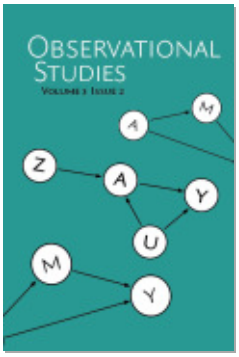
Regression-Discontinuity Analysis: An Alternative to the Ex-Post Facto Experiment



Donald L. Thistlewaite, Donald T. Campbell

Observational Studies, Volume 3, Issue 2, 2017, pp. 119-128 (Article)

Published by University of Pennsylvania Press
DOI: *https://doi.org/10.1353/obs.2017.0000*

Observational Studies 2 (2016) 119-128 Submitted 1960; Published reprinted, 12/16

Regression-Discontinuity Analysis: An Alternative to the Ex-Post Facto Experiment

Donald L. Thistlewaite and **Donald T. Campbell**

Editor's Note: Donald Thistlewaite (1923-1997) was Professor of Psychology at Vanderbilt University and Donald Campbell (1916-1996) was University Professor at Lehigh University. This article was originally published in *Journal of Educational Psychology*, December 1960, Vol. 51, pp. 309-317. At the time the article was published, Donald Thistlewaite was at the National Merit Scholarship Corporation and Donald Campbell was at Northwestern University. The article is now in the public domain. Comments follow by leading current researchers in regression discontinuity designs: Peter Aronow, Nicole Basta, and Betz Halloran; Matias Cattaneo and Gonzalo Vazquez-Bare; Guido Imbens; Alessandra Mattei and Fabrizia Mealli; Jasjeet Sekhon and Rocío Titiunik; and Vivian Wong and Coady Wing.

1. Introduction

While the term "ex post facto experiment" could refer to any analysis of records which provides a quasi-experimental test of a causal hypothesis, as described by Chapin (1938) and Greenwood (1945), it has come to indicate more specifically the mode of analysis in which two groups – an experimental and a control group – are selected through *matching* to yield a quasi-experimental comparison. In such studies the groups are presumed, as a result of matching, to have been equivalent prior to the exposure of the experimental group to some potentially change inducing event (the "experimental treatment"). If the groups differ on subsequent measures and if there are no plausible rival hypotheses which might account for the differences, it is inferred that the experimental treatment has caused the observed differences.

This paper has three purposes: first, it presents an alternative mode of analysis, called regression-discontinuity analysis, which we believe can be more confidently interpreted than the ex post facto design; second, it compares the results obtained when both modes of analysis are applied to the same data; and, third, it qualifies interpretations of the ex post facto study recently reported in this journal (Thistlethwaite, 1959). Two groups of near-winners in a national scholarship competition were matched on several background variables in the previous study in order to study the motivational effect of public recognition. The results suggested that such recognition tends to increase the favorableness of attitudes toward intellectualism, the number of students planning to seek the MD or PhD degree, the number planning to become college teachers or scientific researchers, and the number who succeed in obtaining scholarships from other scholarship granting agencies. The regression-discontinuity analysis to be presented here confirms the effects upon success in winning

scholarships from other donors but negates the inference of effects upon attitudes and is equivocal regarding career plans.

2. Method

2.1 Subjects and Data[1]

Two groups of near-winners–5,126 students who received Certificates of Merit and 2,848 students who merely received letters of commendation–answered a questionnaire approximately 6 months after the announcement of awards in the second National Merit Scholarship program. The C of M group received greater public recognition: their names were published in a booklet distributed to colleges, universities, and other scholarship granting agencies and they received approximately two and one half times more newspaper coverage than commended students. The decision to award some students the Certificate of Merit, which meant greater public recognition, was made chiefly on the basis of "qualifying scores" on the CEEB Scholarship Qualifying Test (SQT). A second aptitude test, the Scholastic Aptitude Test, was used to confirm the high ability of all finalists, i.e., all students scoring above the SQT qualifying score for the state in which the student attended high school.[2] Two hundred and forty-one students who voluntarily withdrew from the program before the second test or whose scores were not confirmed received neither award while 7,255 students who satisfactorily completed the second test received Certificates of Merit. The latter were subsequently screened by a selection committee and 827 of these students were awarded Merit Scholarships. Since the interest is in estimating the effects of honorary awards, questionnaire responses from Merit Scholars are not included in these analyses. As Table 1 shows, response rate did not vary systematically by test score interval, and there is no reason to believe that differential response bias can account for the effects to be described.

2.2 Regression-Discontinuity Analysis

In situations such as the foregoing, where exposure to an experimental treatment (in this case, increased public recognition) is determined by the subject's standing on a single, measured variable, and where the expected effects of the treatment are of much the same nature as would be produced by increasing magnitudes of that variable, examination of the details of the regression may be used to assess experimental effects. The experimental treatment should provide an additional elevation to the regression of dependent variables on the exposure determiner, providing a steplike discontinuity at the cutting score.

The argument–and the limitations on generality of the result–can be made more specific by considering a "true" experiment for which the regression-discontinuity analysis may be regarded as a substitute. It would be both indefensible and infeasible to conduct an

[1] Details of the sample of students, the experimental treatment, and dependent variables are described in the previous report (Thistlethwaite, 1959), and only the essential features of the data collection will be discussed here

[2] Recognition awards in the 1957 Merit program were distributed so that the number of students recognized in each state was proportional to the number of public high school graduates in each state. Since there were marked state differences in student performance on this test, qualifying scores varied from state to state. All SQT scores represented a composite in which verbal scores were weighted twice as heavily as mathematical scores.

Table 1: Participants in 1957 Merit Program Classified by Aptitude Score Interval

Group	Scholarship qualifying test score interval[b]	Number of Merit Scholars	Number in designated sample[a]	Number of respondents	Percentage of designated sample responding	Percentage of C of M winners in each interval awarded Merit scholarships
Commended students	Below 1		419	322	76.8	
	1		318	256	80.5	
	2		368	281	76.4	
	3		320	258	80.6	
	4		407	338	83.1	
	5		324	259	79.9	
	6		333	267	80.2	
	7		280	213	76.1	
	8		301	248	82.4	
	9		256	201	78.5	
	10		262	205	78.2	
Totals			3,588	2,848	79.4	
Certificate of Merit winners	11	17	476	380	79.8	3.4
	12	22	466	370	79.4	4.5
	13	16	399	319	79.9	3.9
	14	17	371	298	80.3	4.4
	15	19	361	300	83.1	5.0
	16	34	358	289	80.7	8.7
	17	13	319	247	77.4	3.9
	18	18	345	256	74.2	5.0
	19	17	254	211	83.1	6.3
	20	23	301	237	78.7	7.1
	Above 20	631	2,778	2,219	79.9	18.5
Totals		827	6,428	5,126	79.7	11.4

[a] Intervals show the student's SQT score relative to the qualifying score in the student's state, e.g., subjects whose scores equaled the qualifying score are classified in Interval 11, those whose scores were one unit less than the qualifying score are classified in Interval 10, etc.

[b] The designated sample for commended students consisted of a 47% random sample of all commended students.

experiment in which a random group of students along the whole range of abilities would be given the C of M award while a randomly equivalent group received merely the letter of

commendation. However, a group of commended students who narrowly missed receiving the higher award might be given the opportunity of receiving extra recognition. Thus students in Interval 10 in Figure 1 might be randomly assigned to the different treatments of C of M award and no C of M award. The two half-circle points at 10 for Line AA' in Figure 1 illustrate a possible outcome for such a true experiment, the solid half-circle representing the award group, and the hollow half-circle the no award group. Alternatively, a similar true experiment might be carried out among students just above the cutting point (Score 11 in Figure 1). For reasons discussed below, the regression-discontinuity analysis attempts to simulate the latter of these two experiments, by extrapolating from the below-cutting-point line to an "untreated" Point 11 value (an inferred substitute for the no award "control group"). Thus the major evidence of effect must be a distinct discontinuity or difference in intercept at the cutting point. Outcomes such as those shown in Line AA' would, of course, be strictly demonstrated only for aptitude intervals adjacent to the cutting point, and inferences as to effects of the C of M award upon persons of other ability levels would be made in hazard of unexplored interactions of award and ability level. Inferences as to what the regression line would have looked like without the C of M award become more and more suspect the further the no award experience of Points 1 to 10 has to be extrapolated. The extrapolation is best for Point 11 and becomes increasingly implausible for Points 12 through 20.

To better illustrate the argument several hypothetical outcomes are shown in Figure 1. Line AA' indicates a hypothetical regression of the percentage exhibiting Attribute A as a function of score on the decision variable. The steplike discontinuity which begins at the point where the experimental treatment begins to operate would be convincing evidence that the certificate has had an effect upon Attribute A. Similarly, outcomes such as those shown by Lines BB' and CC would indicate genuine treatment effects. Line DD' is a pure case of no effect. Lines EE' and FF' are troublesome: there seems to be a definite change in the regression lines, but the steplike discontinuity at the cutting point is lacking. Consequently the points could merely represent continuous, curvilinear regressions. It seems best not to interpret such ambiguous outcomes as evidence of effects.

In applying this mode of analysis to the present data, the qualifying score in each state was used as a fixed point of reference, and students were classified according to the number of score intervals their SQT score fell above or below the qualifying score in their state. For example, in Figure 2 all students whose scores equaled the qualifying score in their state have been classified in Interval 11, while all those whose scores were one less than the relevant qualifying score have been classified in Interval 10. Data were analyzed only for subjects whose scores placed them within 10 score intervals of the relevant cutting point. Because of nonresponse to particular questionnaire items the *N*s for percentages and means in Figures 2-4 differ slightly from those shown in Column 4 of Table 1.

3. Results

3.1 Graphic Presentation of Results

Figures 2, 3, and 4 present the results for five variables, with least squares linear regression lines fitted to the points. In Figure 2, both regression lines for scholarships received seem to show a marked discontinuity at the cutting point. The persuasive appearance of effect

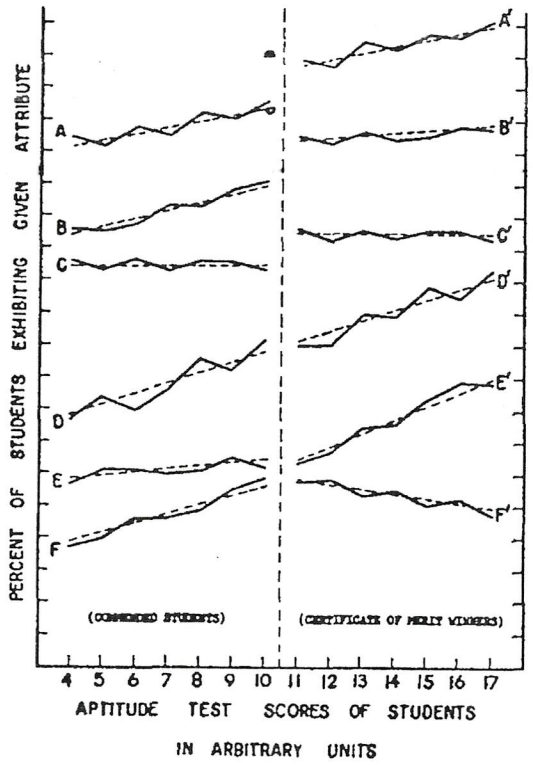


FIG. 1. Hypothetical outcomes of a regression-discontinuity analysis.

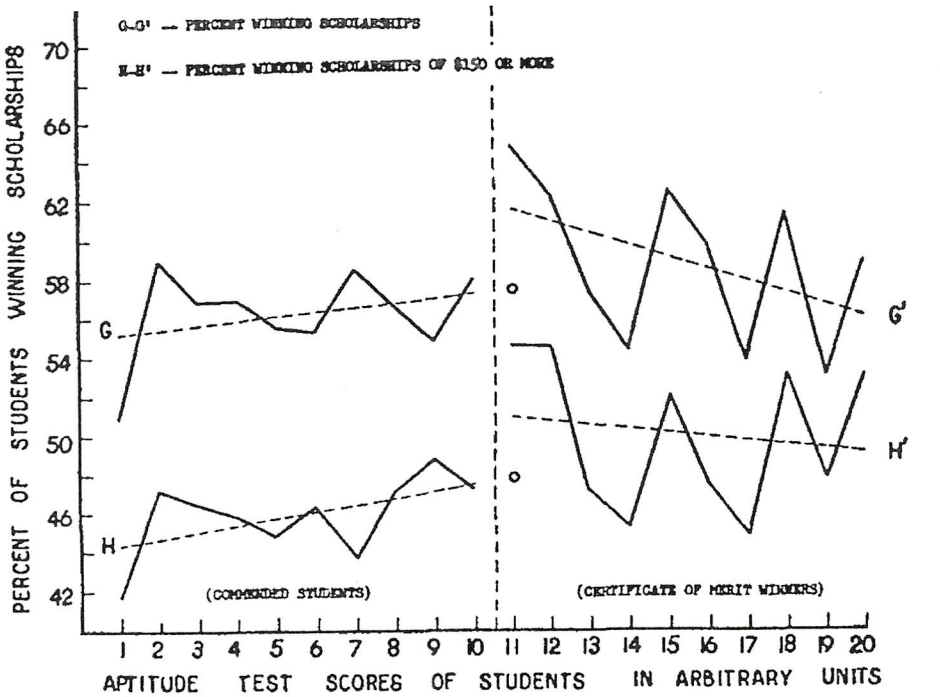


FIG. 2. Regression of success in winning scholarships on exposure determiner.

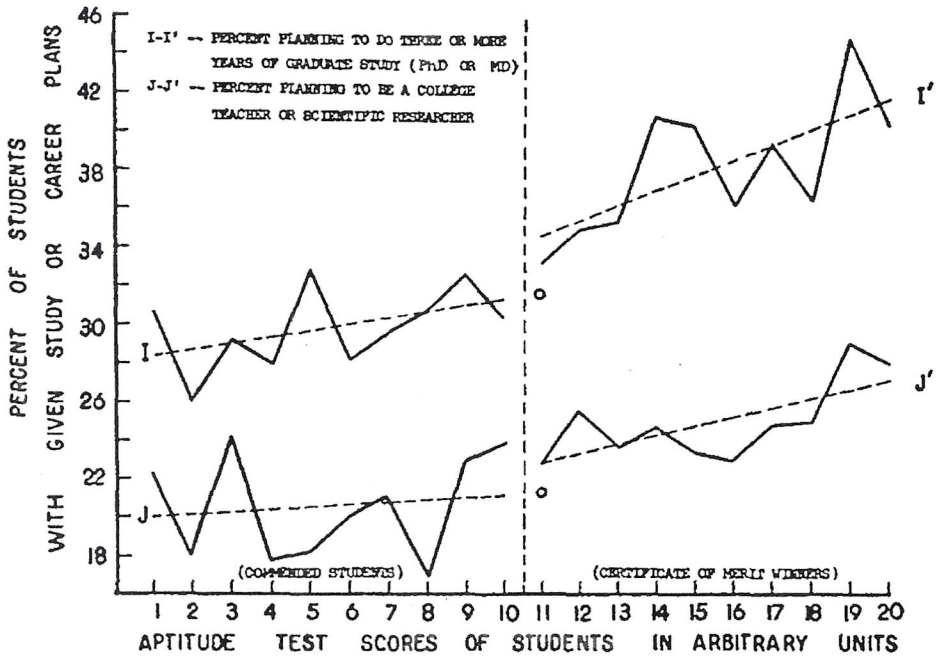


FIG. 3. Regression of study and career plans on exposure determiner.

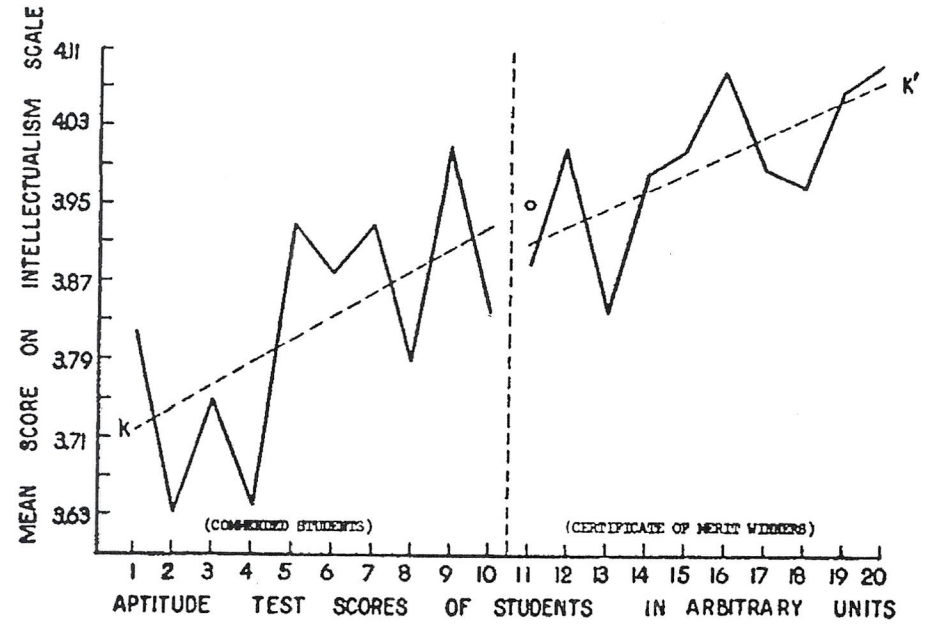


FIG. 4. Regression of attitudes toward intellectualism on exposure determiner.

is, however, weakened by the jaggedness of the regression lines at other points, particularly to the right of the cutting score. In addition, the slopes of the right-hand lines indicate that the effects are specific to students near the cutting score. The downward trend with high scores is presumably a result of eliminating from consideration those receiving Merit Scholarships. Where those of high aptitude test scores are passed over for National Merit Scholarships, it is usually for undistinguished high school grades, which likewise affect the scholarship awards by other agencies as plotted in Figure 2. Table 1 shows that, in general, larger proportions of C of M winners in the highest score intervals were selected for Merit Scholarships.

The two plots in Figure 3 show less discontinuity at the cutting point: there is little or no indication of effect. In II' the difference between observed values at 10 and 11 is small, and while in the hypothesized direction, is exceeded by five other ascending gaps. In JJ' the observed 10-11 jump is actually in the wrong direction. On the other hand, it is confirming of the hypothesis of effect that all of the observed Points 11 through 20 lie above the extrapolated line of best fit for Points 1 to 10, in both II' and JJ'. But this could well be explained by the rival hypothesis of an uninterrupted curvilinear regression from Points 1 to 20. The picture is ambiguous enough to leave us skeptical as to the effects upon the student's study and career plans. The analysis neither confirms nor denies the ex post facto findings.

In Figure 4 no such ambiguity remains. It is inconceivable in view of this evidence that the Certificate of Merit award has increased favorableness of attitudes toward intellectualism, a finding clearly contradicting the ex post facto analysis.

3.2 The Problem of Tests of Significance

In discussing tests of significance in this case, it is probably as important to indicate which tests of significance are ruled out as to indicate those which seem appropriate. Again, reference to the pure cases of Figure 1 will be helpful. A simple t test between Points 10 and 11 is excluded, because it would show significance in an instance like DD' if the overall slope were great enough. That is, such a test ignores the general regression obtained independently of the experimental treatment. Such a test between adjacent points is likewise ruled out on the consideration that even if significant in itself, it is uninterpretable if a part of a very jagged line in which jumps of equal significance occur at numerous other places where not expected. Similarly, a t test of the difference between the means of all points on each side of the cutting point would give significance to cases such as DD' or EE', which would be judged irrelevant. Furthermore, covariance tests applied to the regression lines (e.g., Walker & Lev, 1953, pp. 390-395) are judged inappropriate, because of the differential sample bias for the score intervals arising from the exclusion of Merit Scholars. Even in the ideal case, if the hypothesis of common slope is rejected (as it would be for lines such as EE' and FF') we presumably could not proceed further with a simple linear version of the covariance model.

Mood (1950, pp. 297-298) provides a t test appropriate for testing the significance of the deviation of the first experimental value beyond the cutting point (i.e., the observed Point 11) from a value predicted from a linear fit of the control values (i.e., the encircled point in Figures 2, 3, and 4, extrapolated from Point 1 through 10). As applied here, each plotted

point has been treated as a single observation. On this basis, both of the plots in Figure 2 show a significant effect at Point 11. For GG', $p < .025$; for HH', $p < .01$ (one-tailed tests). Thus the Certificate of Merit seems to have significantly increased chances of obtaining scholarships from other sources. For none of the other figures does this test approach significance. The test in this form fails to make use of the potentially greater stability made available by considering the trend of all of the Values 11 through 20. Potentially the logic of the Mood test could be extended to provide an error term for the difference between two extrapolated points at 10.5, one extrapolated from Points 1 through 10, the other from Points 11 through 20. In many applications of the regression discontinuity analysis, this would be the most appropriate and most powerful test. In our present instance, we have judged it inappropriate because of the differential sampling bias felt to exist in the range of Points 11-20, as explained above.

4. Discussion

A critic may easily question the results of an ex post facto experiment by supposing that one or more relevant matching variables has been inadequately controlled or entirely overlooked. In contrast the regression discontinuity analysis does not rely upon matching to equate experimental and control groups, hence it avoids the difficulties of (a) differential regression toward-the-mean effects, and (b) incomplete matching due to failure to identify and include all relevant antecedent characteristics in the matching process.

Edwards (1954, pp. 279-282) has shown how pseudo effects may be produced in ex post facto designs through differential regression effects. Suppose, for example, we were to match, with respect to aptitude test scores, a group exposed to recognition and a group not exposed to recognition. Since exposure to recognition tends to be positively correlated with aptitude test score we expect that the matched experimental subjects will have low aptitude scores relative to other exposed subjects, while the matched control subjects will have high aptitude scores relative to other unexposed subjects. To the extent that there are errors of measurement on the aptitude variable, however, our experimental group is apt to contain subjects whose aptitude scores are too low through error, while our control group is apt to contain subjects whose aptitude scores are too high through error. Simply on the basis of regression effects, then, we can predict that the matched experimental group will excel the matched control group on a subsequent administration of the aptitude test and on any other variable positively correlated with aptitude. Following Thorndike (1942, pp. 100-101), who discussed a similar problem, one might attempt to match individuals on the basis of predicted true score on the background traiti.e., score predicted by the regression equation between original test and a retest at the time of the experimental comparison. However, the predicted true score for each individual must be determined from the regression equation for his own population, and for groups when the special treatment is not applied. Unfortunately such matching is usually impossible in situations where we wish to use the ex post facto design, since we typically cannot obtain pretest and posttest measures on control variables for "experimental" groups from which the special treatment has been withheld. Indeed if we had the power to withhold the treatment from some subjects we would usually be able to test our causal hypotheses by an experiment with true randomization. In short, the

suggested procedure for controlling regression effects in ex post facto studies presupposes knowledge which we typically cannot obtain.

In the present analysis exposed and unexposed groups are subdivided according to their closeness to receiving a treatment other than the one they have received. Background traits correlated with the probability of exposure to recognition (e.g., rank in high school graduating class, scholastic aptitude, etc.) presumably vary systematically with the score intervals which represent the student's nearness to the cutting point. All of these traits contribute to the observed slopes of the regression lines plotted in Figures 2-A. Since there is no reason to believe that the composite effect of all relevant background traits fluctuates markedly at the cutting point, regression discontinuities emerging at the 10-11 gap must be attributable to the special experimental treatment-the only factor which assumes an abrupt change in value in this region. Thus the new analysis seems to provide a persuasive test of the presence or absence of experimental effects.[3]

The value of the regression-discontinuity analysis illustrated here is that it provides a more stringent test of causal hypotheses than is provided by the ex post facto design. Admittedly the class of situations to which it is applicable is limited. This class consists of those situations in which the regression of dependent variables on a single determiner of exposure to an experimental treatment can be plotted. Whenever the determiners of exposure are multiple or unknown this mode of analysis is not feasible. Of the five variables described in Figures 2-4 the regression-discontinuity analysis indicated significant effects only for those shown in Figure 2. The ex post facto experiment, on the other hand, indicated significant effects for all variables except HH' (success in winning a freshman scholarship of $50 or more). For six other variables, not reported here, neither analysis indicated a significant effect.[4] Considering the regression-discontinuity analysis to be the more definitive, it appears that the ex post facto experiment underestimated effects for one variable and wrongly indicated effects for three variables.

We conclude that increased public recognition tends to increase the student's chances of winning scholarships. There is no clear-cut evidence in the present analysis that such recognition affects the student's career plans, although an effect upon plans to seekgraduate or professional degrees is not ruled out. In this regard, Thistlethwaite (1961) has reported that when near-winners in a subsequent National Merit program were asked, "How did winning a C of M help you?" approximately two out of every five reported that it "increased my desire for advanced training (MA, PhD, MD, etc.)." In short, while other evidence indicates that the hypothesis of effect upon study plans may be correct, the present analysis does not provide confirmation.

[3]Background traits uncorrelated with the probability of exposure to recognition will, of course, not vary systematically with score intervals, but these traits are irrelevant. Even if partialed out they would not affect the correlation between the dependent variable and degree of exposure to recognition.

[4]No significant differences were found with respect to the percentages enrolling in college immediately, well satisfied with their choice of college, believing their college offers the best training in their field of study, going to college more than 250 miles from home, applying for two or more scholarships, or receiving encouragement from their high school teachers and guidance counselors to go to college.

5. Summary

The present report presents and illustrates a method of testing causal hypotheses, called regression-discontinuity analysis, in situations where the investigator is unable to randomly assign subjects to experimental and control groups. It compares the results obtained by the new mode of analysis with those obtained when an ex post facto design was applied to the same data. The new analysis suggested that public recognition for achievement on college aptitude tests tends to increase the likelihood that the recipient will receive a scholarship but did not support the inference that recognition affects the student's attitudes and career plans.

References

Chapin, F.S. (1936) Design for social experiments. *American Sociological Review*, 3: 786-800.

Edwards, A.L. (1954). Experiments: Their planning and execution. In G. Lindzey (Ed), *Handbook of social psychology.* Vol. 1. Cambridge, Mass : Addison-Wesley.

Greenwood, E. (1945). *Experimental sociology: A study in method.* New York: King's Crown.

Mood, A.M. (1950). *Introduction to the theory of statistics.* New York: McGraw-Hill.

Thistlewaite, D.L. (1959). Effects of social recognition upon the educational motivation of talented youth. *Journal of Educational Psychology*, 50: 111-116.

Thistlewaite, D.L. (1961). The recognition of excellence. *College and University*, 36: 282-295.

Thorndike, R.L. (1942). Regression fallacies in the matched group experiment. *Psychometrika*, 7: 85-102.

Walker, H.M. and Lev, J. (1953). *Statistical Inference.* New York: Holt.