DOCUMENT RESUME

ED 243 937 TM 840 250

AUTHOR Ludlow, Larry H.
TITLE Diagnostic Techniques in Research Synthesis.
PUB DATE Apr 84
NOTE 23p.; Paper presented at the Annual Meeting of the American Educational Research Association (68th, New Orleans, LA, April 23-27, 1984).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Data Analysis; *Effect Size; Estimation (Mathematics); Graphs; *Meta Analysis; Research Methodology
IDENTIFIERS Data Interpretation; *Residuals (Statistics)

ABSTRACT

One purpose for combining research studies is to estimate a population treatment effect. The internal validity of a model for how effect size estimates should be computed and combined will hinge upon the homogeneity of the effect size variation. Effect size variation may be assessed in the form of a summary fit statistic, and a direct consideration of the extent of individual effect variation from the population estimate. This paper presents some diagnostic techniques that facilitate the analysis of effect size variation. Bivariate plots of effect size residuals can aid in detecting sources of variation inconsistent with the model. Particularly, plotting the standardized residual of each study against the homogeneity of the sample if that study were removed is of interest for assessing the extent of heterogeneity contributed by individual studies. It is emphasized that the use of diagnostic techniques is useful for revealing why a lack of fit occurred, and is not advocated for the ad hoc purpose of finding a best-fitting subset of studies. (BW)

Diagnostic Techniques in Research Synthesis

by

Larry H. Ludlow
Boston College

April 1984

In the design and execution of a research project, an early step forward is taken when one proposes the mathematical model for how the observed data are the result of a set of hypothesized influences. This model will state how each influence enters as a parameter (e.g., item difficulty, treatment effect), how parameters interact (e.g., additive, multiplicative), and what conditions are assumed. A useful model provides a potential for predicting similar observations.

Models, then, serve as frames of reference for understanding the unseen, but hypothesized, forces operating in the world around us. Though useful, models are never perfect in any application, some experimental error is inevitable. This is due to our inability to completely control and measure the variables of interest. Consequently, the limits of models must be tested continually as we search to reveal their generality.

Since models can not serve as exact blueprints, there must be provided a mechanism that can assess how well models do function at least as rough outlines. In general the problem can be addressed by focusing on discrepancies in the observed data and their modelled estimates. The behavior of these discrepancies, or residuals, can usually be determined for situations ranging from near-perfect fit to extreme lack of fit. Once a baseline for the distribution of residuals is established it may be possible to analyze any new set of residuals for surprising, unexpected features. Unexpected features might be traced back to some anomoly in the original data or to an oversight in the parameterization of the model. They also might not have a reasonable explanation.

This simple preface on model building and model testing is relevant when one conducts a quantitative research synthesis. One purpose for combining research studies is to estimate a population treatment effect (Hedges, 1982a). The internal validity of a model for how effect size estimates should be computed and combined will hinge upon the homogeneity of the effect size variation. Effect size variation may be assesed in the form of a summary fit statistic, and a direct consideration of the extent of individual effect variation from the population estimate. This paper presents some diagnostic techniques that facilitate the analysis of effect size variation.

Effect size estimates may reflect a variety of comparisons, e.g. means, correlations, proportions (Cohen, 1977). For illustrative purposes we consider the case where the estimate represents the standardized difference between a pair of means. If we let the effect size estimate for a single study be

$$g_i = \frac{(\bar{Y}_i^E - \bar{Y}_i^C)}{S_i} \qquad \text{EQ. 1}$$

where:

$i = 1,2,\ldots,k$ studies

$\bar{Y}_i^E$ = experimental group mean

$\bar{Y}_i^C$ = control group mean

with S_i^2 as the pooled estimate of the variance, where

$$S_i^2 = \frac{(n_i^E-1)(S_i^E)^2 + (n_i^C-1)(S_i^C)^2}{(n_i^E+n_i^C-2)} \qquad \text{EQ. 2}$$

and n_i^E = experimental group n; n_i^C = control group n, then we can specify a weighted mean estimate of the population effect as

$$g. = GDOT = \frac{\sum_{i=1}^{k} \frac{g_i}{\sigma_{g_i}^2}}{\sum_{i=1}^{k} \frac{1}{\sigma_{g_i}^2}} \qquad \text{EQ. 3}$$

where

$$\sigma_{g_i}^2 = VARG = \frac{n_i^E + n_i^C}{n_i^E n_i^C} + \frac{g_i^2}{2(n_i^E + n_i^C)} \qquad \text{EQ. 4}$$

The estimate GDOT has variance

$$\sigma_{g.}^2 = VGDOT = \frac{1}{\sum_{i=1}^{k} \frac{1}{\sigma_{g_i}^2}}, \quad SDGDOT = \sqrt{VGDOT} \qquad \text{EQ. 5}$$

A 95% asymptotic confidence interval can be expressed as

$$LCI = GDOT - 1.96 * SDGDOT$$

$$UCI = GDOT + 1.96 * SDGDOT$$

The model for GDOT assumes that individual studies are independent estimates of a common population parameter, or in hypothesis testing terms: $H_0: \delta_i = \delta$, i = 1,2,...,k.

A test of this hypothesis can be expressed as

$$H = \sum_{i=1}^{k} \left(\frac{g_i^2}{\sigma^2_{g_i}} \right) - \frac{\left(\sum_{i=1}^{k} \frac{g_i}{\sigma^2_{g_i}} \right)^2}{\sum_{i=1}^{k} \frac{1}{\sigma^2_{g_i}}} \qquad \text{EQ: 6}$$

(The interested reader will find the complete presentation of Eq. 1 thru Eq. 6 in Hedges, 1982a.)

If H_0 is true, then the test statistic H has an asymptotic chi-square distribution given by $H \sim \chi^2_{k-1}$. If H is not significant, then we accept H_0. If H is significant, then we reject the hypothesis that all δ_i are equal.

With the finding of a significant H the analytic problem becomes one of determining which effect estimates contributed to the lack of fit. Between-group differences can be tested with the categorical or continuous model fitting techniques proposed by Hedges (1982b, 1982c), provided one has the information necessary to specify the groups. Those techniques help explain a lack of fit by revealing when effects are homogeneous within groups but heterogeneous between groups.

It may be the case that classification variables are not available or are not apparent. In this situation the investigator may wish to address the residuals from the model fitting process. Residuals are the differences between individual effects and the population estimate (Hedges and Olkin, 1984). A typical representation of an estimated residual and its standardized form is

$$\varepsilon_i = URES = g_i - g_\cdot$$

$$SRES = \frac{g_i - g_\cdot}{\sqrt{\sigma^2_{g_i} - \sigma^2_{g_\cdot}}} \qquad \text{EQ. 7}$$

When the model holds, ε_i has an expected value and variance of

$$E(\varepsilon_i) = 0, \quad V(\varepsilon_i) = 1$$

$$V(\varepsilon_i) = \sigma^2_{g_i} + \sigma^2_{g_\cdot} - 2\,cov(g_i, g_\cdot) = \sigma^2_{g_i} - \sigma^2_{g_\cdot}$$

This residual is computed as the difference between the j'th effect and the population estimate, where the population estimate includes the j'th effect.

If we want the difference between a particular effect and the other members of the sample, then we might be more interested in representing an estimated residual and its standardized form as

$$\tilde{\varepsilon}_i = URESJ = g_i - g_{\cdot(i)} \qquad \text{EQ. 8}$$

$$SRESJ = \frac{g_i - g_{\cdot(i)}}{\sqrt{\sigma^2_{g_i} + \dfrac{1}{\left(\sum_{i=1}^{k} \frac{1}{\sigma^2_{g_i}}\right) - \frac{1}{\sigma^2_{g_i}}}}}$$

where $g_{\cdot(i)}$ means the population estimate does not contain the j'th effect in the calculation. When the model holds, this residual has an expected value and variance of

$$E(\tilde{\varepsilon}_i) = 0 \; ; \; V(\tilde{\varepsilon}_i) = 1$$

$$= \sigma^2_{g_i} + \sigma^2_{g_{\cdot(i)}} - 2\,\mathrm{cov}(g_i, g_{\cdot(i)})$$

$$= \sigma^2_{g_i} + \sigma^2_{g_{\cdot(i)}}$$

Note that the numerators and denominators in Eq. 7 and Eq. 8 differ. In Eq. 8 the numerator will reflect a larger discrepancy between a given effect and the population estimate than will the difference computed with Eq. 7. The denominator, too, will be larger but the rate of change will be less than that of the numerator. The overall result is that residuals computed under Eq. 8 are larger than their counterparts computed under Eq. 7. Table 1 illustrates the difference between the two alternatives. The SRES residuals are computed from the population estimate with all studies included (GDOT). The SRESJ residuals are computed based on the population estimate with the j'th study removed (GDOTJ). The differences in the pairs of standardized residuals (SRESJ-SRES) are listed in the DIFF column. In each instance SRESJ is more extreme than SRES i.e, |SRESJ|-|SRES|=>0. It is asserted that these differences indicate that SRESJ is a more sensitive statistic than SRES for detecting heterogeneous variation. The remainder of this discussion addresses standardized residuals computed according to Eq. 8.

Regardless of the form of the residual the observed sum of the residuals is not likely to equal zero. Although each residual has an expected value of zero when the data fit the model and, therefore, the expected value of the sum of those residuals is zero, there is no algebraic requirement that estimated residuals must sum to zero when they are computed relative to a weighted estimator.

When an analysis of residuals is undertaken it is reasonable to decide first on an analytic approach that takes into account the number of studies under consideration. This is because residuals from an analysis based on 10 or fewer studies do not normally require the techniques that are useful when 10 or more studies are involved. This is a relevant consideration because sub-analyses of effect size data frequently involve fewer and fewer studies. For a small analysis it is usually sufficient to construct a table containing the original estimates, their residuals, the H statistic computed if that study were removed from the analysis (HJ), and the upper and lower 95% confidence interval if that study were removed. (It is relatively easy during the initial pass through the data to compute the second-step statistics that result when a given study is removed.)

An example of a summary and diagnostic table is presented as Table 2. The observed homogeneity statistic (H=8.92, df=6, p>.10), for the population estimate (GDOT=.326), is consistent with the hypothesis of homogeneous effects. The 95% confidence interval (.230 to .423) does not include zero. The iterated estimator of GDOT shows only a slight improvement, as expected when the data fit the model

(Hedges, 1982a, Eq. 14). In the diagnostic statistics section the residual for Study 3 (SRESJ=-2.01) results from a relatively small effect estimate (G=-.03), a fact reflected in the H statistic computed if this study were removed (a decline from H=8.92 to HJ=4.87). We also note that removal of this study would result in a rise in the estimate of GDOT from .326 to .35 (GDOTJ). None of the individual 95% confidence intervals include zero. We conclude that the individual estimates fit the model and the population estimate is significantly greater than zero.

The analytic situation changes when more than 10 or 20 studies are involved. The basic problem is that it is difficult to scan long columns of values in any systematic manner. It is, however, relatively easy to construct a few simple bivariate plots to aid in detecting sources of variation inconsistent with the model. The purpose of these plots is to focus on the continuity and the range in the distribution of the residuals. That is, do the residuals form a narrow, unbroken pattern or do they tend to form clusters separated by recognizable gaps with occasional outliers lying a considerable distance from the main body?

The plots presented in this paper do not address whether or not the distribution of residuals fits a specific hypothesized form, i.e., the standard normal. The techniques developed for assessing the statistical distribution of a set of residuals are appropriate here but they have yet to attract serious attention. Perhaps this is because the analytic question which these techniques address has not demonstrated its practical significance in a quantitative synthesis context. In part, this is because techniques such as probability plotting have been shown to be highly variable when relatively few data points are plotted (cf. Daniel and Wood, 1980, Appendix 3A).

The following discussion is based on data collected in 1983 by Professor George Hillocks, Jr. His purpose was to determine if instructional strategies lead to a significant improvement in writing composition skills, and if so, is there a significant difference between the effects of the alternative strategies. A total of 39 studies were included. They represent six instructional strategies: grammer (k=5), models (k=7), sentence combining (k=5), scales (k=6), inquiry (k=6), and free-writing (k=10). The fit of these data to the model was inconsistent with the hypothesis of homogeneous effect size variation (H=84.48, df=38, p<.0001). An analysis of the residuals was undertaken.

It is evident that the relation between an effect estimate and its residual is necessarily linear. In fact, the relation will be perfectly linear for G versus URES or G versus URESJ plots. This relation is demonstrated in Figure 1. Linearity may, however, be less than perfect for G versus SRES or for G versus SRESJ plots.

A less than perfect linear relation is possible because of the influence of sample size in computing the variance of an effect estimate. More weight, in the form of a smaller error term, is attributed to studies with the larger samples. Thus, it is possible for the more extreme of two effect estimates to have the smaller estimated residual. This means that: a) sole consideration of the largest and smallest effect estimates is not sufficient, and b) the preferred choice of residual will usually be one that has been standardized. Obviously, if the sample sizes are identical the unstandardized and standardized residual plots will be identical.

A plot of G versus SRESJ can be interesting because the heterogeneous G estimates will be found in the tails of the distribution. Our attention will be drawn to gaps in the distribution or clusters in the tails. An example is provided in Figure 2. Two features are noteworthy. The first is the less than perfect linear relation (though r=.94). Study A has the most negative effect estimate (G=-.27) but its residual (SRESJ=-2.32) is not as extreme as the residual for study B (G=.05, SRESJ=-3.10). This is because of the difference in sample sizes. Study A contains samples of 41 and 36 persons. Study B consists of samples with 420 and 371 persons. The difference between the study B estimate and the population estimate was accorded more weight than the difference between study A and the population because study B was based on considerably larger samples.

The tails of the distribution define the second interesting feature. The three studies in the negative tail represent "free-writing" strategies. Two of the three studies in the positive tail belong to the "inquiry" category.

What consequences do these extreme estimates have upon the overall fit of the data to the model? A natural plot to consider would be G versus HJ, each effect estimate plotted against the homogeneity of the sample if that study were removed. The problem with this particular plot is that it does not fully represent the adverse impact of a potentially heterogeneous study. This is because a plot with G as an axis does not take into account the sample size. Thus, it is the standardized residual (SRESJ) versus HJ plot which is of interest for assessing the extent of heterogeneity contributed by individual studies.

Figure 3 illustrates the relation between the standardized residuals and the improvement of fit to the model if their respective studies are removed (SRESJ versus HJ). The plot is necessarily quadratic because increasingly larger and smaller effects diverge from the population estimate. The interesting regions are the extremes of the curve where either the curve extends for a substantial distance or gaps occur.

The two studies with the greatest over-estimate of the population effect and two of the studies in the next cluster of points involve "inquiry" strategies. The four studies with the greatest under-estimate of the population effect involve "free-writing" strategies. These findings supported the decision to group and analyze common instructional strategies separately (see Hillocks, 1984).

There can be a problem with this type of plot. Depending on how wide the plot boundaries are defined, a relatively slight difference can be transformed into a large gap. One could attempt to produce a standardized graph by dividing each HJ by the degrees of freedom in the analysis. This plot still looks quadratic but now the values tend to be identical if the data fit, or the values tend to deviate from the main cluster of points by only a slight margin. An example is presented in Figure 4 (SRESJ versus HJ/DF). How to construct a more useful standardized plot remains to be discovered.

In conclusion, I emphasize that the use of diagnostic techniques is not advocated for the ad hoc purpose of finding a best-fitting subset of studies. Such a purpose is clearly meaningless for estimating a population effect. The techniques are, however, useful for revealing why a lack of fit occurred. The issue of whether individual studies should be removed from consideration or should be formed into subsets for separate analysis must be based on methodological considerations that are consistent with the original criteria stated for including or excluding studies from the original design. That is, in the initial stages of a project it is possible that studies have been included that the investigator accepts as marginally relevant but which are believed to be consistent with the studies of direct interest. This tactic is taken occasionally when one seeks to increase the number of studies in the analysis. It may also be the case that the first test of the data will be to determine the extent of heterogeneity, given that differences are assumed to exist but one wants to verify that is indeed the situation. If specific studies or groups of studies do not fit the omnibus analysis and if there was some a priori awareness that they might not, then their exclusion from consideration or the construction of a subset based on diagnostic results does seem warranted.

REFERENCES

Cohen, J. Statistical power analysis for the behavioral sciences (Revised edition). New York: Academic Press, 1977.

Daniel, C. & Wood, F.S. Fitting equations to data. New York: Wiley, 1980.

Hedges, L.V. Estimating effect size from a series of experiments. Psychological Bulletin, 1982, 92, 490-499. (a)

Hedges, L.V. Fitting categorical models to effect sizes from a series of experiments. Journal of Educational Statistics, 1982, 7, 119-138. (b)

Hedges, L.V. Fitting continuous models to effect size data. Journal of Educational Statistics, 1982, 7, 245-270. (c)

Hedges, L.V. & Olkin, I. Statistical methods in meta-analysis. New York: Academic Press, 1984.

Hillocks, G. Jr. A meta-analysis of experimental treatment studies in the teaching of composition: A summary of results. Chicago: University of Chicago, Department of Education, 1984.

Ludlow, L.H. HSTAT: A Fortran program for computing the homogeneity of a quantitative research synthesis. 1983, Boston College, School of Education.

Diagnostic Techniques in Research Synthesis
Larry H. Ludlow, Boston College, AERA, 1984

(HANDOUT)

1 $$G = \frac{(\overline{Y}^e_j - \overline{Y}^c_j)}{S_j}$$, j = 1,2,...,k studies; $\overline{Y}^e$ = experimental group mean

$\overline{Y}^c$ = control group mean

(G is the effect estimate for study J)

2 $$S_j^2 = \frac{(n_j^e - 1)(S_j^e)^2 + (n_j^c - 1)(S_j^c)^2}{(n_j^e + n_j^c - 2)}$$; n_j^e = experimental group n

n_j^c = control group n

3 $$GDOT = \frac{\sum_{j=1}^{k} \frac{G}{VARG}}{\sum_{j=1}^{k} \frac{1}{VARG}}$$ (GDOT is the weighted mean estimate of the population effect parameter)

4 $$VARG = \frac{(n_j^e + n_j^c)}{n_j^e n_j^c} + \frac{G^2}{2(n_j^e + n_j^c)}$$ (VARG is the variance of G)

5 $$VGDOT = \frac{1}{\sum_{j=1}^{k} \frac{1}{VARG}}$$, $SDGDOT = \sqrt{VGDOT}$ (VGDOT is the variance of GDOT)

LCI = GDOT - 1.96 * SDGDOT
UCI = GDOT + 1.96 * SDGDOT

6 $$H = \sum_{j=1}^{k} \left(\frac{G^2}{VARG}\right) - \frac{\left(\sum_{j=1}^{k} \frac{G}{VARG}\right)^2}{\sum_{j=1}^{k} \frac{1}{VARG}}$$ (H is the test statistic for the homogeneity of the G estimate. It is distributed as χ^2_{k-1}.)

7 URES = G - GDOT

$$SRES = \frac{URES}{\sqrt{VARG - VGDOT}}$$ (SRES is the standardized residual when study J is included in GDOT)

8 URESJ = G - GDOTJ

$$SRESJ = \frac{URESJ}{\quad}$$ (SRESJ is the standardized residual when study J is not included in

STUDY	N1	N2	VARG	G —	GDOT =	URES	SRES	G —	GDOTJ =	URESJ	SRESJ	DIFF
1	16	11	.1573	.459	.119	.34	.86	.459	.09	.37	.90	.04
2	30	30	.0669	.181	.119	.06	.24	.181	.10	.08	.27	.03
3	30	30	.0689	-.521	.119	-.64	-2.48	-.521	.26	-.79	-2.70	-.22
4	44	40	.0478	.097	.119	-.02	-0.09	.097	.13	-.03	-0.12	-.03
5	37	55	.0462	.425	.119	.31	1.48	.425	.00	.42	1.68	.20

TABLE 1.--Differences in SRES depending on whether or not j'th study is included in g. estimate (These data are reported in Hedges, 1982a.)

SUMMARY STATISTICS

GDOT	VGDOT	SDGDOT	LCI	UCI	H	DF
.326	.002	.049	.230	.423	8.92	6

THE ITERATED ESTIMATOR OF GDOT

GDOT	LCI	UCI
.328	.231	.424

DIAGNOSTIC STATISTICS

STUDY	N1	N2	VARG	G	URESJ	SRESJ	GDOTJ	LCIJ	UCIJ	HJ
1	70	65	.0308	.56	.25	1.39	.31	.206	.407	7.00
2	71	49	.0345	.05	-0.30	-1.54	.35	.247	.447	6.54
3	75	49	.0337	-.03	-0.38	-2.01	.35	.254	.454	4.87
4	136	266	.0113	.35	.03	.25	.32	.211	.429	8.86
5	97	266	.0142	.31	-0.02	-0.15	.33	.224	.436	8.90
6	142	266	.0109	.34	.02	.15	.32	.213	.432	8.90
7	100	266	.0140	.45	.15	1.15	.30	.195	.407	7.61

TABLE 2.--Summary and diagnostic statistics for determining homogeneity of a small quantitative research synthesis (These data are reported in Hillocks, 1984. This table and its statistics are produced by the computer program HSTAT by Ludlow, 1983).

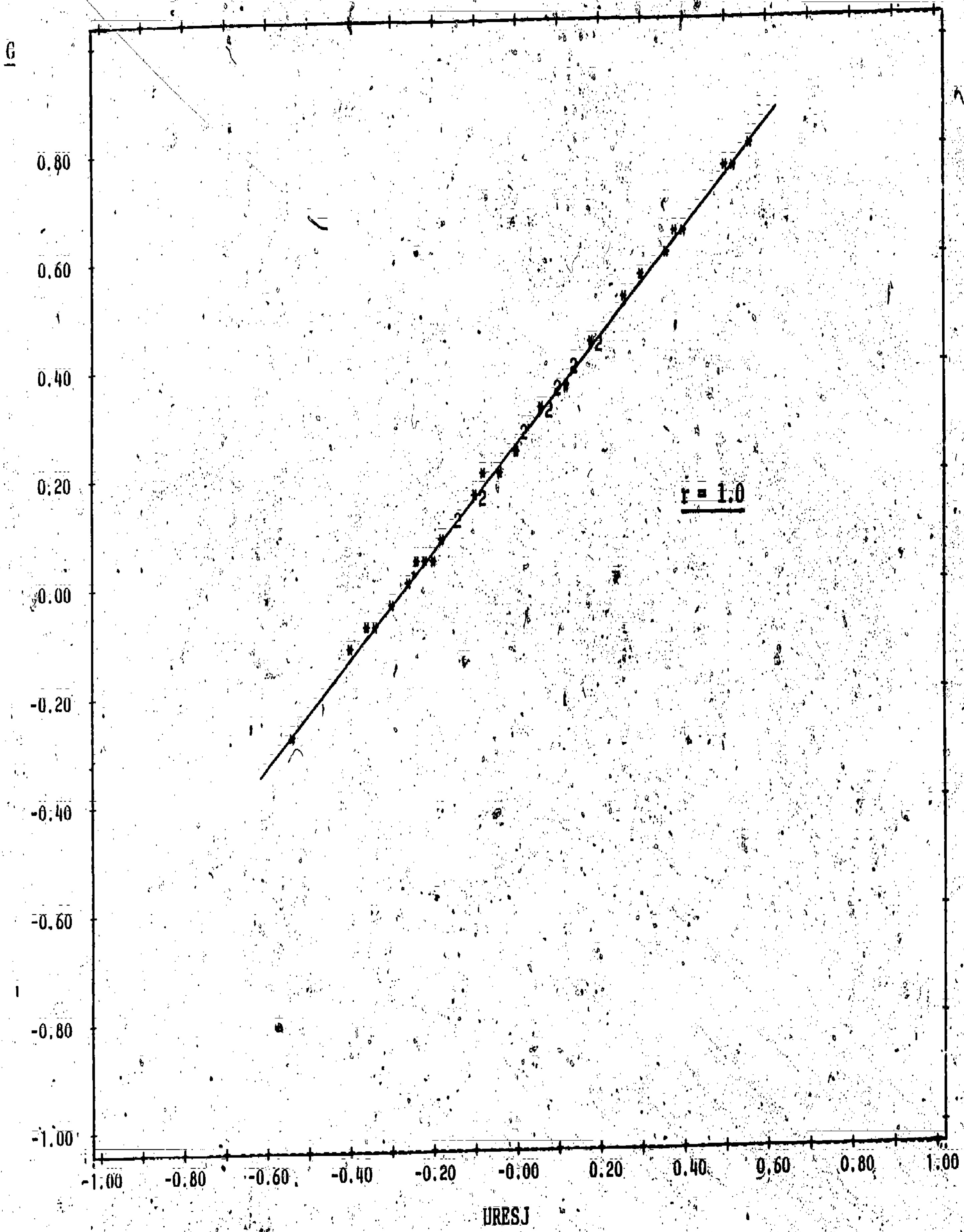

Figure 1.-- G versus URESJ

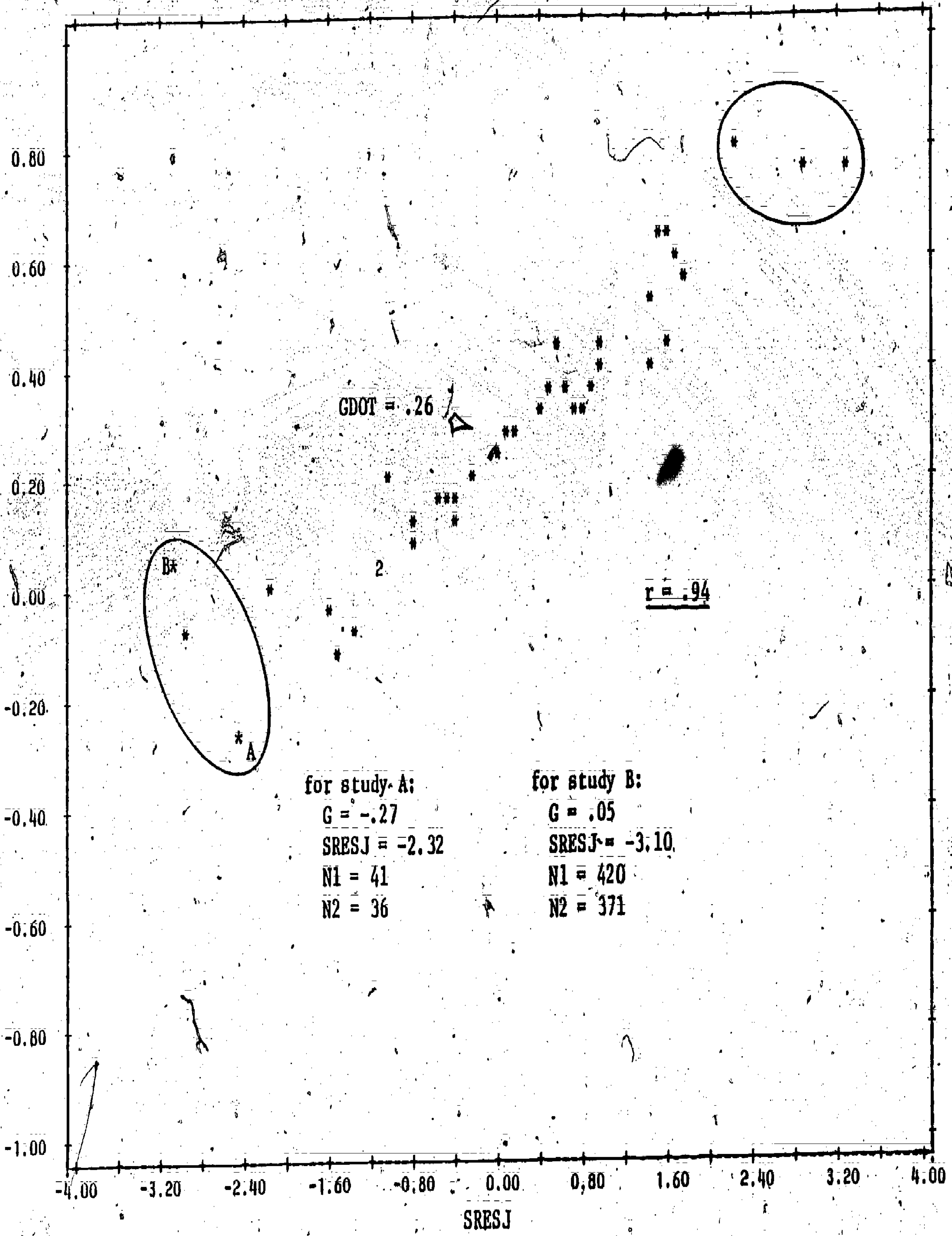

Figure 2.-- G versus SRESJ

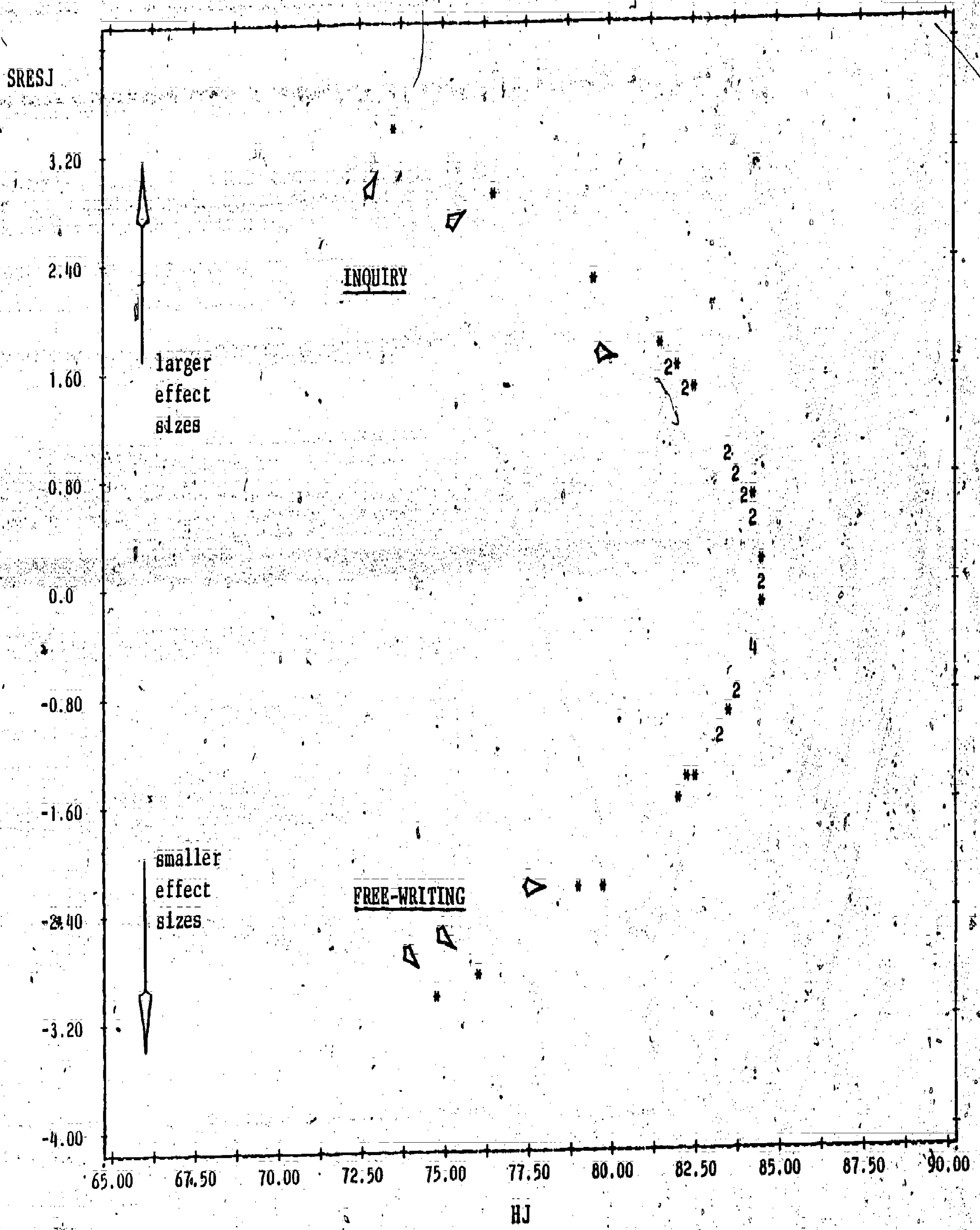

Figure 3.-- SRESJ versus HJ