STATISTICAL PROBLEMS ARISING IN THE ESTABLISHMENT OF PHYSICAL STANDARDS

W. J. YOUDEN
NATIONAL BUREAU OF STANDARDS

1. Introduction

Scientific research, modern technology and commerce depend on the establishment of adequate physical standards and reliable values for the physical and chemical properties of an ever growing list of materials. Workers undertaking to establish these values are inevitably forced to distinguish between precision and accuracy. In fact the task of these workers is to make available such satisfactory standards and values of the properties that the vast majority of measurements made by others involve only comparative experiments. Tremendous care coupled with scientific ingenuity are the traditional earmarks of work directed to the determination of the constants that fill the pages of scientific and engineering handbooks. One might have expected that, along with this care and ingenuity, there would be found an active utilization of statistical techniques. In fact, only in the last decade or so has statistics been given much opportunity to contribute to these exacting scientific tasks.

There are two important reasons why relatively little use was made of statistical techniques. In the first place the early workers often assigned to their work calculated estimates of errors that were subsequently proven to have been most optimistic. The evidence was unmistakable because a later value for a constant, obtained by improved scientific techniques, would differ from the earlier value by possibly fifty times the error assigned to the earlier value. Naturally enough these estimates of the error were soon regarded by experimenters as relatively worthless. The other important reason for the disinterest in statistical techniques is that statisticians themselves were conscious that the successful applications of statistical techniques involved only comparative measurements. The most striking early examples had to do with agricultural field trials. A further obstacle to the use of modern statistical tools was that laboratory experimentation differed in many respects from the agricultural experimentation that was familiar to most applied statisticians. Often physical and chemical apparatus is complex and not easily understood without some background in the physical sciences. Recognition of the opportunities for the successful application of statistical ideas required a good understanding of the complex apparatus used in the experiments.

This essay undertakes to point out some important differences between agricultural experimentation and laboratory experimentation even when both are concerned with comparative measurements. In addition there will be pointed out some of the ways in which statistics can serve the experimenter who is trying to establish the correct values of the quantities he is measuring.

2. Classical Experimental Design

Statisticians use the expression "experimental design" in connection with certain statistical requirements for the sound interpretation of experimental data. Prominent among these requirements are the notions of randomization and replication. Randomization refers to the ordering, either spatially or temporally, or both, of the units corresponding to the experimental values. Randomization provides protection against both known and unknown possible sources of bias that might otherwise unduly distort the experimental values. Replication provides an estimate of the experimental error and hence the means for evaluating the experimental data. One of the early tenets in this connection was that each experimental project would make available an estimate of the experimental error applicable to that particular project. The data obtained in the study could then be interpreted without drawing upon other sources of experimental data.

For many years the development of experimental designs reflected the particular circumstances that were associated with the application of experimental design in agricultural field trials. The pronounced heterogeneity of soil fertility over quite modest areas had long been a major stumbling block in the testing of crop varieties, fertilizers and techniques of cultivation. The variation in soil fertility forced additional replication and this in turn increased the size of the experimental area. The larger area brought with it increased variability in soil fertility and consequently was, to some extent, self-defeating. The arrangement of complete replications in compact blocks permitted extensive replication without increase in the experimental error. This method of "local control" often brought dramatic successes in place of inconclusive or misleading results formerly obtained.

The undoubted success of experimental design in this area of application led to extensive elaboration in the arrangements devised to keep the block size small even when the number of experimental treatments was large. Very little attention was paid to certain characteristics of agricultural field trials that were not common to many other fields of experimentation.

3. Differences between field and laboratory experimentation

Notable among the differences between field experimentation and most laboratory experimentation is the planting early in the season of all the experimental plots and the recording of all the yields, or other records, in the harvest season. That is, all the experimental items marched abreast through time. The requirements of the growing season did not permit any other procedure and a suitably large experimental area made this procedure feasible. Thus all the experimental values are obtained at one time, or certainly after all the units have been started. Laboratory experimentation usually involves more or less elaborate and expensive equipment of which there may be only one complete assembly. The equipment is used repeatedly and the experimental values are obtained in a series. Usually an experimental value for a particular item is obtained before the next item is processed. The laboratory experimenter sees the results in a sequence and it is inevitable that the results will be examined one by one as they become available.

The agricultural experimenter has the winter to plan an elaborate and far-reaching program. If he makes any errors of omission in his program, another year will be needed before the omitted items can be studied. The laboratory experimenter may soon perceive a fruitless direction or a promising lead and may alter his program immediately (and usually does). Workers in these two areas are bound to see the problem of experimentation from quite different points of view.

There are other substantial differences between field and laboratory experiments. Suppose that r replications of t different tests are made, each replication being arranged in a compact block. The usual breakdown of the $rt - 1$ degrees of freedom is shown in table I.

TABLE I

Tests	$t - 1$
Blocks	$r - 1$
Residual	$(t - 1)(r - 1)$
Total	$rt - 1$

In agricultural field trials the sum of squares associated with blocks has little interest other than to show the extent to which grouping in blocks has reduced the experimental error. No permanent value attaches to the mean square for blocks because the area may be used next year for a different purpose.

In the laboratory the blocks may be time blocks, or refer to certain features of the equipment. For example, meter bars are compared in pairs in a long narrow chamber. It is essential that a uniform temperature be maintained throughout the chamber. The experimental design may schedule the position of the bars so that, after a certain number of comparisons, all of the bars will have been placed once in the left end of the chamber, and all of them placed once in the right end. Consequently the left end and the right end of the chamber constitute two blocks, each with a complete replication. The comparison of the bars has, by this means, been automatically corrected for any slight persistent difference in temperature between the two parts of the chamber.

The attitude of the experimenter in the physical sciences is revealed by his reaction to this situation. The reduction in error and the automatic correction for a deficiency in the equipment does not impress the experimenter nearly as much as the fact that a defect in the equipment has been detected and an estimate made of the magnitude of the error introduced by this defect. Almost certainly the experimenter will undertake to correct this defect in the equipment. The statistician will wisely go along with this project, even though the correction of the defect may appear to remove the need for experimental design. Second thought, on the part of the statistician, will lead him to realize that, when the defect is corrected, the orthogonal contrast, formerly used for the two ends, may be assigned to some other grouping of the meter bars, such as, for example, a choice of two positions for the optical assembly. A vista opens up to the statistician of a continuing opportunity to "clean up" the equipment by a series of investigations that can be fitted into the regular schedule of testing. The statistician is aided by the fact that the contrast used to detect the defect is often a simple dichotomy which marshals a considerable number of values behind each alternative.

At this point the statistician sees that the "blocks" which were of little or no interest in field trials are the important items in laboratory experimentation. They are important because the apparatus and the environment is not a one time event and anything learned about the equipment and the method of using it is of great value to the experimenter.

Laboratory work has still other problems for the statistician. Mention has already been made of the availability of laboratory measurements in a sequence. Suppose an experimenter wishes to investigate the property of a certain material as a function of temperature. A parallel situation in field trials would be the yield of a crop as a function of the amount of fertilizer applied. In the latter case the investigator must pick the several rates of application at the very start because he must plant all the plots at the same time. Randomization of the different rates of application over the planting area presents no difficulty. The laboratory worker will prefer to try a few selected temperatures, with the idea of interpolating additional temperatures or extending the range of temperatures after examining the results obtained with the first temperatures. Randomization with respect to time is now quite impossible and a new quirk is added to the formation of "blocks."

In some cases the number of replications that the experimenter has in mind is very limited. Sometimes cost is a factor. More often replication is not needed because laboratory measurements are usually much more precise than crop trials. The investigator may be willing to duplicate only a few of his tests so that this introduces further complications in devising appropriate experimental designs.

Finally the laboratory experimenter has a special advantage over the agricultural experimenter. The experimental error in a field trial depends on the locality, the year, that is, the weather, the area of the block, the crop, as well as other more or less unpredictable hazards. Sufficient replication must be incorporated to provide an estimate of the experimental error that is unique to this particular program. In the laboratory a new and complicated apparatus is rarely used immediately to obtain results for the record. Before undertaking any important investigation numerous measurements will have been made to check out the equipment and make sure that it is operating properly. In the course of these preliminary measurements the experimental error is usually pretty well established. In fact the experimental error is an indicator that the apparatus is functioning properly and that drifts and other troubles have been successfully overcome. Provided that provision is made to verify that the equipment remains in good operating condition, the experimental error as established by considerable preliminary work may be a better estimate than one based on the very few degrees of freedom available from a few duplications in a particular program.

The items discussed above by no means exhaust the list of ways in which laboratory experiments pose problems not provided for by classical experimental designs. Some progress has been made in devising designs appropriate for these problems but much remains to be done. The next section deals with some specific examples.

4. Some designs for comparative measurements

4.1. *Latin square applications.* Few things in statistics have caught the imagination as much as the Latin square so widely used in experimental work. The original application in field trials made possible the elimination of fertility differences between blocks running east and west as well as between blocks running north and south. Four fertilizers, A, B, C and D would be applied to 16 plots arranged in a square as shown in table II.

TABLE II

A	*B*	*C*	*D*
C	*A*	*D*	*B*
D	*C*	*B*	*A*
B	*D*	*A*	*C*

The blocks that contain all four fertilizers are not so compact but are unavoidably rather elongated. Nevertheless the similarity of adjacent or nearly adjacent plots made this arrangement very effective. Simple as the design is, early users were not always aware of the dependence upon the continuity of the blocks. One early worker reported, in a seminar, his 6 × 6 Latin square. No field was large enough so three rows of the square were located in one field and the other three in a field nearly a mile distant.

In the laboratory the rows and columns of the square may correspond to actual physical entities. A striking example of this was an early application in tobacco virus studies. Concentrations of virus selections could be compared by rubbing leaves with the solutions and counting the lesions that appeared three or four days later. Leaves from different plants varied in their susceptibility. And for any given lot of plants the susceptibility of the leaves depended on the position of the leaf on the plant. It was natural enough then to compare five virus solutions, A, B, C, D and E by making sure that a given solution was used once on each of five plants and once in each of the five available leaf positions. Table III shows the counts obtained and the analysis of variance for a test in which the five solutions were identical. The identification of columns with plants, and of rows with leaf positions, brings out the central idea of the Latin square.

TABLE III

LESION COUNTS ON LEAVES RUBBED WITH SAMPLES A, B, C, D AND E

Leaf position	Plant number					Leaf totals	Sample totals
	1	2	3	4	5		
1	18(A)	25(E)	36(D)	35(C)	23(B)	137	204(A)
2	11(B)	33(A)	29(E)	43(D)	22(C)	138	140(B)
3	5(C)	44(B)	49(A)	78(E)	28(D)	204	146(C)
4	24(D)	36(C)	27(B)	50(A)	15(E)	152	207(D)
5	5(E)	76(D)	48(C)	35(B)	54(A)	218	152(E)
Totals	63	214	189	241	142	849	849

ANALYSIS OF VARIANCE

Source of Variance	Degrees Freedom	Sum of Squares	Mean Square
Plants	4	3914	979
Leaf Positions	4	1179	295
Samples	4	865	216
Error	12	2395	200
Total	24	8353	

Other than agricultural workers have attempted to use the Latin square arrangement for purposes for which it was never intended. A chemist may wish to ascertain the yield of a chemical process using different temperatures, pressures and catalysts. Quite a number have yielded to the temptation to draw up a $k \times k$ "design" using the k rows for k different temperatures, the k columns for k different pressures and the letters A through K for the catalysts. Here the thought is that the k^2 tests specified by the square can be used instead of the k^3 required for the complete factorial. This is not at all a good fractional factorial; nevertheless an applied statistical text for engineers published in 1958 suggested exactly this approach. The scheme fails because the effect of a change in the pressure will almost certainly depend upon the temperature and the catalyst; that is, the effects are not expected to be additive. The chemical investigation should, in fact, be directed to ascertaining the extent of nonadditivity and taking advantage of this property of the system.

A more successful application concerned some experiments performed to test the identity of several specimens of a device used to set up a reference temperature. The device, or cell, consisted of a sealed glass container with about a pound of a highly purified chemical. The cells were warmed until the crystalline chemical just melted and were allowed to cool very slowly in an insulated box. Thermometer wells extending into the chemical permitted the introduction of platinum resistance thermometers to measure the temperature of the triple point which was closely maintained for a day or so. It was necessary to compare the cells and a number of thermometers were available. Thermal equilibrium between thermometer and cell contents was slow so that only one thermometer could be used on each melting. Readings spread over a number of days would be vulnerable to possible changes in the measurements of the electrical resistance of the thermometers.

The cells were made as nearly alike as possible and the thermometers carefully calibrated. On any one day measurements could be made on several thermometer-cell combinations. Disagreement among the temperatures recorded might arise either through inequality of the cell temperatures, differences among the thermometers, or both. The very narrow range of values encountered made it obvious that if there was a difference in the temperatures set by two cells, the difference would be independent of the thermometer used to measure the two cells, so long as both cells were read with the same thermometer. An error in thermometer calibration would drop out when the difference was taken. Thermometers, cells and measuring apparatus are purposely chosen to be as nearly identical as possible. The departure by any of these factors from the average value will be so slight that the effect on the system may be considered to be directly proportional to the departure. Additivity is assured in contrast to the "interaction" effects that are characteristic of chemical reactions.

The Latin square arrangement suggested a precise manner of making the comparisons. The pairing of the cells and thermometers is shown in table IV along with the data. The letters refer to the four days corresponding to the four replications. All readings agreed to two decimal places so only the third and fourth decimal places in degrees centigrade are shown.

Ordinarily rows and columns are the blocks and the letters identify the items tested. In this scheme rows and columns are the experimental factors and the letters are blocks, nevertheless the essential character of the Latin square is nearly perfectly realized in this experiment. Cells and thermometers are physical objects which are highly permanent and maintain their characteristics over lengthy periods of time. The four readings made on a day were made within a few minutes so the local circumstances that might influence a reading would apply to all four readings. The analysis of the data shows that there is some evidence of a day to day effect upon the readings. The thermometers show better agreement than do the cells; a not unexpected result because cell 2 contained a

TABLE IV

PAIRING OF CELLS AND THERMOMETERS. LETTERS REFER TO DAYS. NUMBERS GIVE CODED TEMPERATURE READINGS

	Thermometer				
Cell	I	II	III	IV	Totals
1	*A* 36	*B* 38	*C* 36	*D* 30	140
2	*C* 17	*D* 18	*A* 26	*B* 17	78
3	*B* 30	*C* 39	*D* 41	*A* 34	144
4	*D* 30	*A* 45	*B* 38	*C* 33	146
Totals	113	140	141	114	508

ANALYSIS OF VARIANCE

Source of Variance	Degrees Freedom	Sum of Squares	Mean Square
Between thermometers	3	182.5	60.8
Between cells	3	805.0	268.3
Between days	3	70.0	23.3
Residue	6	43.5	7.25
Total	15	1101.00	

less purified lot of the chemical. In subsequent work there was no hesitation in using all the available orthogonal contrasts for experimental comparisons.

4.2. *Linear trend design.* Corrections to the readings obtained with a thermometer may be obtained by taking readings when the thermometer is placed in an environment of known temperature. A standard thermometer may be used to determine the temperature of the bath. Actual calibrations are not made in quite such a simple manner. Instead of attempting to regulate the bath temperature to as nearly constant a value as possible, the heating element is adjusted so that a very gradual rise in bath temperature takes place. The rate is of the order of a few thousandths of a degree per minute and permits reading the level of the mercury on a "rising meniscus." A number of thermometers, including the reference, are arranged in a device that can be rotated to bring each thermometer stem in turn before an optical device. The thermometers are first read in order and then in reverse order. The readings are called off to an assistant. A set of readings is given in table V. On the assumption of a steady rise in bath temperature and of equal intervals between reading it is taken that the *average* of the two bath temperatures associated with a thermometer is the same for all thermometers. Comparisons among the thermometers are therefore free of the change in bath temperature. A special feature of this scheme is that the reader is able to obtain two independent readings for each thermometer without being influenced by the operator's memory of the first reading. The use of this

TABLE V

THERMOMETER READINGS (MINUS 18° C) IN A BATH WITH RISING TEMPERATURE

Reading order	Thermometer identification							
	A	*B*	*C*	*D*	*E*	*F*	*G*	*H*
$A \to H$	.107	.104	.074	.072	.083	.107	.098	.091
$H \to A$	.121	.115	.084	.080	.089	.113	.103	.094

design is traditional in laboratories calibrating thermometers and is a stimulating example of a good design developed without statistical assistance.

The data also provide a means of testing the combined assumption of steady temperature rise and equally spaced readings. There is one time interval between the two readings for the last thermometer, three time intervals between the two readings for the next to last thermometer, and, if there are k thermometers, there are $2k - 1$ time intervals between the two readings for the first thermometer. The differences between the two readings for a thermometer may be plotted against the number of time intervals between the readings. If the assumptions are met and the readings made without error all the points should lie on a straight line through the origin whose slope gives the rate of temperature rise for the bath. When a line is fitted the deviations from the line provide an estimate of the error of the thermometer comparisons. It is interesting to contrast the above clean-cut procedure with some alternative such as making a number of paired comparisons in a bath regulated for a constant temperature. The pairs would afford two immediately adjacent readings but the adjustments to the averages would be more tedious to make.

4.3. *Designs concealed within designs.* In planning a study of the effect of temperature on the transverse breaking strength of synthetic sapphire rods the question was raised as to whether specimens from different rods varied more than specimens taken from the same rod. This seemed a possible complication and the experimental design was based on using each rod as a block. Each rod was long enough to furnish three test specimens. The investigator suggested that seven different temperatures would suffice to define the strength-temperature curve. A standard balanced incomplete block design seemed exactly suited to this problem. The proposed arrangement was as shown in table VI, where the letters denote the test temperatures.

TABLE VI

Position in rod	Rod number						
	1	2	3	4	5	6	7
Top	*A*	*B*	*C*	*D*	*E*	*F*	*G*
Middle	*B*	*C*	*D*	*E*	*F*	*G*	*A*
Bottom	*D*	*E*	*F*	*G*	*A*	*B*	*C*

The investigator soon perceived the fact that the design made sure that any two temperatures could always be compared using specimens from the same rod. He protested, however, that he did not wish to commit himself to all the test temperatures in advance. He would prefer to try some temperatures and, after seeing these results, either interpolate additional temperatures or extend the range of temperatures. The experimenter wanted to insure a spacing of the temperatures that would provide points where the temperature effect is most pronounced or possibly locate the temperature of maximum strength. The question was put whether the preliminary study could also have the within rod comparisons without disturbing the final program. Scrutiny disclosed that, if the letters A, C, D, and E were assigned to a study with four temperatures the pattern in table VII emerged.

TABLE VII

Position in rod	Rod number						
	1	2	3	4	5	6	7
Top	*A*		*C*	*D*	*E*		
Middle		*C*	*D*	*E*			*A*
Bottom	*D*	*E*			*A*		*C*

This is also a balanced incomplete design for blocks of size two. The six pairings of the four letters appear on six of the seven rods. The preliminary data, therefore, could also be used to obtain a preliminary temperature curve unobscured by differences among the rods. This improvement would permit a more satisfactory selection of the remaining three temperatures which would be assigned to the letters B, F, and G as in the original schedule.

This property of a design concealed within a design is a general one and appears to be true for many partially balanced designs as well. It is interesting to observe that the field trial applications apparently did not prompt any search for this property. It was only when a use for such a property appeared in laboratory applications that any attention was given to it.

5. The problem of absolute measurements

The preceding section on comparative measurements used some examples to reveal at once the diversity of experimental situations encountered in laboratories and the flexibility of experimental design to fit these situations. The primary objective of these designs is to improve the precision of comparative measurements. Good precision is indispensable but does not guarantee correct values for the physical constants measured in the laboratories. The undoubted existence of errors that remain constant throughout a series of repeat measurements, and are therefore not revealed by the random errors associated with precision, has long appeared to be a stumbling block in the application of statistics to absolute measurements. These errors that remain constant arise from various sources. An error may be inherent in the method of measurements or in the theory on which the method is based. Other errors are peculiar to various parts of the particular set of equipment brought together for making the measurement. Some may even be associated with the locality, or the time of day. Experimenters carefully go over every one of the possible sources mentioned above and take whatever precautions their scientific knowledge and experience suggests. At the completion of the work an effort is usually made to evaluate every relevant item that has remained constant during the study and estimate its likely maximum contribution to the error in the final result.

There may be in the apparatus a tube or an aperture whose diameter, although small, is chosen arbitrarily by the designer of the equipment. The only requirement is that the diameter must be known and certainly a very careful determination will be made of the diameter. The contribution of the maximum error in the diameter to the error in the final result will be calculated. In any complex piece of equipment there are several arbitrary choices such as dimensions, resistances, and so forth. Usually these dimensions, resistances and similar quantities must be known. The whole principle of the method for establishing the physical constant rests upon the concept that the final result should be independent of these arbitrary choices if the values chosen are known to a sufficient degree of accuracy.

The catch in practice is that so much care is expended on each item that goes into the final assembly that rarely is more than one arbitrary choice exercised. The apparatus will therefore employ a tube of one arbitrarily chosen diameter and no other choice provided. One chamber, whose volume must be known, will be prepared and no other will be available. It would certainly not double the cost of the equipment to provide a second choice for certain critical items. Usually the second item could be made at a considerable saving over the first item and not all items need duplication. This additional equipment cost is probably not the main objection to constructing the duplicate pieces. The experimenter quickly sees that if only seven items are duplicated there will be 2^7 or 128 possible assemblies that can be constructed. Perhaps putting together each assembly is time consuming and each individual, very carefully made, measurement may also be time consuming. There is no possibility of undertaking 128 careful measurements so the whole notion is promptly dismissed.

It appears that experimenters are in general unaware of the statistical resources for meeting this problem and that statisticians are even more unaware of the uniquely ideal way such a problem meets their statistical assumptions. Suppose all 128 assemblies were constructed and readings obtained with each. By the experimenter's own authority all 128 results should be nearly identical. Here is a homogeneity trial never approached in ordinary circumstances. Additivity of the very tiny effects associated with the change from one component to its alternative choice is a virtual certainty. No transformations of data will be needed to cope with a range of values. Homogeneity of variance is assured. Indeed the situation is a statistical paradise.

Once this Eden is entered there is absolutely no need to even consider putting together 128 different assemblies. The most highly fractionated factorial will serve without arousing any of the usual misgivings. Weighing designs are perfectly suited to this problem. Often the number of assemblies may be held to just one more than the number of items duplicated—eight assemblies, if there are seven items duplicated.

A particular set of eight different assemblies is shown in table VIII. The seven items are designated by letters, the capital letter is used to label one choice and the lower case the alternative choice for that item. For example, the item might be a chamber and the two chambers would be made somewhat different in volume.

TABLE VIII

SCHEDULE FOR CONSTRUCTING EIGHT ASSEMBLIES WHEN TWO CHOICES, CAPITAL AND LOWER CASE, ARE AVAILABLE FOR EACH OF SEVEN COMPONENTS. ASSEMBLY NUMBER

1	2	3	4	5	6	7	8
A	*A*	*A*	*A*	*a*	*a*	*a*	*a*
B	*B*	*b*	*b*	*B*	*B*	*b*	*b*
C	*c*	*C*	*c*	*C*	*c*	*C*	*c*
D	*D*	*d*	*d*	*d*	*d*	*D*	*D*
E	*e*	*E*	*e*	*e*	*E*	*e*	*E*
F	*f*	*f*	*F*	*F*	*f*	*f*	*F*
G	*g*	*g*	*G*	*g*	*G*	*G*	*g*

Any two assemblies have three choices in common so that only four items need to be changed at a time. If some items are difficult to interchange a judicious choice of the order of constructing the assemblies will keep down the work. Thus if items *A* and *a* and *D* and *d* are tedious to interchange the order shown in the table would minimize the changes for these items.

Important advantages flow from this program. First it is presumed that two or more readings will be made with each assembly so that a good estimate of the precision will be available. Second the schedule is extremely efficient in detecting the effect of replacing an item with its alternative choice. The effect of replacing *A* by *a* is revealed by comparing the averages for the results obtained with each item. If only duplicate readings are available for each assembly there will still be eight readings for each average. This makes for a sensitive test for the effect of the substitution and the proper error to apply to judge any difference between these averages is the precision error as revealed by the duplicate measurements. Results with the alternative choice provide a convincing check on the calibration of the component parts.

The schedule provides that whenever the four assemblies with a particular capital letter are matched against the other four assemblies with the corresponding lower case letter all other items are neatly balanced off against each other. Each squad of four assemblies always contains two capital and two lower case for each of the other six letters. Thus the effects of the other choices cancel out. The consequence is that the total weight of all the data is marshalled to detect the effect of a change in each of the seven items. These effects would reveal any inadequacies in the measured values used for dimensions, volumes, and so forth, of these component items. By this means some light would be thrown on the uncertainty in the value obtained for the physical constant which is the primary objective. Certainly the gross variation among the results associated with the different assemblies would provide a *minimum* measure of the uncertainty in the final result. Any defect in the principle of the method will not be revealed and can only be detected by measuring the quantity by a different method.

The minimum number of assemblies required to evaluate the constants is equal to $N + 1 - k$, where N is the total number of parts and k is the number of items for which choices are available. If there are but two choices for each of k items the minimum number of assemblies would be $k + 1$ but this minimum cannot always be attained in an efficient manner. The four assemblies for $k = 3$ are ABC, Abc, aBc and abC. The eight assemblies for $k = 7$ have already been given. Eight assemblies provide efficient comparisons when k is 4, 5 or 6. They may be obtained from the assemblies given for $k = 7$ by ignoring the choices for those items for which no choice is available. A paper by Plackett and Burman [5] takes up this problem.

If there are three choices available for each of four items (A, B, C and D) a 3×3 Graeco-Latin square specifies the nine required assemblies. Let lower case letters with subscripts denote the choices.

a_1	b_1	c_1	d_1	a_1	b_2	c_2	d_2	a_1	b_3	c_3	d_3
a_2	b_1	c_2	d_3	a_2	b_2	c_3	d_1	a_2	b_3	c_1	d_2
a_3	b_1	c_3	d_2	a_3	b_2	c_1	d_3	a_3	b_3	c_2	d_1

When some of the items are available in just two choices, and other items are available in three or more choices the selection of combinations tests the ingenuity. The goal is to specify the assemblies that allow for the most efficient evaluation of the differential effects associated with the choices. A simple example will serve to illustrate the point.

Consider item A available as a_1 or a_2, item B available as b_1, b_2 or b_3, and item C available as c_1, c_2 or c_3. Table IX shows one selection of six from the possible eighteen assemblies. The letters u, v, w, x, y and z in parentheses denote the observed measurement for each assembly.

The estimate of the *difference* in the effects of choices a_1 and a_2 is given by the expression $(u + v + w - x - y - z)/3$. The quantities u, v and w, taken together, contain one effect from each of the choices b_1, b_2, b_3, c_1, c_2, and c_3, and this is also true for the three measurements x, y and z. Consequently the effects of

TABLE IX

	b_1	b_2	b_3
c_1	$a_1(u)$	$a_2(y)$	
c_2	$a_2(x)$		$a_1(w)$
c_3		$a_1(v)$	$a_2(z)$

the B and C choices cancel out. The estimate of the difference in the effects of b_1 and b_2 is given by $(2u - w - v + x - 2y + z)/3$. In this expression the effects of choices other than b_1 and b_2 all cancel out. Similar expressions apply for the difference between any two choices among the three alternatives for an item.

It is interesting to note that these experimental designs may be regarded as an extension of the weighing design problem using a spring scale [3], [4], [5]. Given N weights which are classified into k types. Let the number of weights of each type be $n_1, n_2, \cdots, n_k$, where $\sum n = N$. The spring scale is always loaded with exactly k weights, there being one weight of each type. The problem is to specify the several selections so that efficient estimates of the differential effects of the weights will be obtained. It is desirable that the number of selections be as close to the minimum, $N + 1 - k$, as possible.

The analogy with the spring scale is instructive. In these particular physical measurements, all the assemblies should have the same total "weight" so that an error in the calibration of the scale, if present, will appear in all the results. This is the error of the method. If some particular weight is "heavy" with respect to the alternatives of its type, the proper groupings of the weighings will provide a sensitive test for this shortcoming. The real saving comes from the simultaneous evaluation of several features of the equipment with but little more work than would be required to study equally well just one feature.

The problem of studying the experimental equipment is discussed by Dorsey [1] on page 11 of a 110 page paper on the velocity of light. Dorsey remarks: "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 the result, and some that do not, in every case pushing the change well beyond any that seems at all likely to occur accidentally." This remark is included (page 107) in a brief paper by Dorsey and Eisenhart [2] that is based upon excerpts from the lengthy Dorsey article. The point of the present discussion rests on the phrase "one by one" in the above quotation. A planned approach, changing several items at once, is a more efficient way to determine the effects of changing these items of the apparatus than the traditional practice of changing just one thing at a time.

Different principles of measurement usually give results that disagree more than would be anticipated considering the precision as revealed by repeat measurements made with just one assembly. Perhaps the discrepancy would be more understandable if the errors were based on the variation shown by results obtained with *different* assemblies. Even the same method, in the hands of another investigator, usually gives a result that differs from an earlier result by more than the indicated precisions would permit. Obviously the later worker changed *all* the components in the equipment. Any attempt to single out a particular component as responsible for the disagreement falls in the realm of speculation. Each laboratory has the opportunity and the responsibility to ascertain the errors that are associated with its equipment as used in the final measurement. In this task statistics can play a cooperative role, both in the apt selection of the assemblies, and in the evaluation of the measurement errors as revealed by these different assemblies.

REFERENCES

[1] N. E. Dorsey, "The velocity of light," *Trans. Amer. Philos. Soc.*, Vol. 34 (1944), pp. 1–110.

[2] N. E. Dorsey and C. Eisenhart, "On absolute measurements," *Sci. Monthly*, Vol. 77 (1953), pp. 103–109.

[3] K. Krishen, "On the design of experiments for weighing and making other types of measurements," *Ann. Math. Statist.*, Vol. 16 (1945), pp. 294–300.

[4] A. M. Mood, "On Hotelling's weighing problem," *Ann. Math. Statist.*, Vol. 17 (1946), pp. 432–446.

[5] R. L. Plackett and J. P. Burman, "The design of optimum multifactorial experiments," *Biometrika*, Vol. 33 (1946), pp. 305–325.