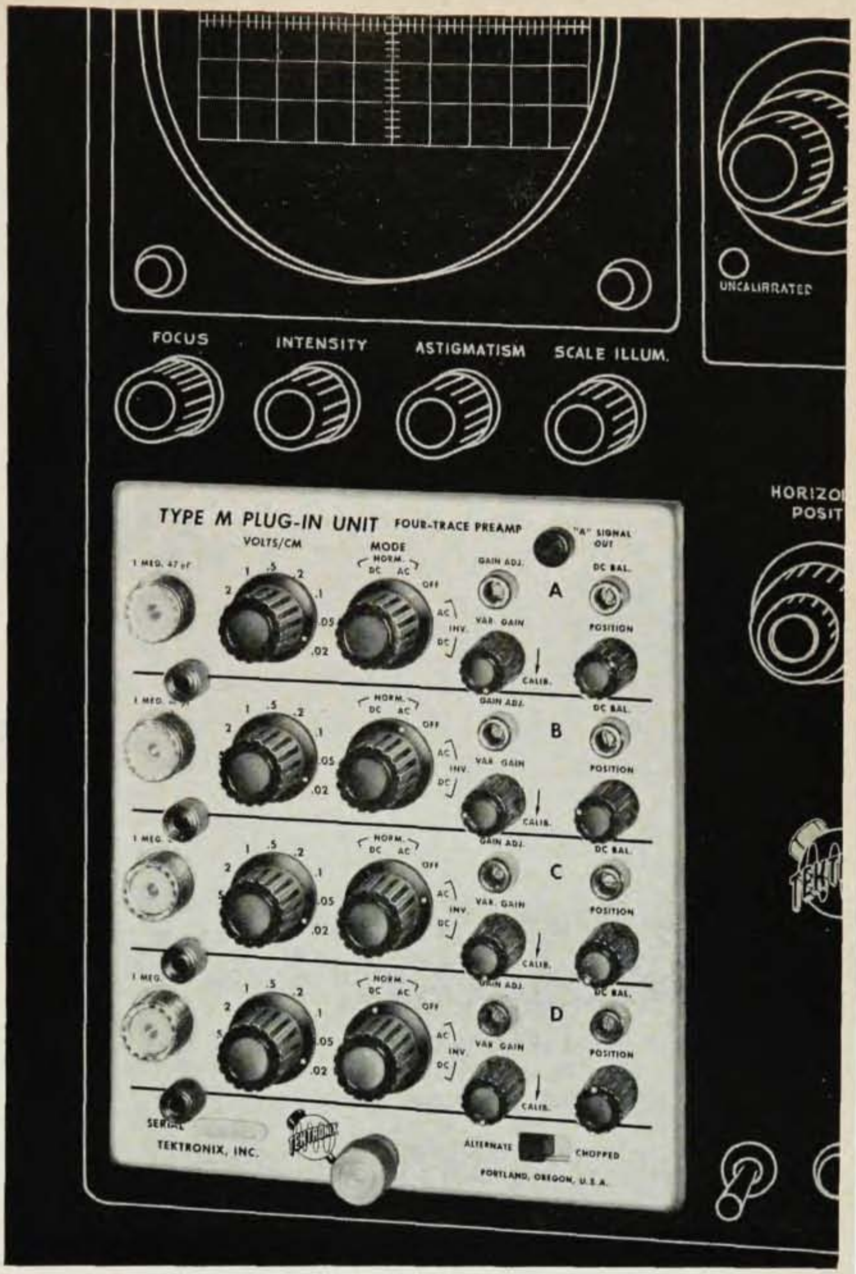
NEW

4-TRACE

PREAMPLIFIER

for Tektronix Oscilloscopes that accept letter-series plug-in units





TYPE M UNIT Seventeenth in the letter-series of plug-in units, the new Type M Unit adds multiple-trace displays to the wide range of applications possible with your Tektronix oscilloscope.

With a Type M Unit, you can observe up to four signals—either separately, or in any combination.

Independent controls for each amplifier channel permit you to position, attenuate, invert input signals as desired.

Other convenient preamplifier features—such as triggered or free-running electronic switching . . . ac-coupling or dc-coupling . . . and, after initial hookup, little or no cable switching—ideally suit the new Type M Unit for multiple-trace presentations in the laboratory or in the field.

For more information about this new 4-Trace Preamplifier, please call your Tektronix Field Engineer.

CHARACTERISTICS

Operating Modes—Any combination of one to four channels electronically switched—at the end of each sweep or at a free-running rate of about 1 mc (1 μsec width samples). Or each channel separately. *Channel Sensitivity*—20 mv/cm to 10 v/cm in 9 calibrated steps. Continuously variable uncalibrated between steps, and to 25 v/cm. *Channel A Signal*—available on front panel for optimum triggering in some applications. *Frequency Response and Risetime*—With Type 540-Series and Type 550-Series Oscilloscopes dc to 20 mc, 17 nsec. With Type 531A, 533A, 535A Oscilloscopes dc to 14 mc, 25 nsec. *Constant Input Impedance*—at all attenuator settings.

Type M Plug-in Unit $455.00
f.o.b. factory

Tektronix, Inc. *P. O. BOX 500 · BEAVERTON, OREGON / MItchell 4-0161 · TWX—BEAV 311 · Cable: TEKTRONIX*

TEKTRONIX FIELD OFFICES: Albuquerque, N. Mex. • Atlanta, Ga. • Baltimore (Towson) Md. • Boston (Lexington) Mass. • Buffalo, N.Y. • Chicago (Park Ridge) Ill. • Cleveland, Ohio • Dallas, Texas • Dayton, Ohio Denver, Colo. • Detroit (Lathrup Village) Mich. • Endicott (Endwell) N.Y. • Greensboro, N.C. • Houston, Texas • Indianapolis, Ind. • Kansas City (Mission) Kan. • Los Angeles, Calif. Area (East Los Angeles, Encino • West Los Angeles) • Minneapolis, Minn. • Montreal, Quebec, Canada • New York City Area (Albertson, L.I., N.Y. • Stamford, Conn. • Union, N.J.) • Orlando, Fla. • Philadelphia, Pa. • Phoenix (Scottsdale) Ariz. Poughkeepsie, N.Y. • San Diego, Calif. • San Francisco, Calif. Area (Lafayette, Palo Alto) • St. Petersburg, Fla. • Syracuse, N.Y. • Toronto (Willowdale) Ont., Canada • Washington, D.C. (Annandale, Va.).

TEKTRONIX ENGINEERING REPRESENTATIVES: Hawthorne Electronics: Portland, Oregon; Seattle, Washington • Kentron Hawaii Ltd., Honolulu, Hawaii. Tektronix is represented in twenty overseas countries by qualified engineering organizations.

'n Europe please contact Tektronix International A.G., Terrassenweg 1A, Zug, Switzerland, Phone (042) 4-91-92, for the address of the Tektronix Representative in your country.

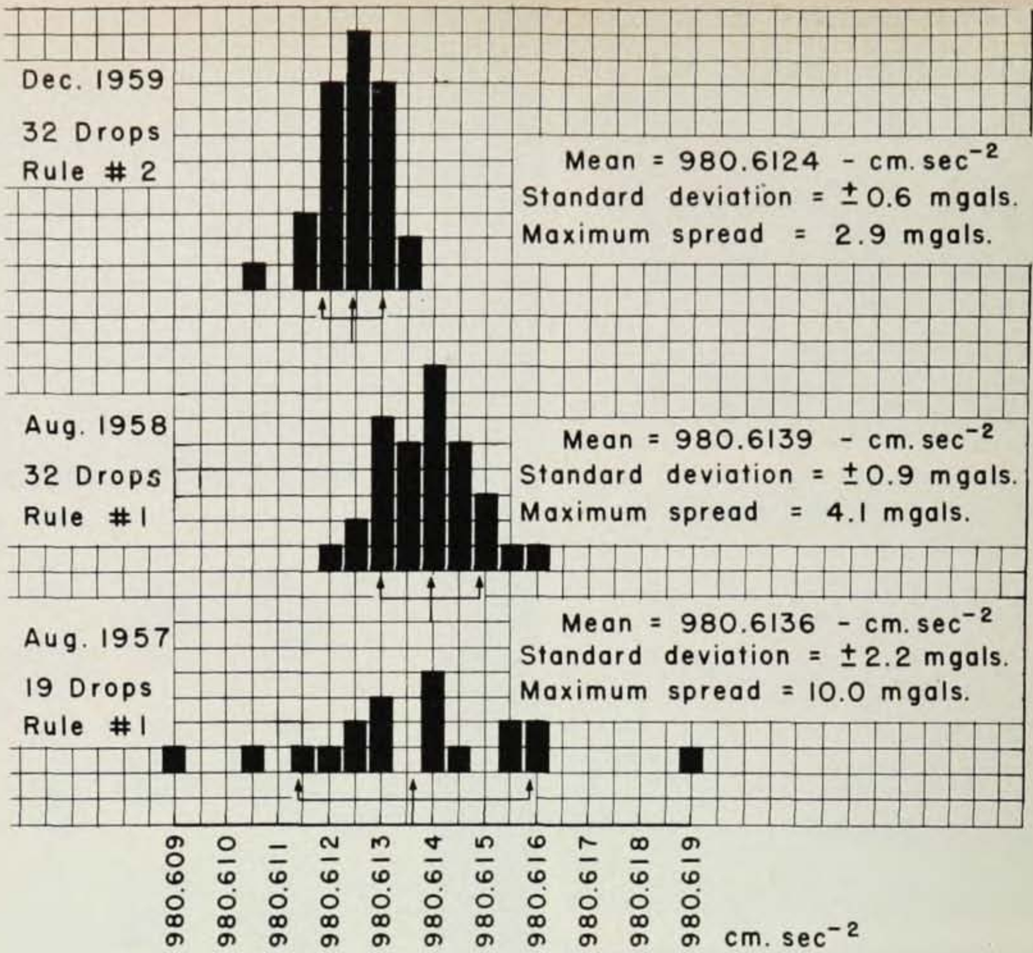


Fig. 1. Measurement of gravity constant

direct their best efforts to the detection and reduction of the larger increments because this is the most effective way to reduce the average size of the increments.

Detection of Increments of Systematic Errors

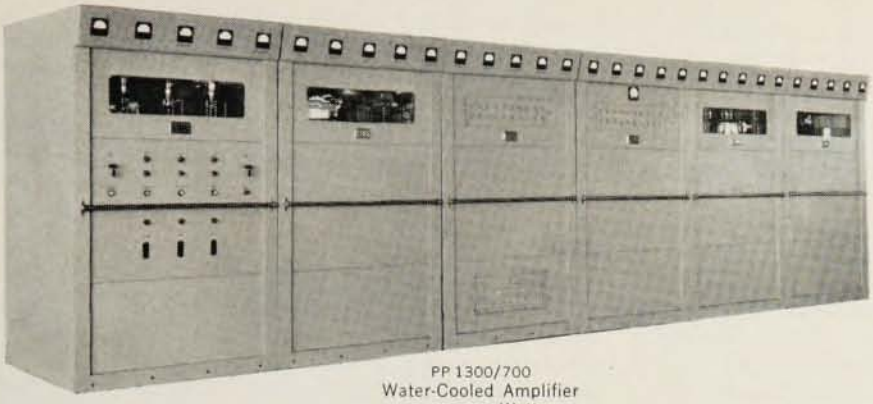
WE have seen that an aggregate of systematic errors, all of them individually relatively small, can nevertheless sum up in such a fashion as to produce a substantial net displacement from the correct result. The detection of small systematic errors, and by that I mean errors comparable to the precision error, requires a considerable number of repeat measurements. Fig. 1 shows 32 repeat measurements of the gravitation constant g with each of two different rules. The repeat measurements with a rule cluster around a central value for that rule and offer convincing evidence that there is a real difference between the averages for these two rules. The shape of the scatter of the measurements around their average is what would be expected on the basis of the normal distribution of errors. Suppose the difference Δ between the averages for the two rules is equal to s, the standard deviation of the repeat measurements. Then, reference to tables for the normal distribution shows that it is necessary to make at least eight repeat measurements on each rule before we can conclude, with 95% confidence limits, that the rules differ at all in their mean values.

The important thing here is, that within one laboratory, the *precision* measure of error is the proper measure to use in evaluating differential effects of such substitutions of components of the apparatus, or in evaluating effects of changing environmental conditions. Dorsey, in a lengthy paper published in 1944,[3] gives on pages 10 and 11 some pointed remarks on the necessity of examining the effect of changing the adjustments of the apparatus. I quote one sentence.

> Readjusting the apparatus, he (the experimenter) will proceed to change, one by one, every condition he can think of that seems by any chance likely to affect his result, and some that do not, in every case pushing the change well beyond any that seems at all likely to occur accidentally.

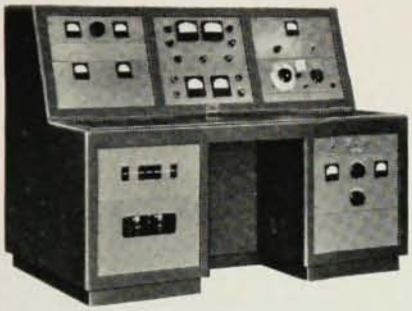
Excerpts from Dorsey's 110-page article are given in a paper by Dorsey and Eisenhart.[4]

The single sentence quoted above is particularly interesting because Dorsey saw the direction in which progress was to be made. In the nearly twenty years since Dorsey prepared his remarks we have made considerable progress in the direction he indicated. We see that not only should the adjustments be changed, but whenever possible there should be at least duplicate components for certain vital parts of the apparatus. The use of two rules, as exhibited in Fig. 1, shows how much the results are at the mercy of a single rule. Clearly the only thing to do is to take the average for

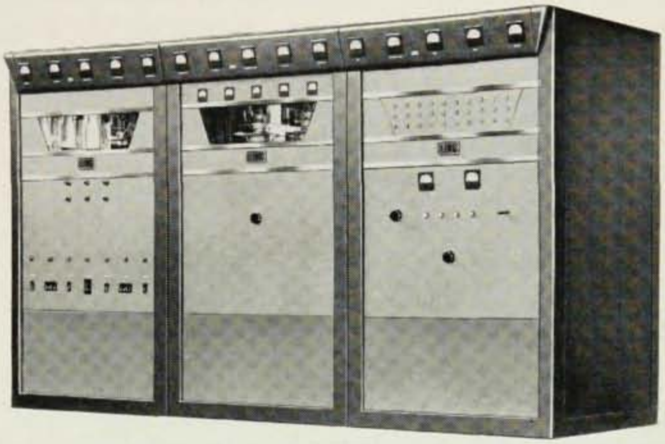


PP 1300/700
Water-Cooled Amplifier
1,300,000 Watts

PP 10/15 Modular
Sonar Transmitter



CP 5/6 5000 VA Amplifier



PP 120/150
120,000 VA Amplifier

the *two* rules, and there are *only* two rules. The prediction as to what might happen with more rules throws us right back to the discussion in the first part of this paper. Incidentally, Dorsey's recommendation that substantial changes be introduced in the conditions indicates that he found it difficult to detect the effects of small changes.

The previously-quoted sentence contains the phrase, "one by one". Change the adjustments "one by one" is the way we all learned to experiment. The interpretation is easy then because, for example, if we merely substitute one rule for another, any effect is obviously to be credited to the substitution of one rule for the other. In the intervening years since Dorsey wrote there has been a good deal of activity in the devising of more efficient programs for evaluating the effect of just such changes in adjustments or substitutions of components in the apparatus.

If there are a number of possible adjustments and components to investigate, the total number of measurements may become very large because a considerable number of repeat measurements must be made for each assembly and each adjustment. There are really two parts to this problem. If, for example, the experimenter winnows his choices down to seven alternatives (including both adjustments and substitutions for components) does that mean that he need try all 2^7, or 128, possible combinations? Experimenters have already answered this question. They designate some standard initial assembly and set of adjustments and then proceed to change, one by one, the seven items under consideration. Some measurements are made under the initial state; an item is changed, and another set of measurements made. Whatever was changed is put back to the initial state and a second item changed. There will be eight such sets and a goodly number of measurements are required in each set.

Today, as a result of some purely theoretical inquiries into what statisticians term *weighing designs*, we know that seven variables could have been equally well evaluated with one fourth the usual number of measurements. Or, and the prospect is enticing, we could have detected, and perhaps corrected, systematic effects only half as large as those just detectable under the "one by one" approach. I say that these were theoretical statistical inquiries because statisticians were mainly concerned with biological and chemical problems that involved *major* changes in the variables. In such investigations there are mutual interactions of the variables that pose quite different problems. Here the changes in the variables are minute. The differential effect of substituting one rule for another almost identical rule (as in Fig. 1) would be virtually unaltered even if some other set of initial conditions had been chosen.

Statisticians were unaware of the extremely important problems posed in the evaluation of physical constants. Yates was the first statistician to suggest and name "weighing designs" in an incidental paragraph in a paper [5] in 1935. In fact, Yates belittled the designs because he deemed it most unlikely that any problems appropriate for such designs really existed. It is interesting that other statisticians,[6-10] in a purely theoretical way, embellished the idea advanced by Yates. None saw the possibilities that exist for application in very precise physical measurements. And we lack, even now, an adequate exploration of the programs that might serve the needs of those who determine physical constants.

Once it is recognized that the effect of a very small change in a variable does not depend on the other variables, provided that these other variables also are held within very close limits, the way is open to change more than one variable at a time. I illustrate this principle first for the case of three variables x, y, and z, which may be assigned other nearby values x', y', and z'. Let us designate the standard initial condition by x, y, and z and let these serve as the coordinates of the origin in the three-dimensional graph shown in Fig. 2.

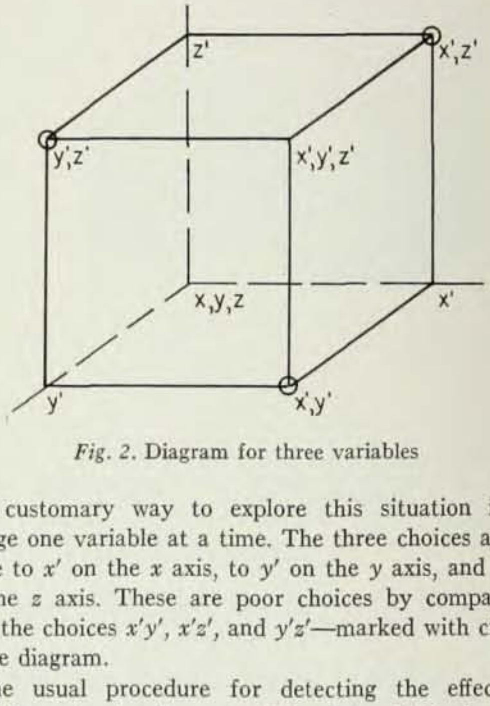


Fig. 2. Diagram for three variables

The customary way to explore this situation is to change one variable at a time. The three choices are to move to x' on the x axis, to y' on the y axis, and to z' on the z axis. These are poor choices by comparison with the choices $x'y'$, $x'z'$, and $y'z'$—marked with circles in the diagram.

The usual procedure for detecting the effect of changing x to x' makes use of the data obtained at the two points x,y,z and $x'yz$. The more efficient method for detecting the effect of changing x to x' makes use of the data obtained at all four points, x,y,z; $x,y'z'$; $x'y'z$ and $x'yz'$. Two of these sets involve x and two involve x' so the data are grouped accordingly.

xyz	$x'y'z$
$xy'z'$	$x'yz'$

The two sets with x include y and z and y' and z'. So the average value for x incorporates the effects associated with y, y', z and z'. This is also visibly true for the two sets with x'. Therefore, the effect of changing x to x' will be given by comparing the *average* of all

the data taken at x with the *average* of all the data taken at x'. Inspection discloses that the four sets may be partitioned into appropriate pairs to detect the effect of changing y to y', or z to z'.

xyz	$xy'z'$	xyz	$xy'z'$
$x'yz'$	$x'y'z$	$x'y'z$	$x'yz'$

The basic idea here is so important that I illustrate it again for the case of just two variables. Recall the 32 measurements made with rule 1 (r) and the additional 32 measurements made with rule 2 (R). Group them in opposing groups as next shown.

rrrr	*RRRR*
rrrr	*RRRR*
rrrr	*RRRR*
rrrr	*RRRR*
rrrr	*RRRR*
rrrr	*RRRR*
rrrr	*RRRR*
rrrr	*RRRR*

If half of the measurements in each group were made with another variable at s and the remainder at S, the measurements may be segregated into four sets.

Set 1		*Set 2*		*Set 3*		*Set 4*	
rs	*rs*	*rS*	*rS*	*RS*	*RS*	*Rs*	*Rs*
rs	*rs*	*rS*	*rS*	*RS*	*RS*	*Rs*	*Rs*
rs	*rs*	*rS*	*rS*	*RS*	*RS*	*Rs*	*Rs*
rs	*rs*	*rS*	*rS*	*RS*	*RS*	*Rs*	*Rs*
rs	*rs*	*rS*	*rS*	*RS*	*RS*	*Rs*	*Rs*
rs	*rs*	*rS*	*rS*	*RS*	*RS*	*Rs*	*Rs*
rs	*rs*	*rS*	*rS*	*RS*	*RS*	*Rs*	*Rs*
rs	*rs*	*rS*	*rS*	*RS*	*RS*	*Rs*	*Rs*

Rule r is present in sets 1 and 2 and rule R in sets 3 and 4. The other variable is put at s in sets 1 and 4 and at S in sets 2 and 3. We may now play both ends against the middle pairs of sets and evaluate the effects of s and S. The data are used twice over. If the set size is reduced from 16 to 8, 7 variables may be studied with these same 64 measurements.

That is, *all* the data taken are used to evaluate the effect of changing each variable. Either fewer repeat measurements are required at each combination, or more variables may be investigated with the same number of measurements. Indeed, the more variables that are investigated in this manner, the more efficient this method becomes. Seven variables lend themselves to an especially elegant sequence of seven partitions of eight sets into contrasting sets of four sets against four sets. This example, shown in Table 2, I am glad to report, is the one first mentioned by Yates twenty-five years ago. You may note that four of the initial conditions are changed each time.

It would be a pleasure indeed if I could include here a small catalog of programs extensive enough to meet the situations likely to occur in practice. I can point out that the program shown in Table 2 can be used for fewer than seven variables by ignoring one or more of the changes. It will still be necessary to work with eight different combinations. Similarly, in a paper by Plackett and Burman [10] schemes for 12, 16, and more variables are given. Variables may be ignored here, too, but the number of combinations is not reduced.

The minimum number of combinations required is one more than the number of variables, if just two alternatives are used for each variable. The substitution of a component is sometimes a tedious affair so there is sure to be interest in programs involving a minimum number of combinations. I have tried my hand at this game and offer the program shown in Table 3 for studying five variables with *six combinations*. Each effect is measured using the results of four of the six combinations, divided two against two.

Table 2. Program for seven variables with eight sets.

1	2	3	4	5	6	7	8
t	t	t	t	T	T	T	T
u	u	U	U	u	u	U	U
v	V	v	V	v	V	v	V
w	w	W	W	W	W	w	w
x	X	x	X	X	x	X	x
y	Y	Y	y	y	Y	Y	y
z	Z	Z	z	Z	z	z	Z

Table 3. Five variables in six sets.

1	2	3	4	5	6	
v	V	v	V	v	V	$v-V=(1+5)/2-(2+6)/2$
w	w	W	W	W	W	$w-W=(1+2)/2-(5+6)/2$
x	X	X	x	X	x	$x-X=(1+4)/2-(2+3)/2$
y	Y	y	Y	Y	y	$y-Y=(3+6)/2-(4+5)/2$
z	z	Z	Z	z	z	$z-Z=(5+6)/2-(3+4)/2$
						$\mathrm{av}=(1+2+3+4)/4$

Table 4. Program for three variables, two with three choices, one with two choices.

Variables

x $\mathbf{x}$ X
y $\mathbf{y}$ Y
z Z

Six sets

1	2	3	4	5	6
x	x	$\mathbf{x}$	$\mathbf{x}$	X	X
y	$\mathbf{y}$	y	Y	$\mathbf{y}$	Y
z	Z	Z	z	z	Z

	x	$\mathbf{x}$	X
y	z 2	Z −2	
$\mathbf{y}$	Z 1		z −1
Y		z −1	Z 1

Above coefficients are weighing factors to estimate x–$\mathbf{x}$

There will be times when more than two choices are possible and of interest for some of the variables. I regret to say that the enumeration of efficient designs for such mixtures of two and three choices has hardly begun. Let me illustrate with a simple case of three choices for each of two variables and two choices for a third variable. There are $2 \times 3 \times 3$ possible combinations, and a minimum of six sets are necessary to separate the individual effects of these variables. The problem is to pick that subset of six from the eighteen available sets that will lead to the most efficient evaluation of the effects of the variables. I suspect that the program shown in Table 4 is as good as any, just on the basis of the appealing symmetry.

In the squares are indicated certain factors and these are the factors to be used in evaluating the effect of changing x to $\mathbf{x}$. Notice that the estimation of the effect of a change in the variable involves a weighted average of the six set results. The best average for the constant gives equal weight to all six sets. Similar sets of constants apply for evaluating $\mathbf{x} - X$, $\mathbf{y} - Y$, etc.

The essential point regarding these illustrative programs is that certain combinations lend themselves to an efficient use of the data, that is, to a more sensitive scrutiny of the possible sources of error. The one-at-a-time technique is one of the least efficient programs. The small individual contributions to error that are associated with uncertainties in values assigned to component quantities are not easy to detect. A planned set of combinations will rank the various sources of error in order of magnitude and reveal where the program is weakest. Statistical techniques will not remove errors but they can help in isolating the important sources of error.

Enduring Values

THERE is more in this discussion than the matter of efficiency. The several variables, chosen by the experimenter because they may influence the result, are actually put to the test. At present the investigator has two ways to arrive at an opinion or guess as to the error introduced by any one of the quantities which he would like to know exactly when he introduces it into his computations. He may, on his judgment, hazard a guess as to the maximum uncertainty in each of the relevant quantities. Alternatively, he may accept the estimates of others—e.g., the estimate of the man who measured the length of the rules used in the determination of g. Thermometers, weights, resistances, purities, standard cells—the list is endless—they may all be obtained with some sort of statement from the calibrating source. It is easy to push responsibility off this way. And we go on getting determinations from different laboratories that disagree much more than anticipated even when the claimed uncertainties in the components are included. Maybe it is time to check these indispensable bits of information. If the Coast and Geodetic Survey measures the distance employed in a determination of the velocity of light—ask them to measure two or three distances. The above schemes will soon put these measurements to the test. A choice of resistances, diaphragm diameters, thermometers—all should be made to run the gauntlet.

Yes, I know, all the resistances *may* be subject to the same bias. The two rules used in the determination of g *may* share a common increment that will not be revealed by the data. But there *is* a difference between the two rules and now our estimate of the limits of error can allow for this difference. We must use more than one rule, or we will not have the data to estimate this source of error.

I return to the summation of the systematic errors associated with the individual components. The use of two or more choices creates the possibility that the choices differ in the signs of their systematic errors. The final value reported will be an *average* of the results obtained with the several sets—each set a unique combination of components and conditions. The individual summations of the separate sets now enter into an average with all the advantages that come from taking an average. Furthermore, the spread of the results for the several sets will surely give a more realistic idea of the uncertainty in the final result than that obtained from hopeful guesses.

There is another matter that cannot be glossed over. Suppose the measurements are made according to some carefully thought out program similar to the suggested weighing designs. Admittedly this limits the freedom of the investigator. The experimenter likes to be free to follow some inspired hunch. He often wants to try some alteration in the apparatus, or in the conditions, on the chance that his spontaneous idea has merit. This might be regarded as the art rather than the science of experimentation. The investigator should consider how often such ideas pay off and also the large number of measurements required to detect small effects, when tempted by such ideas.

I personally hold that allowance should be made for "shots in the dark". If the planned program is allotted, say, around three quarters of the measurement time, there would still be opportunity for imaginative excursions. Even if these isolated shots lack the power that they would have if incorporated in the planned program, they add a lot of zest to experimentation.

We all know that a serious effort to determine a physical constant is not undertaken lightly. The dominating thought in the mind of the investigator is to arrive at an enduring value. What is an enduring value? I suggest that it is a result coupled with a stated zone of uncertainty that includes the value that future workers will converge upon.

References

1. P. W. Bridgman, "Critique of critical tables", Proc. Nat. Acad. Sci. **46**, 1394 (1960).
2. H. Preston-Thomas, L. G. Turnbull, E. Green, T. M. Dauphiness, and S. N. Kalra, "An absolute measurement of the acceleration due to gravity", Ottawa Division of Applied Physics, National Research Council, Ottawa, Canada.
3. N. E. Dorsey, "The velocity of light", Trans. Am. Phil. Soc. **34**, 1 (1944).
4. N. E. Dorsey and C. Eisenhart, "On absolute measurements", Sci. Monthly **77**, 103 (1953).
5. F. Yates, "Complex experiments", Suppl. J. Roy. Stat. Soc. **2**, 181 (1935).
6. H. Hotelling, "Some problems in weighing and other experimental techniques", Ann. Math. Stat. **15**, 297 (1944).
7. O. Kempthorne, "The factorial approach to the weighing problem", Ann. Math. Stat. **19**, 238 (1948).
8. K. Kishen, "On the design of experiments for weighing and making other types of measurements", Ann. Math. Stat. **16**, 294 (1945).
9. A. M. Mood, "On Hotelling's weighing problem", Ann. Math. Stat. **17**, 432 (1946).
10. R. I. Plackett and J. P. Burman, "The design of optimum multifactorial experiments", Biometrika **33**, 305 (1946).