PSYCHOMETRIKA—VOL. 7, NO. 2
JUNE, 1942

REGRESSION FALLACIES IN THE MATCHED GROUPS EXPERIMENT

ROBERT L. THORNDIKE

TEACHERS COLLEGE, COLUMBIA UNIVERSITY

This paper is concerned particularly with certain regression effects which appear whenever matched groups are drawn from populations which differ with regard to the characteristics being studied. It is shown that regression will produce systematic differences between these groups on measures other than those upon which they were specifically matched. The size and direction of these differences depends upon the differences between the parent populations both in the matching and in the experimental variables and upon the correlation between the matching and experimental variables. Formulas are presented for estimating the expected regression effect. Several alternative procedures are suggested for avoiding the erroneous conclusions which the regression effect is likely to suggest.

It is not the purpose of this paper to present any scintillating new statistical ideas. What is said here should be in the nature of a reminder of old truths, rather than a message of startling novelty. The aim is to restate and clarify for research workers in education and psychology some of the errors into which they may lapse when they use the experimental pattern of matched groups. The matched groups experiment is sufficiently prevalent in education and psychology and the use of it sufficiently uncritical, to make it worth our while to enquire into certain sources of error.

The fallacies with which we are here concerned may arise whenever the measure or measures by means of which the groups were matched have less than a perfect correlation with the measure of the experimental variable which is being studied. A more limited example of this is found in the less than perfect correlation between a test and a subsequent retest with the same instrument. However, our argument is more general than this, and holds whenever groups are matched upon one measure or group of measures and then studied with regard to their performance on other measures which do not have a perfect correlation with the matching variable. Since this is universally true in the matched-groups experiment, the points to be raised here are of quite general application.

Whenever the correlation between two measures is less than unity, part of the variance of scores in each measure is independent of variance in the other measure. We can conceive of performance in

each test as made up of two parts—a part common to the two tests and a part unique to the particular test. Those individuals who receive high scores on test X do so in part because they possess large amounts of whatever is common to X and Y, in part because they possess large amounts of whatever enters into X score but not into Y score. (We have no concern, for the present, as to what this specific element is, i.e., whether it is "specific factor" or "error of measurement.") Since the specific element in test X and the specific element in test Y are unrelated, those individuals who possess large amounts of the X specific will, as a group, possess just average amounts of the Y specific. The total group of those found to deviate from the average in test X in a certain amount and direction will also deviate from the mean in test Y in the same direction. They will not, however, deviate in the same amount. Whereas in test X both the specific and the common factor combined to produce the deviations in those who were *selected because of their deviant X score,* in test Y only the common factor is at work. The result is that those selected as falling H standard deviations above the X mean will fall Hr_{xy} standard deviations above the Y mean, and vice versa. The Y scores will regress towards the mean by an amount which is a direct function of the size of the X deviation and an inverse function of the correlation between X and Y.

It is important to point out that the regression of scores upon a second test is toward the *mean of the population from which the cases were selected,* and which they truly represent. If a ninth- and a twelfth-grade population are tested with an intelligence test and a reading test, it will be found that the twelfth graders above the *twelfth grade mean* in intelligence will tend to drop down toward the *twelfth grade mean* in reading score (when scores are expressed in standard deviation units). However, twelfth-graders who are above the total group mean but below the twelfth-grade mean will tend to regress *up* toward the twelfth grade mean—not *down* toward the combined population mean. Similarly, the ninth-graders will regress toward the ninth-grade mean. By the same token, if we study two groups of 10-year-olds, one group drawn from a private school catering to executive and professional families and the other from an orphanage or an institution for retarded children, we must expect the deviant individuals in each group to regress toward the mean of population from which they come, and not toward some hypothetical average of 10-year-olds in general.

This point can be illustrated by an empirical comparison of data from a group of 8- and 9-year-olds and a group of 12- and 13-year-

olds. For each child in these two groups there were available an M.A. on Form L of the Revised Stanford Binet and a score on a brief 15-word vocabulary test (selected from this same material). Data on the two groups for the two tests were as follows:

	8 & 9	*12 & 13*
N	185	138
Mean M.A.	9.42 yrs.	12.65 yrs.
Mean Vocabulary Score	3.29 words	5.54 words
Intercorrelation	.77	.82

From each group were picked out all the cases with M.A. falling between 10-0 and 11-11. This is a group above the average M.A. of the younger group, and below the average M.A. of the older group. The average vocabulary scores for the 8 and 9 and for the 12 and 13 groups are 4.09 and 4.41 respectively. If we reverse the procedure, and determine the average M.A. in each group for those having vocabulary scores of 4 and 5, we get 10.59 and 11.46 respectively. In each case, we see that the groups selected as matched on one variable are not matched on the other. The older children surpass the younger in each case. This is because the older group regresses toward a higher population mean score than the younger group.

In studies using matched groups, we can recognize three patterns.

In the first pattern, two or more matched groups are assembled within a single population or in different sub-populations which may reasonably be thought to be, in all essential features, fractions of the same total population. For example, if our interest is to compare the effectiveness of three different types of materials in developing rapid reading, we may select from the students in a large class three groups that are equated in terms of initial speed of reading, and then try out the three variations in practice material, one on each of the three groups. Or if one teacher instructs three different class groups, all chosen in the same way from the same student body, we may make up one of our matched groups in each of the classes. The crucial point in this example is that the groups are selected from what are, to the best of our knowledge, equivalent populations.

In this type of situation, regression should affect each of our samples in the same way. Since the samples are all taken from the same population, we may expect them all to regress toward the same population mean, and since they have the same distribution of scores on the matching variables, the expected direction and amount of regression is the same. There is no reason why there should be any *systematic* tendency for the regression from matching to experimental

variable to affect one group differently from the other. Of course, chance fluctuations may be expected to disturb in some degree the exactness of the matching, but this effect should be a random one.

The problem then reduces to determining the appropriate standard error for evaluating the obtained difference upon the retest. The appropriate formula must take account of the reduction in chance differences between the means on the experimental variable due to the matching. This problem has recently been reviewed in some detail by McNemar.* If the matched groups have been assembled by matching pairs of individuals, the appropriate formula for the standard error of a difference upon any measure other than that upon which they were matched becomes

$$\sigma_D = \sqrt{\sigma_{\bar{x}_1}^2 + \sigma_{\bar{x}_2}^2 - 2 r_{12} \sigma_{\bar{x}_1} \sigma_{\bar{x}_2}}, \qquad (1)$$

where r_{12} is the correlation between members of a pair in the new measure and $\sigma_{\bar{x}_1}$ and $\sigma_{\bar{x}_2}$ are the standard errors of the two group means for the new measure. If the matching of the two groups is in terms of distribution of scores for the groups as a whole rather than individual pairing, an acceptable formula for the standard error of the difference is

$$\sigma_D = \sqrt{(\sigma_{\bar{x}_1}^2 + \sigma_{\bar{x}_2}^2)(1 - r^2_{xy})}, \qquad (2)$$

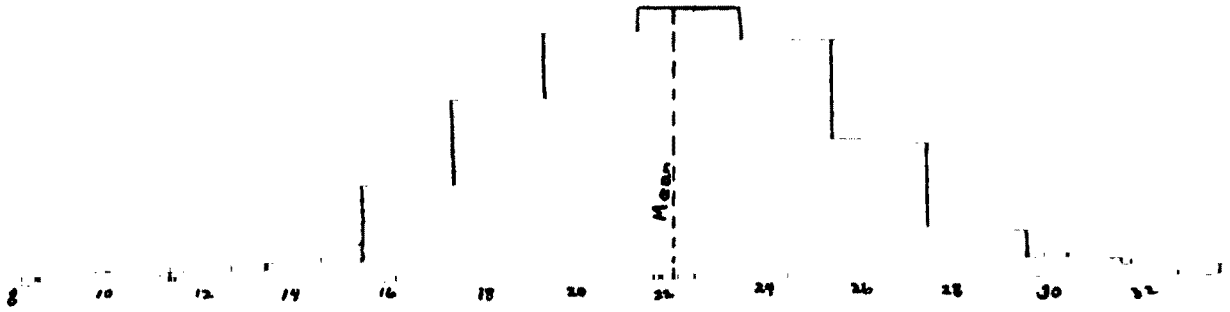
in which r_{xy} is the correlation between the test upon which groups were matched (y) and the experimental test (x).

The second pattern is that in which we are faced with two or more discrete categories of individuals. The categories are differentiated by some characteristic external to but perhaps related to either the measure in terms of which the groups are being matched or the experimental variable or both. We might, for example, study a group of men and a group of women who were matched in performance upon a test of strength of grip. We might plan to work with matched groups of students who do and students who do not take Latin, the basis of matching being score on a test of English vocabulary. Another possibility might be to work with a group of orphanage children and a group of private school children, matched for intelligence test score. In each of these cases we are dealing with matched groups selected from two distinct populations—populations which probably have quite different means upon the tests in terms of which the matched samples were chosen.

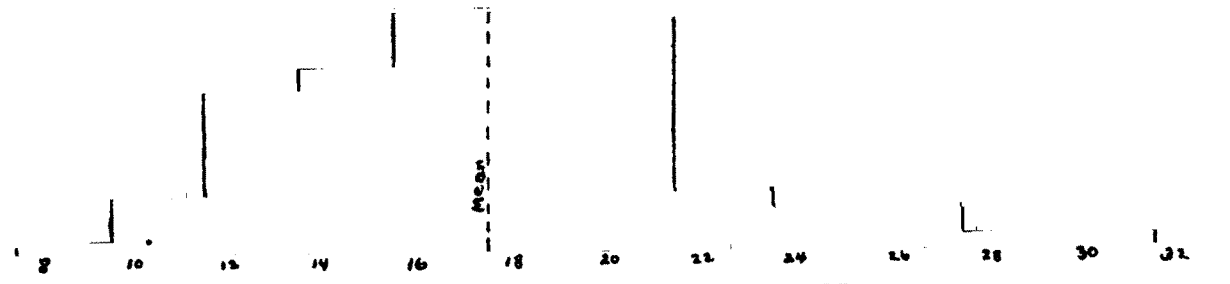
In order to get a matched group when the two populations have

* McNemar, Quinn. Sampling in psychological research. *Psychol. Bull.*, 1940, 37, 331-365.

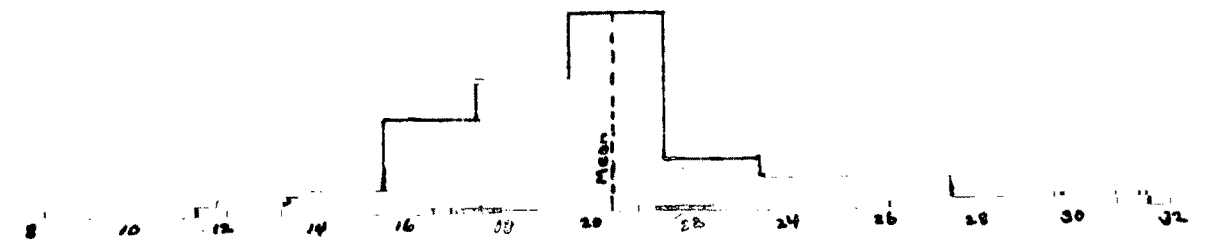
FIGURE 1
Regression in Discrete Populations



"Men"—Total Population—Mean = 22.2



"Women"—Total Population—Mean = 17.5

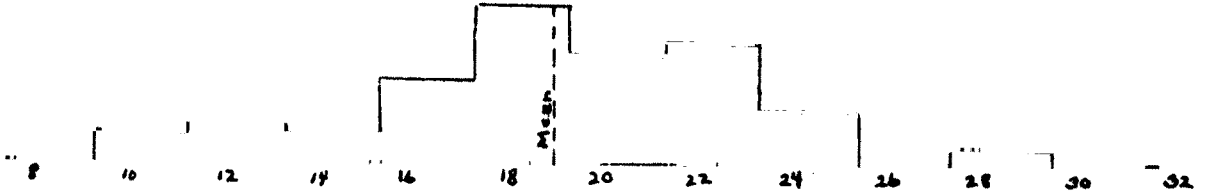


Matched Group—"Men" and "Women"—Mean = 20.3

FIGURE 1

Regression in Discrete Populations

Retest Scores—Matched Group of "Men"—Mean = 21.0



Retest Scores—Matched Group of "Women"—Mean = 19.3

different mean values, we must take individuals who fall relatively high in one population and match them with individuals who fall relatively low in the other. Since the individuals in each group *will regress toward their own population mean*, the regression in the two groups will be different. Upon another test, our groups will no longer be matched.

We can illustrate this with some artificial data constructed from dice throws. Let us suppose that these data represent score upon two strength tests. Suppose that we use seven dice, numbers 1, 2, and 3 representing common ability in the two tests, numbers 4 and 5 represent the factor specific to the first test, and numbers 6 and 7 representing the factor specific to the second test. Score for the first test is the number of spots showing on dice 1, 2, 3, 4 and 5; score on the second test is the number of spots showing on dice 1, 2, 3, 6 and 7. In this way two scores were obtained for 132 "women." Scores for a second population of 132 "men" were gotten in the same way except that the constant amount 5 was added to the number of spots showing to give each score. The distributions of scores for men and for women on the first test are shown as the first two histograms in Figure 1. The theoretical difference between the means of these two populations is 5; empirically, it comes out to be 4.7.

Now a sub-group is made up in each population by selecting cases which can be individually matched with cases in the other population on the basis of score on the first test. This gives 64 matched pairs. The mean is, in each case, 20.3. Let us examine the second test scores of the 64 "men" and 64 "women" in these matched groups. From the last two histograms of Figure 1, we see that the "men" have regressed up so that their mean second test score is 21.0; the "women" have regressed down to a mean score of 19.2. On the second test the two matched groups differ in mean score by 1.8, about 40% of the difference between the means of their parent populations.

If we know the means and standard deviations of the two populations from which our matched samples are drawn, both upon the matching test and upon the experimental test, and the correlations between the two tests in each population, we can determine the amount of difference to be expected on the second test between the means of samples from each group matched on the first test. The formula is

$$
\begin{aligned}
{}_A\bar{\bar{Y}} - {}_B\bar{\bar{Y}} = (\bar{X} - {}_AM_x)\left({}_Ar_{xy}\frac{{}_AS_y}{{}_AS_x} - {}_Br_{xy}\frac{{}_BS_y}{{}_BS_x}\right) \\
- {}_Br_{xy}\frac{{}_BS_y}{{}_BS_x}D + ({}_AM_y - {}_BM_y)\,,
\end{aligned}
\tag{3}
$$

and the derivation of this is shown in the mathematical note below.

MATHEMATICAL NOTE

Given: Two populations, A and B, having different population means in a measure X. The population means are designated ${}_AM_x$ and ${}_BM_x$, respectively. These two means differ by the amount D. The standard deviations in the two populations are ${}_AS_x$ and ${}_BS_x$.

A sample has been selected from each of the populations in such a way that the mean X score in each sample is the same. The mean score in these samples is designated $\bar{X}$.

Required: To determine the expected difference between the means of these two samples upon some other measure Y, when the population means and standard deviations for Y are ${}_AM_y$, ${}_BM_y$, ${}_AS_y$, ${}_BS_y$, respectively, and the coefficients of correlation between X and Y in the two populations are ${}_Ar_{xy}$ and ${}_Br_{xy}$.

Derivation: The X score of an individual from population A may be designated ${}_AX_i$. For this individual, the predicted score on test Y is

$$ {}_A\bar{Y}_i = {}_Ar_{xy}\frac{{}_AS_y}{{}_AS_x}({}_AX_i - {}_AM_x) + {}_AM_y . $$

If we sum over the N_A cases in the matched sample from population A, we get, as an unbiased estimate of the mean of the Y scores,

$$ {}_A\bar{\bar{Y}} = \frac{\sum_{i=1}^{N_A} {}_AY_i}{N_A} = {}_Ar_{xy}\frac{{}_AS_y}{{}_AS_x}\left(\frac{\sum_{i=1}^{N_A} {}_AX_i}{N_A} - {}_AM_x\right) + {}_AM_y $$

$$ = {}_Ar_{xy}\frac{{}_AS_y}{{}_AS_x}(\bar{X} - {}_AM_x) + {}_AM_y . $$

Similarly, for the matched sample from the population B,

$$ {}_B\bar{\bar{Y}} = {}_Br_{xy}\frac{{}_BS_y}{{}_BS_x}(\bar{X} - {}_BM_x) + {}_BM_y $$

$$ = {}_Br_{xy}\frac{{}_BS_y}{{}_BS_x}(\bar{X} - {}_AM_x + D) + {}_BM_y . $$

If, now, we subtract, we get

$$ {}_A\bar{\bar{Y}} - {}_B\bar{\bar{Y}} = (\bar{X} - {}_AM_x)\left({}_Ar_{xy}\frac{{}_AS_y}{{}_AS_x} - {}_Br_{xy}\frac{{}_BS_y}{{}_BS_x}\right) - {}_Br_{xy}\frac{{}_BS_y}{{}_BS_x}D + ({}_AM_y - {}_BM_y). $$

When

$$ {}_Ar_{xy} = {}_Br_{xy} = r_{xy} , $$

$$ {}_AS_x = {}_BS_x = {}_AS_y = {}_BS_y , $$

$$ {}_AM_y - {}_BM_y = D , $$

the foregoing expression reduces to the very simple expression

$$_A\bar{\bar{Y}} - {}_B\bar{\bar{Y}} = (1 - r_{xy})D\,.$$

It is possible, following out just the same line of analysis, to determine the difference to be expected in a third variable Z when groups have been set up matched in terms of two variables, X and Y, and the procedure can be generalized to any number of variables. The resulting formula grows out of the regression equation for predicting Z from X and Y, and involves the partial regression coefficients. The formula becomes

$$\begin{aligned} {}_A\bar{\bar{Z}} - {}_B\bar{\bar{Z}} = {} & ({}_A\bar{X} - {}_AM_x)\,({}_Ab_{zx\cdot y} - {}_Bb_{zx\cdot y}) - {}_Bb_{zx\cdot y}D_x \\ & + ({}_A\bar{Y} - {}_AM_y)\,({}_Ab_{zy\cdot x} - {}_Bb_{zy\cdot x}) - {}_Bb_{zy\cdot x}D_y + ({}_AM_z - {}_BM_z) \end{aligned}$$

If ${}_Ab_{zx\cdot y} = {}_Bb_{zx\cdot y}$ and ${}_Ab_{zy\cdot x} = {}_Bb_{zy\cdot x}$, this reduces to

$$_A\bar{\bar{Z}} - {}_B\bar{\bar{Z}} = ({}_AM_z - {}_BM_z) - (b_{zx\cdot y}D_x + b_{zy\cdot x}D_y).$$

If we are dealing with a test and retest with the same instrument, and if we can assume that (a) the standard deviation for both test and retest, (b) the test-retest correlation, and (c) any gain in mean score from test to retest, are the same for both populations, then the difference to be expected between the means of the two matched groups on the retest reduces to the very simple expression

$$_A\bar{\bar{Y}} - {}_B\bar{\bar{Y}} = (1 - r_{xy})D\,, \qquad (4)$$

where r_{xy} is the test-retest correlation and D is the difference in score between the means of the two populations from which our matched groups were drawn.

Equation (4) presents the simplest possible picture of the effect of regression, uncomplicated by any differences in variability, relation of first to second test, or proneness to gain in the two populations. This simplified picture will be only an approximation in most actual cases, and it will ordinarily be difficult to tell just how reasonable the assumptions involved in this formula are for our data.

It is perfectly possible to develop formulas of the type given in equation (3) for the expected difference when the matching is based upon two or more variables. The formula in the case of two matching variables is given in the mathematical note. The formulas are straightforward but unwieldy. In practice, the chief difficulty which would be encountered would be that some of the statistics with regard to the populations from which the matched samples were selected would be unknown. Excepting as it is possible to compute or estimate these, it is, of course, impossible to solve the equation which provides an indication of the expected difference upon the experimental test.

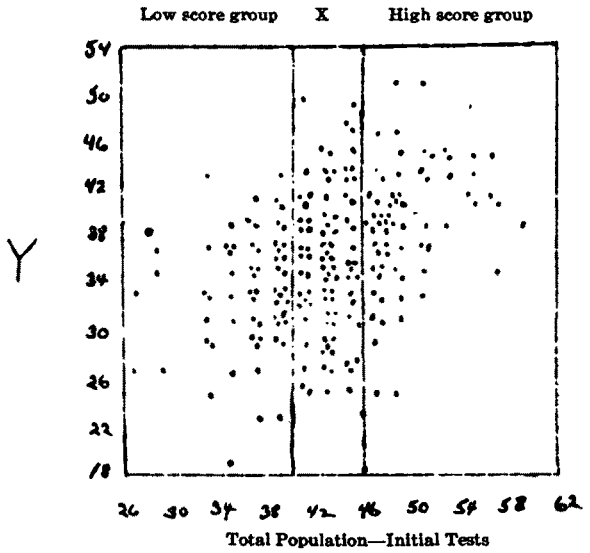
Let me illustrate this regression effect with an actual research

reported in the psychological literature. I select this example without malice—I might have selected any of a number of others—because it is known to me and because it illustrates my point so perfectly. Crissey* has reported an investigation of mental development in orphanages and institutions for the feebleminded. Among other things, he selected from the test records a group in the orphanage and a group in the institution matched in initial I.Q. The average I.Q. of the orphanage population was 85, of the institution population 65. Now a single Binet I.Q. is not infallible, even as an indicator of performance on that test the next day, and is a good deal less so as a prediction of performance a year or two later. We should expect the high scores in *each population* to drop toward the population mean and the low scores in *each population* to rise. Assuming a test-retest correlation of .80 for the unspecified interval of time between tests in this study and assuming the conditions mentioned on page 93, we should expect an I.Q. difference between these two matched sub-samples on the retest of about 4 points of I.Q. Of course, the orphanage children should score the higher. The obtained difference was 6 points. The bulk of the obtained difference needs no other explanation, therefore, than the fact that *scores on a fallible test tend to regress toward the mean of the particular population to which they belong.*

The third pattern arises when we deal with groups which are differentiated with respect to amount of one continuous variable and matched with respect to another correlated variable. This differs from type No. 2, which we have just considered, only in that the differentiating factor in the populations from which we select our matched groups is amount of some quantitative trait rather than membership in one or another discrete category. We might, for example, give a large group of pupils an intelligence test and an arithmetic test. Directing our attention to those cases which fell in the top fourth and those which fell in the bottom fourth in intelligence score, we could—with some difficulty—so select a smaller sample from each of these fourths that the samples were matched in arithmetic test score. If these two groups were then given practice in memorizing nonsense syllables, and were subsequently retested upon another form of the arithmetic test, we might be led to attribute the substantial difference between the two groups on this retest (which we would undoubtedly find) to the differential effect of memorizing nonsense syllables upon the arithmetical ability of bright and dull children.

* Crissey, O. L. Mental development as related to institutional and educational residence. *University of Iowa Studies in Child Welfare*, 1937, 18, No. 1, p. 81.

FIGURE 2
Regression in Continuous Variables



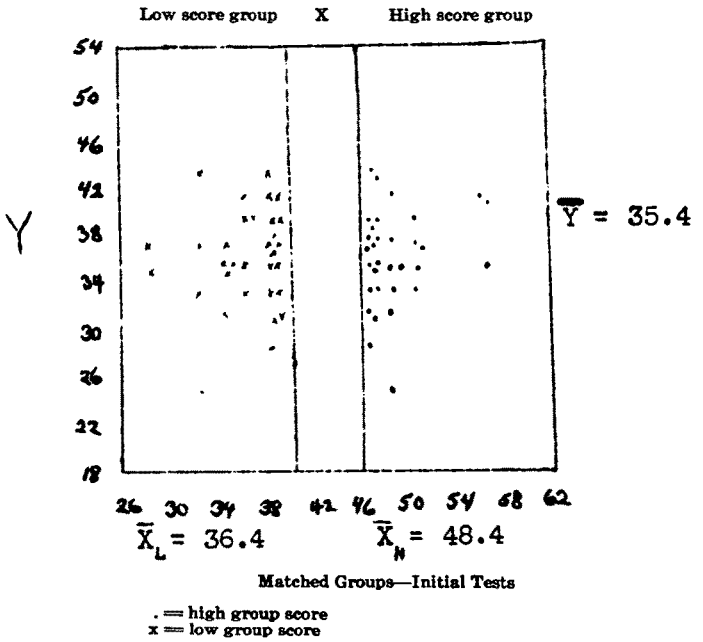
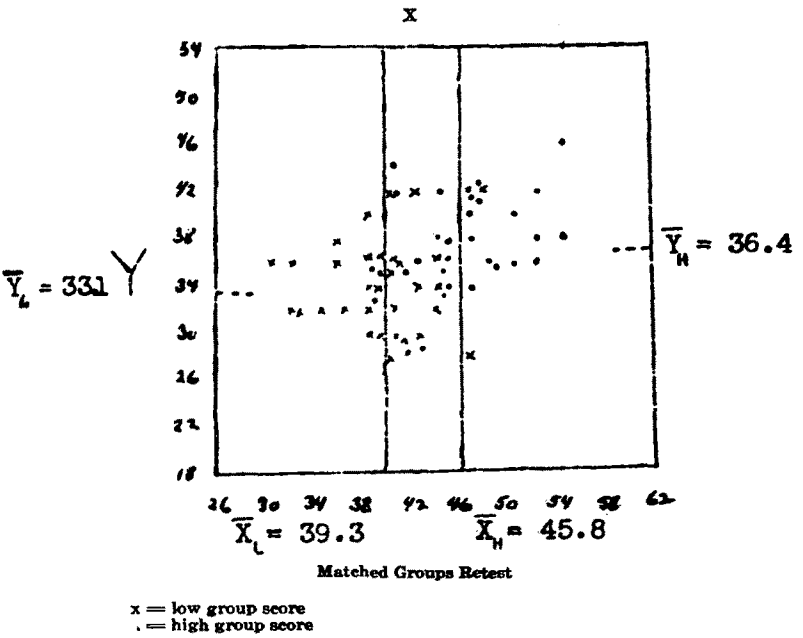


FIGURE 2

Regression in Continuous Variables



Actually, however, the difference in the groups on the third test (in this instance a repetition of the arithmetic test) is a regression phenomenon. The essential feature of our previous situation is retained, in that the two *populations* (children high in intelligence and children low in intelligence) have different means upon the variable which has served as the basis for matching (arithmetic achievement). Upon a retest, each sample mean will move toward its population mean, and the samples will no longer be matched.

Some artificial data to illustrate this point have been prepared, using dice throws as before, and are shown in Figure 2. Scores on a test and retest for two variables, X and Y, were composed, using certain dice for factors general to X and Y, certain ones for factors specific to each variable, and certain ones for errors of measurement on each test. Setting up two populations differentiated in X score, cases were matched upon the basis of Y score. Then the retest scores in both X and Y for the matched groups were examined. The difference in mean X score was 12.0 on the test, and dropped to 6.5 on the retest. The two groups had the same mean Y score on the initial test, but on the retest they differed by 3.3 points. They drew together upon the

test upon which they had been differentiated, and drew apart on the test upon which they had been matched.

Here, again, it is possible to express in a formula the amount of divergence to be expected between the means of the two groups upon a third test (which may or may not be a repetition of the matching test) when they are matched on one variable Y and differentiated on another variable X. Calling the third variable Z, we get the following

$$\bar{\bar{Z}}_H - \bar{\bar{Z}}_L = \frac{r_{xz} - r_{xy}r_{yz}}{1 - r^2_{xy}} \frac{S_z}{S_x} (\bar{X}_H - \bar{X}_L). \tag{5}$$

If $r_{xy} = r_{xz}$, this can be expressed

$$\frac{\bar{\bar{Z}}_H - \bar{\bar{Z}}_L}{S_z} = \frac{r_{xy}(1 - r_{yz})}{1 - r^2_{xy}} \cdot \frac{\bar{X}_H - \bar{X}_L}{S_x}. \tag{6}$$

The derivation of this formula is outlined in the matmematical note. As we examine this formula, we see that the expected amount of separation, in standard deviation units, is

(1) a direct function of the difference between the means of the two differentiated sub-populations;
(2) a direct function of the correlation between the differentiating and the matching variables;
(3) an inverse function of the correlation between the matching and the experimental variable.

MATHEMATICAL NOTE

Given: A population of individuals measured on variables X and Y.

Within the total population two sub-populations are set off, having different distributions and mean values for X. (Usually one sub-population will consist of those scoring high in X and the other of those scoring low.) These sub-populations are designated X_H and X_L, respectively.

Two samples, matched with regard to average score on test Y, are established—one in each of the two sub-populations indicated above.

Required: To determine the expected difference between the means of the two samples upon variable Z.

Derivation: For any pair of values of X and Y, the predicted value of Z is

$$\bar{Z} = b_{zy \cdot x}Y + b_{zx \cdot y}X + C.$$

Summing over the n_H cases in sample H, we get

$$\bar{\bar{Z}}_H = \frac{\overset{n_H}{\Sigma} \bar{Z}}{n_H} = b_{zy \cdot x} \frac{\overset{n_H}{\Sigma} Y}{n_H} + b_{zx \cdot y} \frac{\overset{n_H}{\Sigma} X}{n_H} + C$$

$$= b_{zy \cdot x} \bar{Y}_H + b_{zx \cdot y} \bar{X}_H + C .$$

Similarly, for the n_L cases in group L, we get

$$\bar{\bar{Z}}_L = b_{zy \cdot x} \bar{Y}_L + b_{zx \cdot y} \bar{X}_L + C .$$

Since the groups were matched on the test Y, we know that $\bar{Y}_H = \bar{Y}_L$. Therefore

$$\bar{\bar{Z}}_H - \bar{\bar{Z}}_L = b_{zx \cdot y} (\bar{X}_H - \bar{X}_L)$$

$$= \frac{r_{xz} - r_{xy} r_{yz}}{1 - r^2_{xy}} \cdot \frac{S_z}{S_x} (\bar{X}_H - \bar{X}_L) . \tag{5}$$

If $r_{xy} = r_{xz}$, this can be expressed

$$\frac{\bar{\bar{Z}}_H - \bar{\bar{Z}}_L}{S_z} = \frac{r_{xy}(1 - r_{yz})}{1 - r^2_{xy}} \cdot \frac{\bar{X}_H - \bar{X}_L}{S_x} . \tag{6}$$

Again, the formula can be generalized to any number of matching or differentiating variables. If two groups have been set up which are differentiated on variable X, but matched on Y and W, the expected difference on variable Z is

$$\bar{\bar{Z}}_H - \bar{\bar{Z}}_L = b_{zx \cdot yw} (\bar{X}_H - \bar{X}_L),$$

and the extension of this to any number of matching variables is quite clear. If the groups have been differentiated on X and Y, but matched on W, the expected difference on Z is

$$\bar{\bar{Z}}_H - \bar{\bar{Z}}_L = b_{zx \cdot yw} (\bar{X}_H - \bar{X}_L) + b_{zy \cdot xw} (\bar{Y}_H - \bar{Y}_L),$$

and again the extension to additional variables is straightforward. The practical problems which will arise will concern the feasibility of determining the desired statistics.

Real situations do arise involving just the type of regression which is discussed above. Two have been encountered recently in proposed doctoral dissertations. In one case, a general intelligence test and an analogies test had been administered to a population of students, and two groups were selected which were differentiated in intelligence but matched in analogies score. After some intervening training, another analogies test was given. It was found that the matched group with the higher intelligence did reliably better on the second analogies test. Since the correlations between the intelligence test and analogies tests were quite high, and since the correlation between the two analogies tests was far from perfect, this result could have been predicted, entirely without regard to the intervening experiences.

In the second case, personality characteristics were to be studied in two groups which were matched in intelligence, but sharply differentiated in school achievement. Such groups could be built up, using intelligence test score on the one hand and achievement test score, school grades, and teachers estimates on the other. But it can safely be predicted that upon a subsequent retesting they would be found to be neither accurately matched in intelligence nor so sharply differentiated in school achievement.

Having examined the regression effects which appear when we are dealing with groups from dissimilar populations, we are now led to ask: What are we going to do about them? What adaptations should we make in our experimental design or statistical treatment?

The most usual answer has been: Ignore them. This may, in some cases, be a reasonable expedient. When the populations do not differ greatly in the distribution of the measures being studied, or when the correlation between the matching and experimental variables is very high, the systematic errors introduced by regression will be small, and may very probably be insignificant in comparison with the effect of the other factors which are being studied. But in other cases, when the two populations differ more sharply and the intercorrelations are lower, the systematic regression errors may be of such size as to lead to entirely erroneous conclusions. This possibility must always be kept in view.

A second possibility is to insist that all investigations be carried out and all comparisons based upon groups selected from within the same population. This is certainly an ideal to be striven for. It is desirable not only because it eliminates those regression fallacies in matched group procedures with which we have been concerned, but also because it makes usable other efficient and powerful techniques of treatment. The analysis of covariance technique,* the techniques developed by Johnson and Neyman,* and a procedure suggested by Peters* all make it possible to use every case in each group studied as a basis for determining the effect of experimental treatments. These procedures all involve correcting scores on the experimental variable in terms of differences in background traits, on the assumption that the same regression equation of experimental upon background traits

* For a discussion of analysis of covariance, see Snedecor, G. W. Statistical Methods. Ames, Iowa: Collegiate Press, 1938.
For Lindquist, E. F. Statistical Analysis in Educational Research. New York: Houghton Mifflin, 1940.

* Johnson, P. O. and Neyman, J. Tests of Certain Linear Hypotheses and Their Application to some Educational Problems. *Statistical Research Memoirs*, 1936, 1

* Peters, C. C. A Method of Matching Groups for Experiment With No Loss of Population. *J. educ. Research*, 1941, 34, 606-612.

is appropriate for each group. That is, the assumption must be made that each group is a sample from the same population. These other procedures will probably be generally preferred to the procedures involving matching, since matching may make for either administrative difficulty or the loss of cases with consequent lowering of experimental efficiency.

It must be emphasized that all the methods just mentioned assume, either explicitly or implicitly, that the same regression equation between background traits and experimental variable is appropriate for both, or in the case of more than two, all groups. When this is not the case, those same regression fallacies which we have discussed in the case of matched groups are once more encountered and group differences arise simply because of differential regression effects. Analysis of covariance, the Neyman-Johnson methods, and the procedure suggested by Peters do not make any allowance for differences in the regression equation, arising most commonly out of differences in the means of the populations from which the experimental groups were taken.

Although it would be well, insofar as possible, to avoid investigations involving groups from two populations differing appreciably in the characteristics under study, there may be some cases when data of this sort are the only kind available and must be used. The school achievement of delinquents can be assessed only by comparing it with that of nondelinquents of comparable ability, even though these groups come from quite diverse parent populations. The gains from taking Latin can be studied only by comparing a group of Latin-studying pupils with a group of equivalent non-Latin-studying ones, even though the two total pupil populations may be significantly different in certain academic traits. In cases of this sort, some procedure to take account of differential regression effect is urgently needed.

We can recognize, in the last paragraph, two types of situations calling for samples from different parent populations. In the first type, exemplified by the delinquent- nondelinquent comparison, we are concerned with the effects of mere membership in a particular group or category together with whatever that membership may involve or imply. We wish to determine whether groups which are truly equivalent with regard to some background trait are different with regard to the experimental variable. If the groups are to be equivalent in true score on the background trait, they must be matched on the basis of *predicted true score*—i.e., score predicted by the regression equation between original test on the background trait and a retest at the time of the experimental comparison. Since the regression equations

for the different populations *will not be the same* in the case which we are now considering, the predicted true scores for each individual must be determined from *the regression equation for his own population.* Groups matched in this way will be truly equivalent upon the background trait, and differences between them in the experimental trait must be due to some factor other than background trait differences.

In the second type of situation, exemplified by the Latin-non-Latin comparison, we are interested in studying the effects of a certain type of experimental treatment, but the exigencies of life are such that that experimental treatment is and can be applied only to a population which is selected and atypical of the generality of cases. In this case, our concern is to get groups which would, except for the effect of the experimental treatment, be equivalent in the final test of the ability or trait being measured. We should match members from our two groups in terms of *predicted final test score.* Again the regression equation for predicting final test score will be different for each population and each final test score must be predicted in terms of the regression equation for that population. The regression equation for each population must be the one which holds *when the special experimental treatment is not applied,* or else the effect of the experimental treatment will be absorbed into our regression equation. If regressed values are used as indicated above and the groups are matched, an observed difference in actual final test scores will be attributable to the effect of the experimental treatment.

These procedures are straightforward, but involve quite a burden of computation. The chief difficulty to be encountered will be in determining the regression equations for the different parent populations. An expedient which may often be useful here is to assume that the variability and correlation are the same in the several parent populations. These could then be approximated by summing up the variances and covariances from the sample of each population which we have tested and computing the variance and correlation from these summed values. In that case, only the values for the means would be different in the regression equations for the several populations.

When it is no longer possible to match groups in terms of regressed scores as indicated in the preceding paragraph, it may still be possible to make some allowance for regression effects, making use of formulas (3) and (5) of this paper. Some of the statistics called for in these formulae may not be available, but even where it is not possible to compute the allowance precisely, it may be estimated on the basis of reasonable assumptions about the populations statistics.

Such an estimate will probably yield a much less biased result than the raw experimental differences.

None of the expedients suggested in this paper seems wholly satisfactory. What is really needed is some adaptation of the analysis of covariance to make it applicable to groups taken from populations having different regressions for the traits being studied. It is to be hoped that such an adaptation may soon be supplied to research workers.