DOCUMENT RESUME

ED 262 071 TM 850 548

AUTHOR Skaggs, Gary; Lissitz, Robert W.

TITLE An Exploration of the Robustness of Four Test Equating Models.

PUB DATE Apr 85

NOTE 47p.; Paper presented at the Annual Meeting of the American Educational Research Association (69th, Chicago, IL, March 31-April 4, 1985). Small print in tables 1-3.

PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.

DESCRIPTORS Computer Simulation; *Equated Scores; *Error of Measurement; *Goodness of Fit; Latent Trait Theory; *Mathematical Models; Scaling; Standardized Tests; Statistical Analysis; Statistical Studies

IDENTIFIERS Equipercentile Equating; Linear Models; Rasch Model; Robustness; Three Parameter Model

ABSTRACT

This study examined how four commonly used test equating procedures (linear, equipercentile, Rasch Model, and three-parameter) would respond to situations in which the properties or the two tests being equated were different. Data for two tests plus an external anchor test were generated from a three parameter model in which mean test differences in difficulty, discrimination, and lower asymptote were manipulated. In each case two data sets were generated consisting of responses of 2,000 examinees to a 35 item test plus 15 item anchor test. Each test equating case was comprised of 4,000 examinees and 85 items. The robustness with respect to violations of assumptions was tested for the linear and Rasch equating methods. For equipercentile equating, the results showed how the method responded to various conditions for the three-parameter model, the study primarily tested LOGIST's simultaneous estimation procedure. Results indicated that equipercentile equating was very stable across the cases studied. Linear and Rasch model equating were very sensitive to violations of their models' assumptions. Results for the three-parameter model were disappointing. Paying close attention to test item properties was advised when selecting an equating method. (BS)

An Exploration of the Robustness of Four Test Equating Models

Gary Skaggs
Robert W. Lissitz
University of Maryland

April 1985

Paper presented at the
Annual Meeting of the American Educational Research Association
Chicago, IL
April 1985

AN EXPLORATION OF THE ROBUSTNESS OF FOUR TEST EQUATING MODELS

Gary Skaggs
Robert W. Lissitz
University of Maryland

The application of item response theory (IRT) to many measurement problems has been one of the major psychometric breakthroughs of the past twenty years. IRT methodology is currently being used in many large standardized testing programs, and at the same time, a great deal of research is being done to evaluate the robustness of the procedures under a variety of conditions. One of the most important applications of this model is in the area of test equating.

The purpose of test equating is to determine the relationship between raw scores on two tests that measure the same ability. Equating can be horizontal, between tests of equivalent difficulty and content, or vertical, between tests of intentionally different difficulties.

Most conventional approaches to test equating have been describec as either linear or equipercentile methods (see Angoff, 1971) in contrast to using item response theory (IRT) (Lord, 1980; Wright & Stone, 1979) and are now widely used. In these techniques, a mathematical relationship between raw scores on tests is modeled. This relationship is based on estimates of item parameters from two tests and placement of these estimates on the same scale.

A number of equating studies using IRT methods have appeared in recent years. A majority of these have dealt with the one-parameter logistic, or Rasch, model. This model is the simplest of the IRT models but also the most demanding one in terms of its assumptions.

Several studies have found the Rasch model to be useful and appropriate for item calibration and linking (Tinsley & Dawis, 1975; Rentz & Bashaw, 1977; Guskey, 1981; Forsyth, Saisangjan, & Gilmer, 1981). On the other hand, a number of researchers have noticed problems with the Rasch model for vertical equating

(Whitely & Dawis, 1974; Slinde & Linn, 1978 1979, Loyd & Hoover, 1980; Holmes, 1982).

Some of the inconsistency can be attributed to different equating designs, different types of tests being equated, and different procedures used to analyze the results. Regardless of the cause, there are some fundamental concerns about the Rasch model. The most frequently postulated arguments concern the failure of the model to account for chance scoring, unequal discrimination, and multidimensionality. The last concern applies to more complex IRT models as well.

Research using the three-parameter logistic and other models has not been as plentiful as with the Rasch model, but it suggests the same interpretetive difficulties. Most of the studies have examined the three-parameter model in the context of horizontal equating of general ability tests. This work has generally supported the use of the three-parameter model (Marcho, Petersen, & Stewart, 1979; Kolen, 1981; Petersen, Cook, & Stocking, 1981). For vertical equating, results have been more mixed with some studies finding the there parameter model to be more effective than the Rasch model (Marco et al., 1979; Kolen, 1981). However, the comparison between the models has been shown to depend largely on the content of the tests being equated (Kolen, 1981; Petersen et al., 1981; Holmes & Doody-Bogan, 1983).

With all of this conflicting research, it is very difficult to make decisions about how to use IRT equating or whether to use it at all. The purpose of the present study is to explore how test equating results can be affected by the parameters of the items that make up the tests being equated.

Four methods of equating were chosen representing popular versions of linear, equipercentile, Rasch model, and three-parameter model techniques. Data for two tests plus an external anchor test were generated from a three-parameter model in which mean test differences in difficulty, discrimination, and lower asymptote were manipulated. For Rasch model and linear equating, this study is an exploration of robustness when the model's assumptions are violated. For the three-parameter model, this study amounts to an examination of the parameter estimation strategy. For the equipercentile equating, this study explores its effectiveness under a variety of test conditions.

METHOD

Data

Data in this study were generated from the three-parameter logistic model:

$$P(u_{ij} = 1/ \theta_j, a_i, b_i, c_i) = c_i + (1-c_i)(1 + \exp(-1.702a_i(\theta_j - b_i)))^{-1} \qquad [1]$$

The response to item i by person j, a 0 or 1, was determined by comparing the probability defined by equation 1 to a random number drawn from a (0,1) uniform distribution. If the probability of a correct response exceeded the random number, the item was scored as correct. Otherwise, the item was scored as incorrect. The random numbers were produced from the GGUBS (IMSL, 1980) generator.

In all simulation cases, an external anchor test design was used. In each case, two data sets were generated. Each data set consisted of the responses of 2,000 examinees to a 35 item test plus an anchor test of 15 items. Each test equating case was comprised of 4,000 examinees and 85 items.. This size was chosen to be large enough to provide stable parameter estimates for both IRT models (Lissak, Hulin, & Drasgow, 1982).

Item and Ability Parameters

The item parameters used to generate the data were determined by manipulating lower asymptotes and mean test difficulty and discrimination of the tests being equated. For each reference, the two tests will be referred to as test A and test B.

Item difficulty was studied at three levels: 1) $\overline{b}_A = \overline{b}_B = 0$; 2) $\overline{b}_A = -.5, \overline{b}_B = .5$; and 3) $\overline{b}_A = -1.0, \overline{b}_B = 1.0$. For each test, difficulties were uniformly distributed across a range of +/- 2 logits.

Item discrimination was also examined at three levels: 1) $\overline{a}_A = \overline{a}_B = .8$; 2) $\overline{a}_A = .5, \overline{a}_B = 1.1$; and 3) $\overline{a}_A = 1.1, \overline{a}_B = .5$. In each test, discriminations were uniformly distributed across a range of + - .1. Difficulties and discriminations were randomly paired.

Lower asymptote values were manipulated in four ways: 1) $c_A = c_B = 0$; 2) $c_A = c_B = .2$; 3) $c_A = 0, c_B = .2$, and 4) $c_A = .2, c_B = 0$. In each case, lower asymptote values were the same for all items within a test.

In this study, a complete crossing of all levels produced 36 cells, or cases, of pairs of tests to be equated. All anchor test items had a mean difficulty of zero and a mean discrimination of .8. For vertical equating, the anchor items represented an overlap in difficulty between tests A and B. Lower asymptote values were all zero except in the case where the values were .2 for both tests. In these cases, lower asymptotes were .2 for the anchor test items.

These item parameters were chosen to reflect a typical test equating between tests of either equal or unequal difficulty (to simulate horizontal or vertical equating). The abilities of the examinees was chosen to match that of each test's difficulty, an ideal situation and one found commonly in achievement testing. Each sample of 2,000 examinees was selected from a normal distribution with a mean equal to the mean difficulty of the test and a standard deviation of one. The GGNML (IMSL, 1980) generator was used to generate ability parameters for each sample.

Equating Methods

One linear, one equipercentile, and two item response theory (IRT) equating methods were chosen for this study on the basis of their popularity. In all cases, an external anchor test design was used, and Test B was equated to Test A. That is, for each raw score on Test B, an equivalent was found on the raw score scale of Test A. For vertical equating, Test B was always the more difficulty test.

The linear equating method has been described by Angoff (1971) as Design IVC-1 and is a procedure derived by Levine (1955) for equally reliable tests. Equipercentile equating was accomplished using Levine's (1958) method which has been described by Angoff as Design IVB. Cureton & Tukey's (1951) rolling weighted average method was used to smooth the cumulative distributions.

One of the IRT equating methods is based on the Rasch model. Item parameters were estimated using BICAL (Wright, Mead, & Bell, 1980), and the equating was done using procedures outlined by Wright & Stone (1979).

For three-parameter model equating, parameter estimates were obtained from LOGIST V (Wingersky, Barton, & Lord, 1982). Many versions of this program exist. For this study, a version adapted by ETS for a UNIVAC 1100 was used. For each equating case, item parameters for both tests and the anchor test were estimated simultaneously by employing LOGIST's option for not reached items. The equating then followed Lord's (1980) estimated true score equating procedure. For below chance raw scores, Lord's (1980, p. 210) method of linear extrapolation was used.

Analysis Procedures

Since the data were generated from a known three-parameter model, these initial item parameters were used to develop a criterion for the test equating cases. This criterion was simply a pairing of raw scores corresponding to the same ability estimates:

$$\eta_A = \sum_i P_i(\theta) , \xi_B = \sum_j P_j(\theta)$$

This equating function was then compared to the equating functions produced by the four equating methods. Besides plotting these results, two summary statistics were used to interpret the results. These statistics are very similar to mean square error statistics used in other equating studies, (e.g. Marco, Petersen, &

Stewart, 1979; Petersen, Cook, & Stocking, 1981). These indices are referred to here as the weighted and unweighted mean square error (MSE) and can be stated as follows:

$$\text{weighted(MSE)} = \sum_{i}^{k-1} f_i (X_E - X_{crit})^2 / \sum_{i}^{k} f_i S_B^2$$

$$\text{unweighted(MSE)} = \sum_{i=1}^{k-1} (X_E - X_{crit})^2 / S_B^2$$

where k equals the number of items on Test B, S_B^2 equals the raw score variance for Test B, X_{crit} is the criterion test score equivalent on Test A for raw score i on Test B, X_e is the Test A equivalent for raw score i that is produced by one of the equating methods, and f_i is the frequency of raw score i on Test B. The summation is over raw score values, except that for the weighted MSE, the summation is only across that part of the scale where extrapolation was not necessary. Zero and perfect scores were excluded from all IRT equatings, but included in both conventional equatings.

RESULTS

Raw score means and standard deviations for all data sets are shown in Table 1. Raw score means ranged from approximately 17.5 to 21.3 and standard deviations from

5.0 to 7.7. By looking at data sets generated under similar item parameters, it is clear that the generation procedure produced very consistent results. An examination of the frequency distributions for each data set also revealed a high degree of consistency in the shapes of the distributions. This in turn suggests that there was a high degree of stability in the equatings.

As expected, a higher degree of item discrimination in the generating item parameters produced more dispersion in the raw score distributions. The reverse was true for low discrimination. Non-zero lower asymptotes produced negatively skewed raw score distributions.

The summaries of the two mean square error indices are presented in Tables 2 and 3. The first case, where test difficulties and discriminations were equal and lower asymptotes were zero, was a situation where the data fit the Rasch model. From a psychometric point of view, too, this represented an ideal (easy) equating situation. The best result for all four methods, almost perfect equating, was found in this situation. The worst resuts for all methods occurred where mean test difficulties and discriminations were unequal, where levels of chance scoring were unequal, and where low discrimination was paired with non-zero lower asymptotes on the more difficult test.

In general, the error indices for the equipercentile method were the lowest across all cases. This was followed by the three-parameter model. Values for the Rasch-model and linear equating tended, naturally, to be relatively large in situations where their assumptions were violated.

To aid in the interpretation of Tables 2 and 3 , repeated measures analyses-of-variance were performed on the two MSE indices. The results for the weighted MSE appear in Table 4 and for the unweighted MSE in Table 5. All effects involving a comparison of the four methods were significant. Figures 1 to 4 show cell plots of all means for the first and second order interactions between the four equating methods and independent variables. In each plot, the values shown are means pooled across the variable(s) not included in the plot.

Finally, the actual equating functions for each case are shown in Figure 5. In each plot, the solid line represents the criterion equating based on the initial item parameters. The four broken lines represent the results from the four equating procedures. The criterion equating in most cases was curvilinear, making linear equating clearly inappropriate. In most of the plots, Rasch and linear equating were the most deviant, while the equipercentile equating line was closest to the criterion, thus visually confirming the MSE values in Tables 2 and 3.

DISCUSSION

From a statistical viewpoint, the robustness with respect to violations of assumptions was tested in this study only for the linear and Rasch equating methods. For equipercentile equating, the results showed how the method responded to a variety of conditions. In the case of the three-parameter model, this study was primarily a test of LOGIST's simultaneous estimation procedure.

Linear Equating

The assumptions of the linear equating model are violated whenever the shapes of the raw score distributions differ for the two tests being equated. This occurred when the mean test discrimination and/or level of lower asymptotes differed between the two tests. The appropriateness of linear equating could be gauged by the degree of curvilinearity in the criterion equating function. The total error for linear equating was the smallest for horizontal equating with equally discriminating tests. Chance scoring did not affect the equating in these cases since the criterion equating function was still linear. Linear equating was clearly inappropriate for all vertical equating cases and for horizontal equating where mean test discriminations were unequal.

Equipercentile Equating

As can be seen in Tables 2 and 3, equating error for equipercentile equating was generally the lowest of all four methods. All MSE values were below .25, and it provided the smallest values in 30 of the 36 cases and in all the vertical equating cases. In the most extreme situations, this method was the only one of the four to produce what we would consider acceptable results. Perhaps one reason for this is that it is only one of the four approaches not based on a model. It is simply the best fit of the data at hand. The issue of cross-validation might be important in some situations, but in this case, our preliminary work (not reported here) shows that the results are very stable.

In this version of equipercentile equating, a total group cumulative distribution was estimated for both tests based on the response of the combined samples to the anchor test items. That this estimation in conjunction with a smoothing routine worked so well was somewhat surprising.

Rasch Model Equating

An examination of the results in Tables 2 and 3 suggests that the Rasch model was not very robust to violations of its assumptions. The first case in those tables and in Figure 5 shows a situation where the data fit the Rasch model for horizontal equating. The Rasch model, as expected, performed extremely well as did the other three methods. In the second case in the tables and in Figure 5, all items had a lower assymptote of .2. Yet, the equating was still quite good for all methods.

In subsequent cases, where the level of chance scoring was unequal in the two tests and where test discriminations were unequal, the Rasch model performed very poorly. In situations where low discrimination was paired with non-zero lower asymptotes the total error was relatively large. An explanation for these results can be found by looking at the estimation and linking procedure. When the BICAL program is faced with a data set, a metric is chosen so that mean difficulty equal to zero and all discriminations equal to one. When BICAL runs are done for two tests with different properties, the resulting metrics are different, and estimated item difficulties for one test are more or less compressed than they should be. The use of an equating constant does not alter the underlying metric, and a bias is introduced into the equating.

That this bias can be severe can be seen in the next to last plot in Figure 5 in the case where low discrimination and chance scoring are both present in the more difficult test. The BICAL estimates for this case revealed a range of difficulty of -4.0 to 3.1 for Test A, but for Test B, the range was -1.5 to 1.2. Both tests were generated with a range of +/-2 logits. Obviously the metrics are quite different.

For vertical equating, where the data for each test fit the Rasch model, Rasch equating produced adequate results where the test difficulties differed by one logit. However, where tests differed by two logits in difficulty, the equating was not as good. One possible reason for this involves the anchoring procedure. Since the anchor items represented an overlap in the difficulty ranges of both tests, these items were very difficult for those taking Test A and very easy for those taking Test B. Consequently, estimation was not as accurate. An anchor test with a wider range of difficulty (see Loyd, 1983) might have alleviated this problem.

In the vertical equating cases, the error introduced by unequal mean discrimination and chance scoring was even more pronounced. Even where the same degree of chance scoring occurred on the two tests, Rasch equating was clearly inadequate. These results therefore corroborate from a different methodological perspective, empirical results that advise against using the Rasch model to equate vertically whenever chance scoring is a possibility. These results also advise against using the Rasch model in any situation when mean test discriminations are unequal.

These problems would be especially difficult to overcome when one is constructing alternate forms from an item bank. To ensure that all tests formed from the bank had the same mean discrimination, all items in the bank would have to have the same discrimination, or a complicated algorithm for item selection be programmed. Similarly, chance scoring would not be a problem only if all items had the same degree of chance scoring and the forms to be equated were of comparable difficulty. This is a difficult task for any test developer.

Three parameter Model Equating

Since the data were generated from a three-parameter model, one would expect three-parameter model estimation and equating to be quite accurate. An examination of the values in Tables 2 and 3 indicates that this was not always the case.

The plots in Figures 1 to 4 suggest that three-parameter equating was relatively unaffected by levels of test difficulty or chance scoring. However, the equating was affected by unequal discrimination. There was also an interaction between unequal discrimination and test difficulty and chance scoring. The greatest errors occurred where the more difficult test also had the lower discrimination and where a higher degree of chance scoring was paired with lower discrimination.

Since the data actually fit the model, the LOGIST estimation procedure as programmed should be held responsible for the success of the equating. In this study, simultaneous estimation was used. A single LOGIST run was used for each test equating by employing the "not reached" option. In every case for this study, the LOGIST estimation converged. However, for unequal discrimination and chance scoring, the program typically took at least 35 stages to converge (For practical reasons, it was decided to extend the limit on the number of stages rather than produce continuation runs).

That LOGIST was unable to recover the initial metric can be illustrated by looking at the parameter estimates from one of the cases. For the situation where the initial parameters for tests A and B were as follows: $\bar{b}_A=-.5$, $\bar{a}_A=1.1$, $c_A=.0$ and $\bar{b}_B=.5$, $\bar{a}_B=.5$, $c_B=.2$; the weighted and unweighted MSE's were .616 and .563, respectively. For Test A, the LOGIST difficulty estimates ranged from -3.14 to 1.72, while the original difficulties ranged from -2.5 to 1.5. However, by linearly transforming the LOGIST estimates to the original metric, the estimations ranged from -4.03 to 2.52. The LOGIST discriminations for Test A ranged from .8 to 1.3 compared to the original 1.0 to 1.2. After transformation, the range becomes .6 to .9.

For Test B, the LOGIST difficulties after transformation ranged from -1.37 to 2.55. The original span was from -1.5 to 2.5. However, the difficulties were poorly estimated for the easiest half of the test. The LOGIST discriminations after transformation ranged from .5 to 1.0, the original range being .4 to .6.

Ironically, the lower asymptotes were estimated reasonably well. The default options were used, and default values were obtained for the easiest items on both tests. Yet, no item had a c-value greater than .1 on Test A, and only six items on Test B had c values less than .1.

Another peculiarity was observed in the LOGIST results across all cases. On each test, parameters for a few items (one to three out of 35) were estimated extremely poorly. These tended to occur more frequently on tests with weaker discriminations. No apparent reason for these outliers could be found as all item responses were generated from the same function. Still, an erroneous decision on the quality of an item could be made from these results.

Clearly, in cases of unequal discrimination, LOGIST was unable to reproduce the original metric, and equating was therefore biased. The differences in discrimination were quite severe in this study, and it is not known how well LOGIST would respond to milder differences. On the other hand, the parameters for this study -- sample size, test length, and ability distribution -- were chosen to yield stable, reproduceable estimates. The results suggest therefore that the use of the simultaneous estimation procedure of LOGIST is questionable in circumstances such as these. Some other method for transforming estimates to the same scale should be considered.

Comments on Analysis Procedures

A review of published equating studies reveals a wide variety of evaluation procedures and summary statistics. The degree to which methodology affected conclusions is not known in these studies. In this study, the weighted mean square error statistic was chosen because it has appeared frequently in the literature (e.g. Marco et al., 1979; Petersen et al., 1981). When the results from these statistics were compared to graphs of the equating functions (Figure 5), the weighted MSE values did not seem to represent some of the cases accurately. This was because there were relatively few persons in the raw score ranges where the greatest equating errors occurred, at the lower end of the distribution.

Because of this, the unweighted MSE was also computed. Because each raw score counted equally with this statistic, the values tended to be higher than for the weighted statistics. In Figures 1 to 4, the weighted MSE values appear on the left hand side and the unweighted values on the right. A comparison of the two sets of plots suggests that the two sets of MSE values turned out to be very similar for equipercentile and three-parameter model methods. For Linear and Rasch model equating, quite different results appeared in some of the plots.

Other abnormalities appear when looking at the plots in Figures 1 to 4. For example, in Figure 1, the symmetry in the study's design does not appear in th MSE values for levels of the discrimination and lower asymptote independent variable. Tests A and B alternate between mean discriminations of .5 and 1.1 and lower asymptotes of 0 and .2. Yet, the higher MSE values occur when Test A has the higher discrimination. The same thing occurs when there is more chance scoring on Test B than on Test A. If one examines symmetrical designs (third, fourth, fifth, and ninth plots) in Figure 5, the plots appear to be mirror images of one another.

The paradox can be explained by the fact that the MSE statistic uses vertical distances from the plots. If horizontal distances were used (i.e. Test A equated to Test B), the pattern of results would be reversed. This analysis calls into question the use of MSE statistics for this purpose. A great deal of theoretical statistical work is needed in the area of proper error indexing.

CONCLUSIONS

The purpose of this study was to examine how four commonly used test equating procedures would respond to situations in which the properties of the two tests being equated were different. The results indicated that equipercentile equating was very stable across the cases studied. Linear and Rasch model equating were very sensitive to violations of their models' assumptions. Rasch model equating showed robustness only for horizontal equating where the degree of chance scoring was the same for both tests.

When data fit the Rasch model, three-parameter model equating and Rasch equating achieved comparable and accurate results. In all other cases, three-parameter equating was far better than the Rasch model but generally not as good as equipercentile equating. The results for three-parameter model equating were disappointing since the data were generated from a three-parameter model.

Simultaneous estimation using LOGIST seemed unable to recover the original metric, especially when mean test discriminations were unequal.

All of the equating methods were affected by some situations. Where the tests being equated differed in difficulty, mean discrimination, and in their degree of chance scoring, the equating error was the largest for all four methods. This suggests that equating tests should not be attempted under such extreme conditions. None of the equating methods could completely overcome the effect of such divergence in item type.

The use of the MSE statistics produced several paradoxes in the results. These could be resolved by examining the equating functions themselves. Certainly, more statistical work needs to be done in the comparison of test characteristic curves.

Finally, all of the data for this study were generated from a unidimensional three-parameter model. Real data do not exactly conform to this model, although it seems reasonable in a wide variety of situations. How these methods would respond to multidimensional data is not known, but problems for both IRT methods were uncovered in the unidimensional case.

This study supported other research findings that found the Rasch model inappropriate for use in vertical equating situations. The three-parameter model procedure used here also did not generally produce acceptable results in more complex situations where we might have expected it to do so. The best advice at this point would seem to be to pay very close attention to the properties of the test items. If the tests differ very much in their properties, then classic equipercentile equating is suggested.

REFERENCES

Angoff, W.H. (1971) Scales, Norms and Equivalents Scores. In R. L. Thorndike (ed.). Educational Measurement (2nd ed.). Washington, D.C.: American Council on Education.

Cureton, E.E. & Tukey, J.W. (1951) Smoothing frequency distributions, equating tests, and preparing norms. American Psychologist, 6, 404. (abstract)

Divgi, D.R. (1981) Does the Rasch model really work? Not if you look closely. Paper presented at annual meeting of American Educational Association, Los Angeles.

Forsyth, R., Saisangjan, U., & Gilmer, J. (1981) Some empirical results related to the robustness of the Rasch model. Applied Psychological Measurement, 5, 175-186.

Guskey, T.R. (1981) Comparison of a Rasch model scale and the grade-equivalent scale for vertical equating of test scores. Applied Psychological Measurement, 5, 187-201.

Holmes, S.E. (1982). Unidimensionality and vertical equating with the Rasch model. Journal of Educational Measurement, 19, 139-147.

Holmes, S.E. & Doody-Bogan, E.N. (1983) An empirical study of vertical equating methods using the three-parameter logistic model. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal.

Hulin, C.L., Lissak, R.I., & Drasgow, F. (1982) Recovery of two- and three-parameter Logistic item characteristic curves: a Monte Carlo study. Applied Psychological Measurement, 6, 249-260.

IMSL, Inc. (1980) International Mathematical and Statistical Libraries Reference Manual. Houston, TX: IMSL, Inc.

Kolen, M.J. (1981) Comparison of traditional and item response theory methods for equating tests. Journal of Educational Measurement, 18, 1-11.

Levine, R.S. (1955) Equating the score scales of alternate forms administered to samples of different ability. Research Bulletin No. 23. Princeton, NJ: Educational Testing Service.

Levine, R.S. (1958) Estimated national norms for the Scholastic Aptitude Test. Statistical Report No. 1. Princeton, N.J.: Educational Testing Service.

Lord, F.M. (1980) Practical application of item response theory. Hillsdale, NJ: Lawrence Erlbaum.

Loyd, B.H. (1983) A comparison of the stability of selected vertical equating methods. Paper presented at the annual meeting of the American Educational Research Association, Montreal.

Loyd, B.H. & Hoover, H.D. (1980) Vertical equating using the Rasch model. Journal of Educational Measurement, 17, 179-193.

Macro, G.L., Petersen, N.S., & Stewart, E.E. (1979) A test of the adequacy of curvilinar score equating model. Paper presented at the Computerized Adaptive Testing Conference, Minneapolis.

Petersen, N.S., Cook, L.L., & Stocking, M.L. (1981) IRT versus conventional equating methods: A comparitive study of scale drift. Paper presented at annual meeting of American Educational Research Association, Los Angeles.

Slinde, J.A. & Linn, R.L. (1978) An exploration of the adequacy of the Rasch model for the problem of vertical equating. Journal of Educational Measurement, 15, 23-35.

Slinde, J.A. & Linn, R.L. (1979) A note on vertical equating via the Rasch model for groups of quite different ability and tests of quite different difficulty, Journal of Educational Measurement, 16, 159-165.

Tinsley, H.E. & Dawis, R.V. (1975) An investigation of the Rasch simple logistic model: Sample-free item and test calibration. Educational and Psychological Measurement, 35, 325-339.

Whitely, S.E. & Dawis, R.V. (1974) The nature of objectivity with the Rasch model. Journal of Educational Measurement, 11, 163-178.

Wingersky, M.S., Barton, M.H., & Lord, F.M. (1983) LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters. LOGIST 5, version 1. Princeton, NJ: Educational Testing Service

Wright, B. D., Mead, R. J., & Bell, S. R. (1980) BICAL: Calibrating items with the Rasch model. Research Memorandum No. 23c Chicago: Statistical Laboratory, Department of Education, University of Chicago.

Wright, B. D. & Stone, M.H. (1979) Best test design. Chicago, IL: Mesa Press.

Table 1

Raw Score Means and Standard Deviations for Generated Data

Test A					Test B				
b	a	c	Mean	S.D.	b	a	c	Mean	S.D.
0	.8	.0	17.54	7.23	0	.8	.0	17.58	7.27
0	.8	.2	21.08	5.89	0	.8	.2	21.15	6.10
0	.8	.0	17.79	7.12	0	.8	.2	20.95	6.00
0	.8	.2	21.25	5.98	0	.8	.0	17.67	6.99
0	.5	.0	17.31	6.13	0	1.1	.0	17.16	7.72
0	.5	.2	20.99	4.98	0	1.1	.2	21.04	6.24
0	.5	.0	17.79	6.00	0	1.1	.2	21.08	6.38
0	.5	.2	20.85	5.09	0	1.1	.0	17.38	7.50
0	1.1	.0	17.48	7.55	0	.5	.0	17.66	6.03
0	1.1	.2	20.83	6.36	0	.5	.2	21.14	5.09
0	1.1	.0	17.62	7.62	0	.5	.2	20.99	5.02
0	1.1	.2	21.29	6.36	0	.5	.0	17.63	5.90
-.5	.8	.0	17.54	7.08	.5	.8	.0	17.42	7.05
-.5	.8	.2	20.92	5.94	.5	.8	.2	20.99	5.96
-.5	.8	.0	17.57	7.21	.5	.8	.2	21.01	5.84
-.5	.8	.2	21.04	5.87	.5	.8	.0	17.39	7.17
-.5	.5	.0	17.63	5.93	.5	1.1	.0	17.41	7.60
-.5	.5	.2	21.16	5.07	.5	1.1	.2	20.98	6.37
-.5	.5	.0	17.64	6.04	.5	1.1	.2	21.16	6.27
-.5	.5	.2	20.97	5.10	.5	1.1	.0	17.55	7.70
-.5	1.1	.0	17.51	7.40	.5	.5	.0	17.58	6.07
-.5	1.1	.2	21.03	6.32	.5	.5	.2	21.07	4.98
-.5	1.1	.0	17.49	7.60	.5	.5	.2	20.63	5.15
-.5	1.1	.2	20.88	6.42	.5	.5	.0	17.39	6.09
-1.0	.8	.0	17.41	7.00	1.0	.8	.0	17.37	7.19
-1.0	.8	.2	20.84	5.90	1.0	.8	.2	20.81	6.02
-1.0	.8	.0	17.26	7.06	1.0	.8	.2	21.04	5.95
-1.0	.8	.2	21.11	5.88	1.0	.8	.0	17.63	7.08
-1.0	.5	.0	17.47	6.12	1.0	1.1	.0	17.68	7.49
-1.0	.5	.2	21.00	5.14	1.0	1.1	.2	20.79	6.32
-1.0	.5	.0	17.43	6.05	1.0	1.1	.2	20.85	6.41
-1.0	.5	.2	21.14	5.14	1.0	1.1	.0	17.60	7.41
-1.0	1.1	.0	17.69	7.65	1.0	.5	.0	17.63	5.95
-1.0	1.1	.2	20.98	6.46	1.0	.5	.2	20.97	5.15
-1.0	1.1	.0	17.55	7.46	1.0	.5	.2	20.94	5.01
-1.0	1.1	.2	20.88	6.32	1.0	.5	.0	17.40	6.00

Table 2

Unweighted Mean Square Error for Test Equating Cases

$\bar{b}_A = \bar{b}_B = 0.0$

Test A		Test B		CASE			
$\bar{a}$	c	$\bar{a}$	c	LINEAR	EQ % ILE	RASCH	3-PARA
.8	.0	.8	.0	.000	.000	.000	.000
.8	.2	.8	.2	.003	.004	.000	.001
.8	.0	.8	.2	.088	.010	.231	.045
.8	.2	.8	.0	.003	.016	.172	.037
.5	.0	1.1	.0	.037	.043	.501	.148
.5	.2	1.1	.2	.086	.036	.309	.133
.5	.0	1.1	.2	.041	.042	.140	.082
.5	.2	1.1	.0	.028	.039	.890	.204
1.1	.0	.5	.0	.094	.074	.854	.271
1.1	.2	.5	.2	.195	.079	.422	.214
1.1	.0	.5	.2	.705	.132	1.955	.563
1.1	.2	.5	.0	.060	.041	.158	.111
$\bar{b}_A = -.5;\ \bar{b}_B = .5$							
.8	.0	.8	.0	.227	.009	.054	.050
.8	.2	.8	.2	.369	.055	.195	.043
.8	.0	.8	.2	.391	.023	.826	.120
.8	.2	.8	.0	.183	.008	.120	.058
.5	.0	1.1	.0	.160	.040	.494	.199
.5	.2	1.1	.2	.479	.069	.210	.133
.5	.0	1.1	.2	.387	.078	.370	.157
.5	.2	1.1	.0	.128	.036	.596	.210
1.1	.0	.5	.0	.605	.107	.945	.288
1.1	.2	.5	.2	.427	.095	.982	.270
1.1	.0	.5	.2	.819	.175	2.817	.563
1.1	.2	.5	.0	.426	.040	.352	.131
$\bar{b}_A = -1.0;\ \bar{b}_B = 1.0$							
.8	.0	.8	.0	1.326	.011	.161	.133
.8	.2	.8	.2	1.693	.072	.672	.142
.8	.0	.8	.2	2.383	.112	1.787	.228
.8	.2	.8	.0	1.109	.013	.183	.117
.5	.0	1.1	.0	.916	.059	.443	.275
.5	.2	1.1	.2	1.386	.179	.420	.182
.5	.0	1.1	.2	1.807	.090	1.174	.343
.5	.2	1.1	.0	.740	.059	.442	.212
1.1	.0	.5	.0	3.099	.122	1.288	.319
1.1	.2	.5	.2	2.455	.107	2.000	.477
1.1	.0	.5	.2	3.967	.234	4.585	.478
1.1	.2	.5	.0	2.168	.032	.807	.216

Table 3

Weighted Mean Square Error for Test Equating Costs

$\overline{b}_A = \overline{b}_B = 0.0$

TEST A		TEST B		CASE			
$\bar{a}$	c	$\bar{a}$	c	LINEAR	EQ % ILE	RASCH	3-PARA
.8	.0	.8	.0	.000	.000	.000	.000
.8	.2	.8	.2	.001	.002	.000	.001
.8	.0	.8	.2	.004	.004	.132	.020
.8	.2	.8	.0	.002	.001	.066	.005
.5	.0	1.1	.0	.026	.023	.382	.104
.5	.2	1.1	.2	.028	.023	.256	.119
.5	.0	1.1	.2	.026	.022	.120	.073
.5	.2	1.1	.0	.019	.019	.521	.097
1.1	.0	.5	.0	.086	.085	.632	.256
1.1	.2	.5	.2	.081	.065	.390	.244
1.1	.0	.5	.2	.178	.151	1.430	.560
1.1	.2	.5	.0	.040	.029	.128	.101

$\overline{b}_A = -.5;\ \overline{b}_B = .5$

TEST A		TEST B		CASE			
$\bar{a}$	c	$\bar{a}$	c	LINEAR	EQ % ILE	RASCH	3-PARA
.8	.0	.8	.0	.108	.010	.058	.062
.8	.2	.8	.2	.113	.002	.096	.059
.8	.0	.8	.2	.198	.014	.285	.109
.8	.2	.8	.0	.104	.008	.144	.057
.5	.0	1.1	.0	.132	.031	.444	.190
.5	.2	1.1	.2	.100	.025	.273	.164
.5	.0	1.1	.2	.191	.039	.358	.193
.5	.2	1.1	.0	.111	.024	.396	.160
1.1	.0	.5	.0	.183	.067	.496	.258
1.1	.2	.5	.2	.190	.039	.512	.245
1.1	.0	.5	.2	.417	.166	1.413	.616
1.1	.2	.5	.0	.126	.034	.281	.114

$\overline{b}_A = -1.0;\ \overline{b}_B = 1.0$

TEST A		TEST B		CASE			
$\bar{a}$	c	$\bar{a}$	c	LINEAR	EQ % ILE	RASCH	3-PARA
.8	.0	.8	.0	.870	.014	.140	.151
.8	.2	.8	.2	.775	.008	.194	.155
.8	.0	.8	.2	1.486	.028	.466	.226
.8	.2	.8	.0	.779	.013	.230	.123
.5	.0	1.1	.0	.830	.048	.564	.376
.5	.2	1.1	.2	.557	.028	.358	.278
.5	.0	1.1	.2	1.047	.036	.792	.410
.5	.2	1.1	.0	.661	.038	.480	.272
1.1	.0	.5	.0	1.317	.036	.340	.265
1.1	.2	.5	.2	1.001	.033	.543	.338
1.1	.0	.5	.2	2.136	.093	1.142	.485
1.1	.2	.5	.0	.960	.014	.387	.157

Table 4

Analysis of Variance of Unweighted Mean Square Error

Source	df	ms	F
Among Ss			
Difficulty (DI)	2	5.9676	3140.842
Discrimination (DS)	2	3.6789	1936.263
Lower Asymptote (L)	3	1.7128	901.474
DI x DS	4	.4974	261.789
DI x L	6	.3305	173.947
DS x L	6	.5190	273.158
DI x DS x L	12	.0019	.132
Within Ss			
Method (m)	3	5.2413	363.979
DI x m	6	2.3454	162.875
DS x m	6	.7985	55.451
L x m	9	.4800	33.333
DI x DS x m	12	.2033	14.118
DI x L x m	18	.1141	7.924
DS x L x m	18	.2231	15.493
DI x DS x L x m	36	.0144	

Table 5

Analysis of Variance of weighted mean square error

Source	df	MS	F
Among Ss			
Difficulty (DI)	2	1.345	336.657**
Discrimination (DS)	2	.712	174.510**
Lower Asymptote (L)	3	.412	100.980**
DI x DS	4	.002	.490
DI x L	6	.034	8.333**
DS x L	6	.156	38.235**
DI x DS x L	12	.004	
Within Ss			
Method (m)	3	1.178	136.028**
DI x m	6	.762	87.991**
DS x m	6	.124	14.319**
L x m	9	.055	6.351**
DI x DS x m	12	.041	4.734**
DI x L x m	18	.026	3.002**
DS x L x m	18	.030	3.464**
DS x DI x L x m	36	.009	

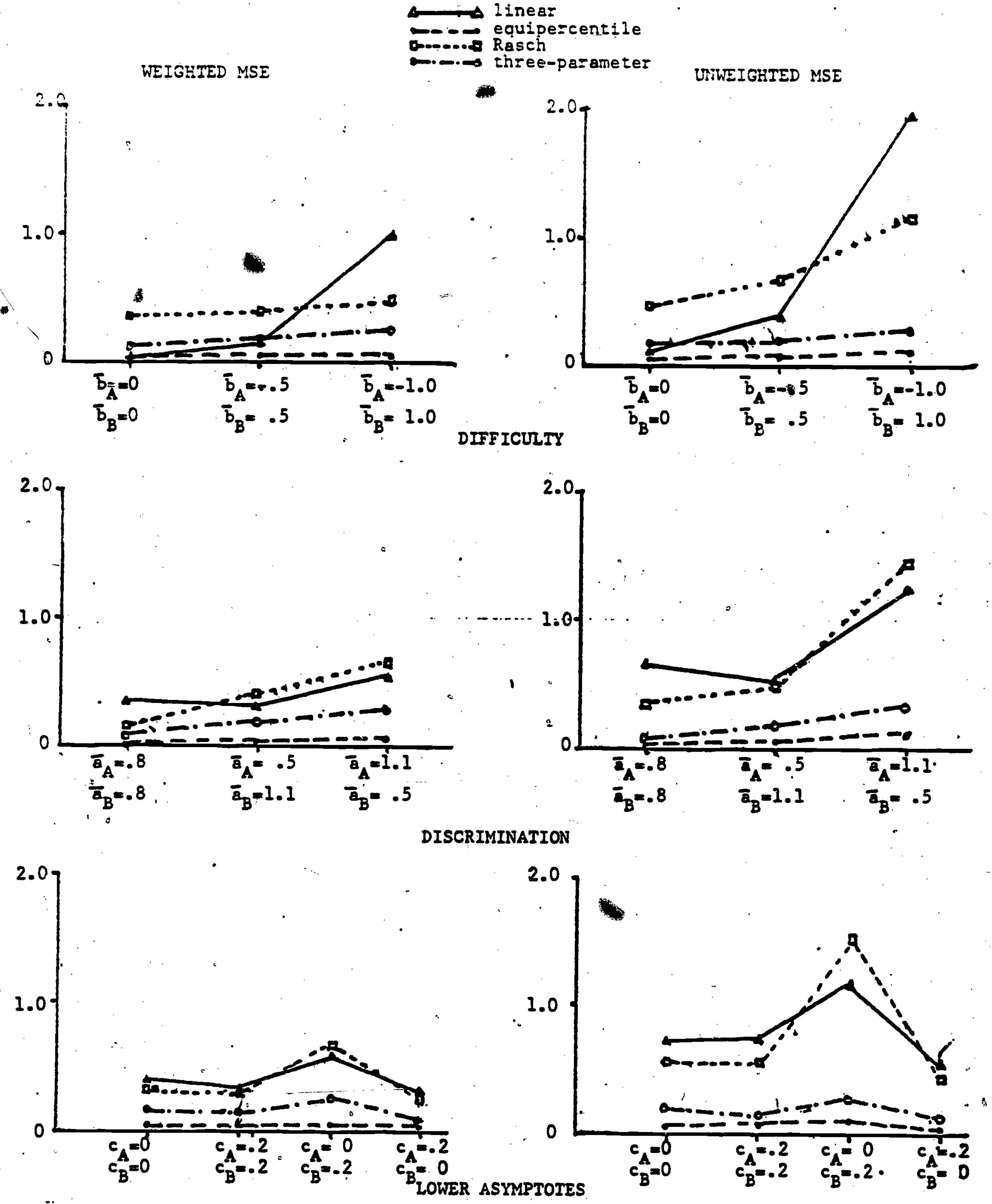


Figure 1. Cellplots of interactions between equating methods and difficulty, discrimination, and lower asymptote

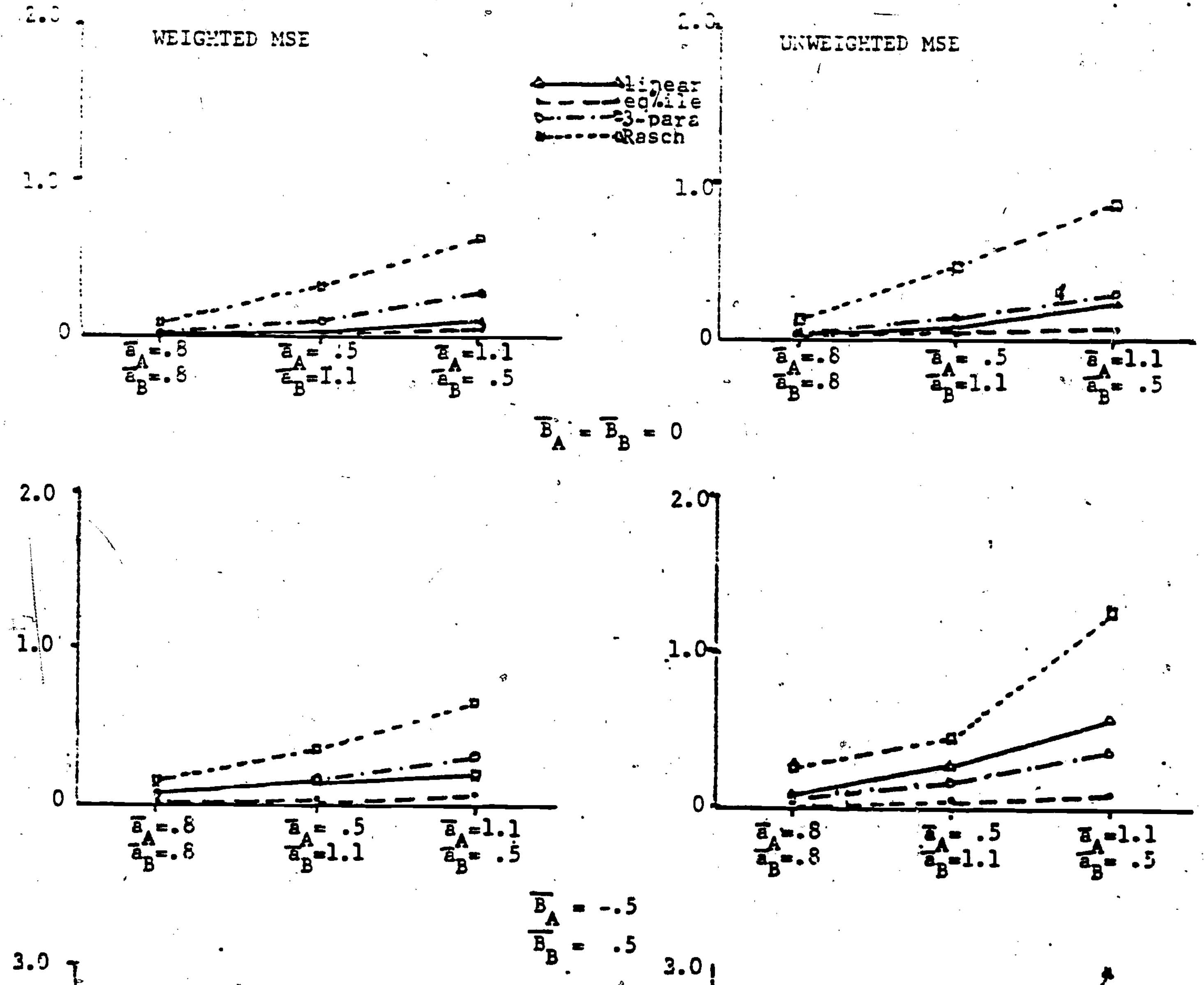


Figure 2. Cellplots of interaction between equating methods, difficulty, and discrimination

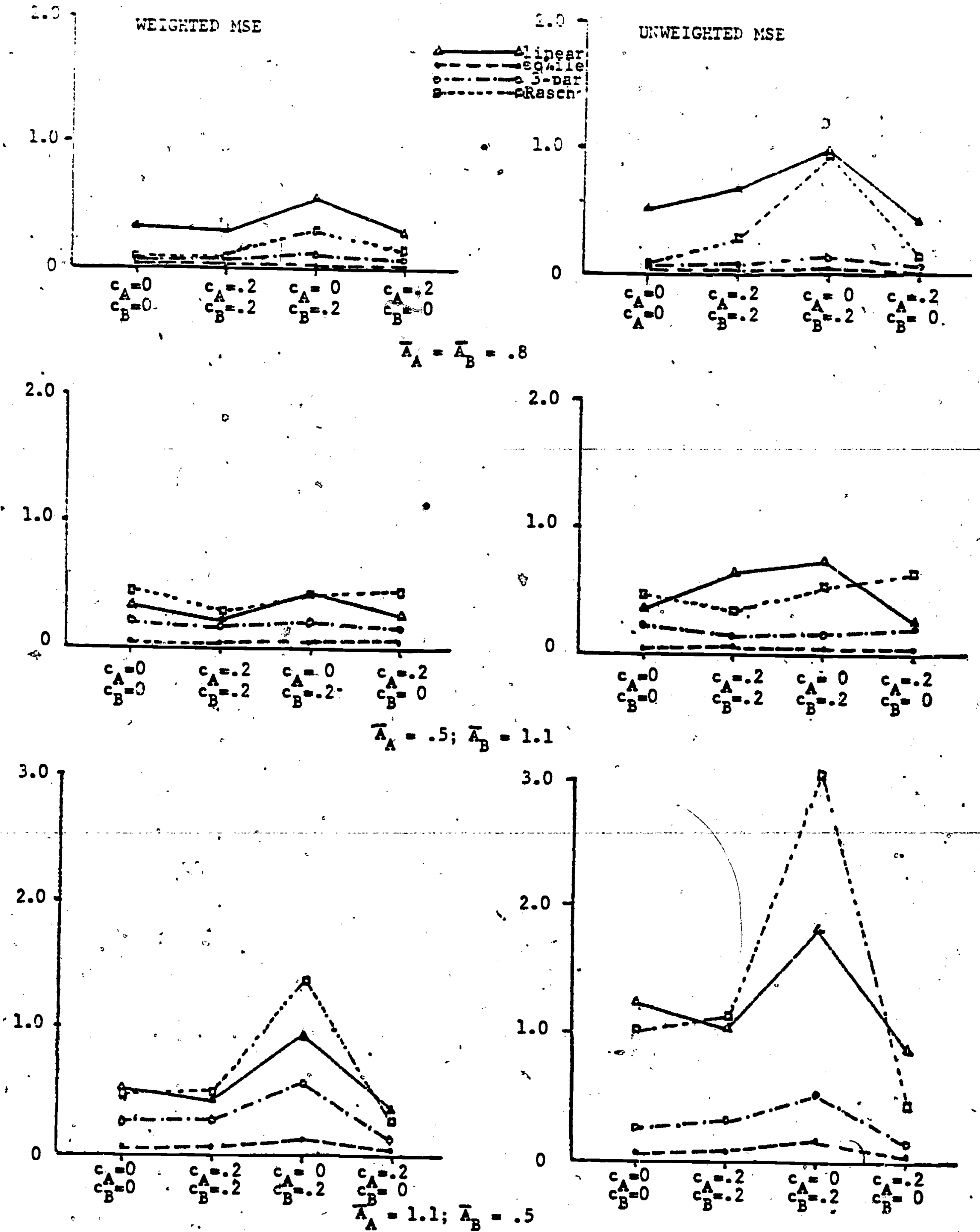


Figure 3. Cellplots of interaction between equating methods, discrimination and lower asymptote

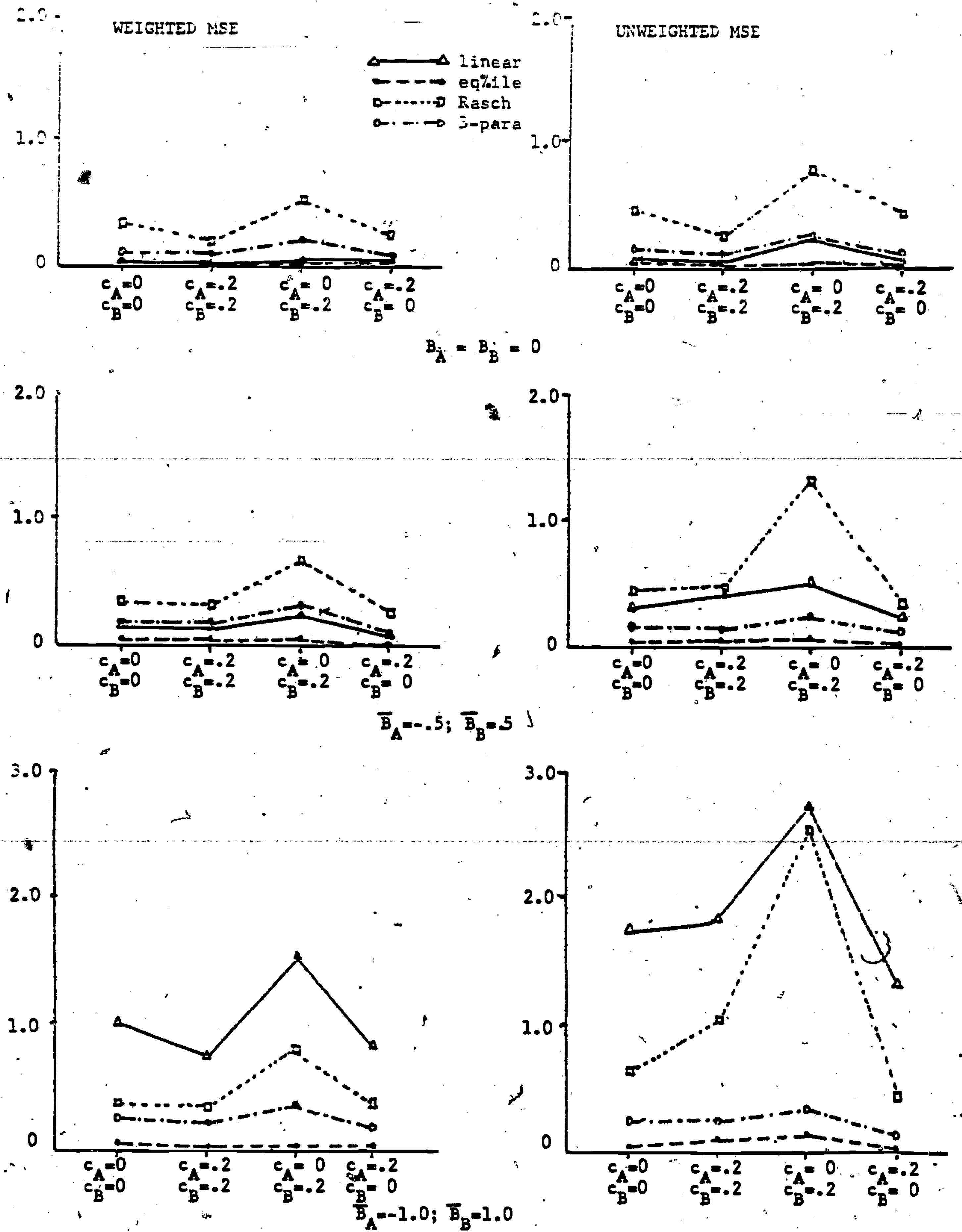
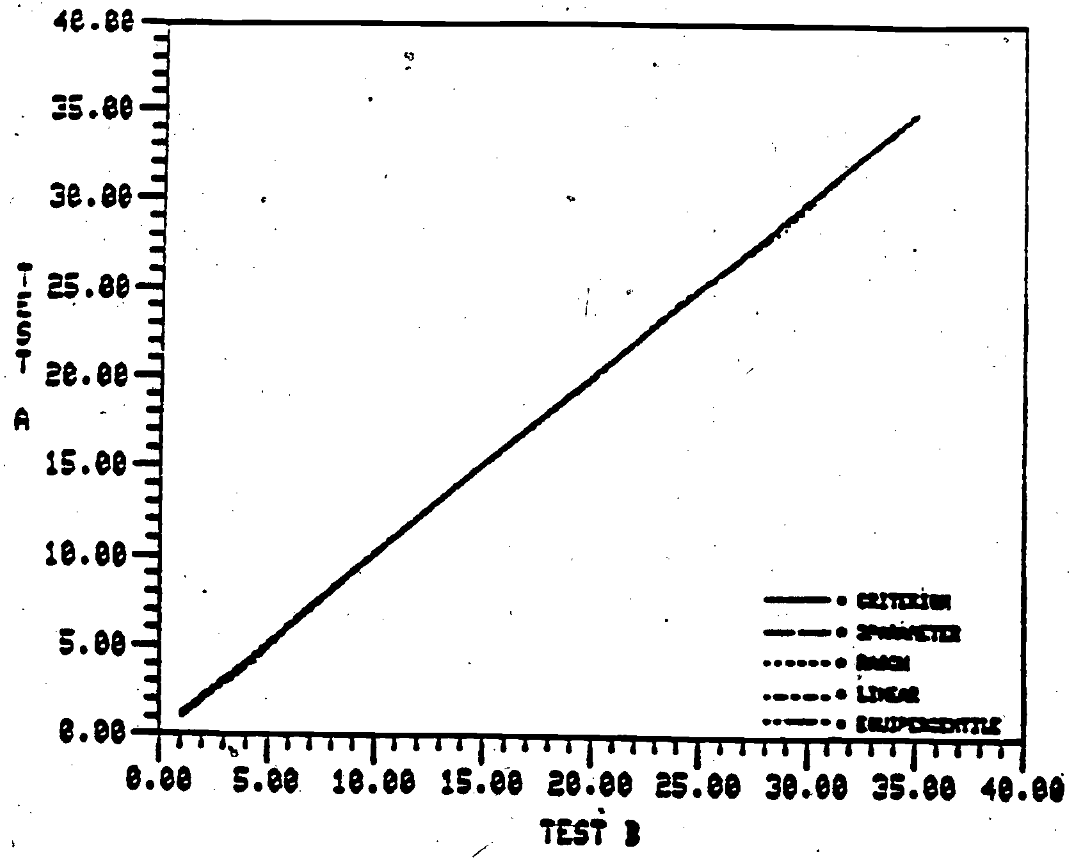


Figure 4. Cellplots of the interaction between equating methods, difficulty, and lower asymptote

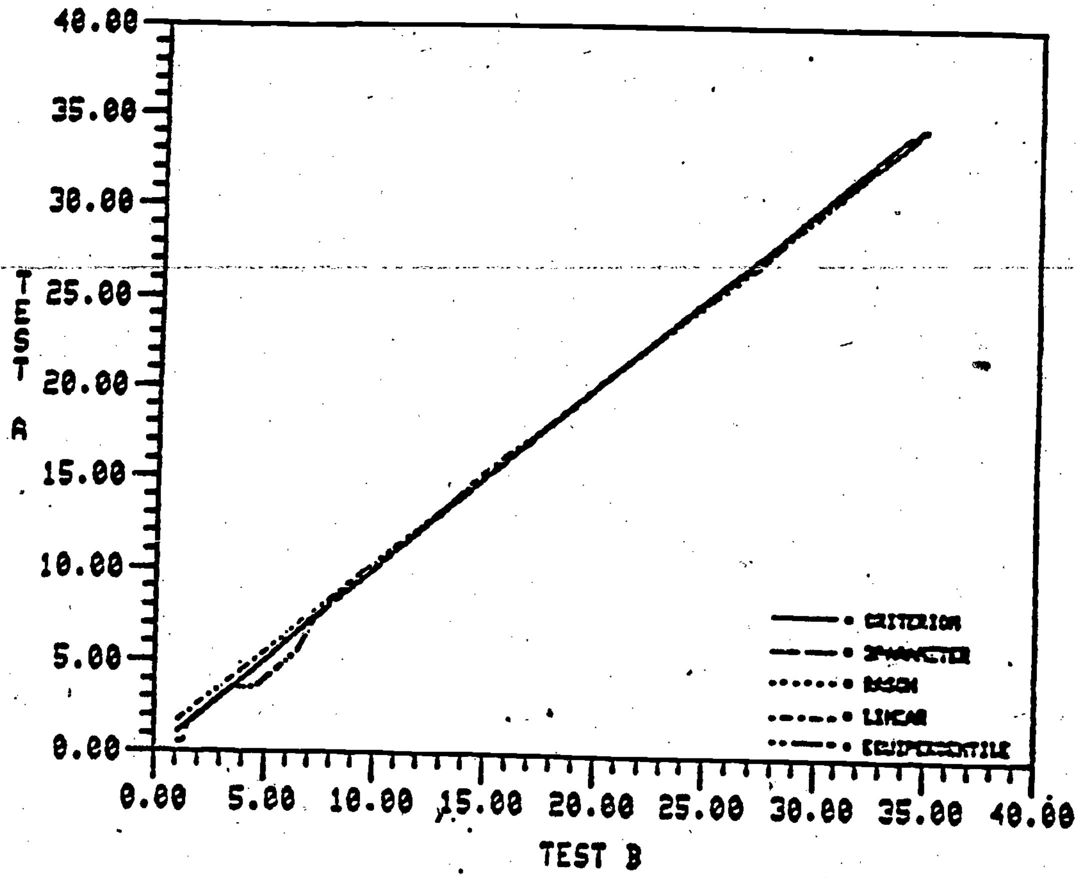
Figure 5

Equating functions for four equating methods

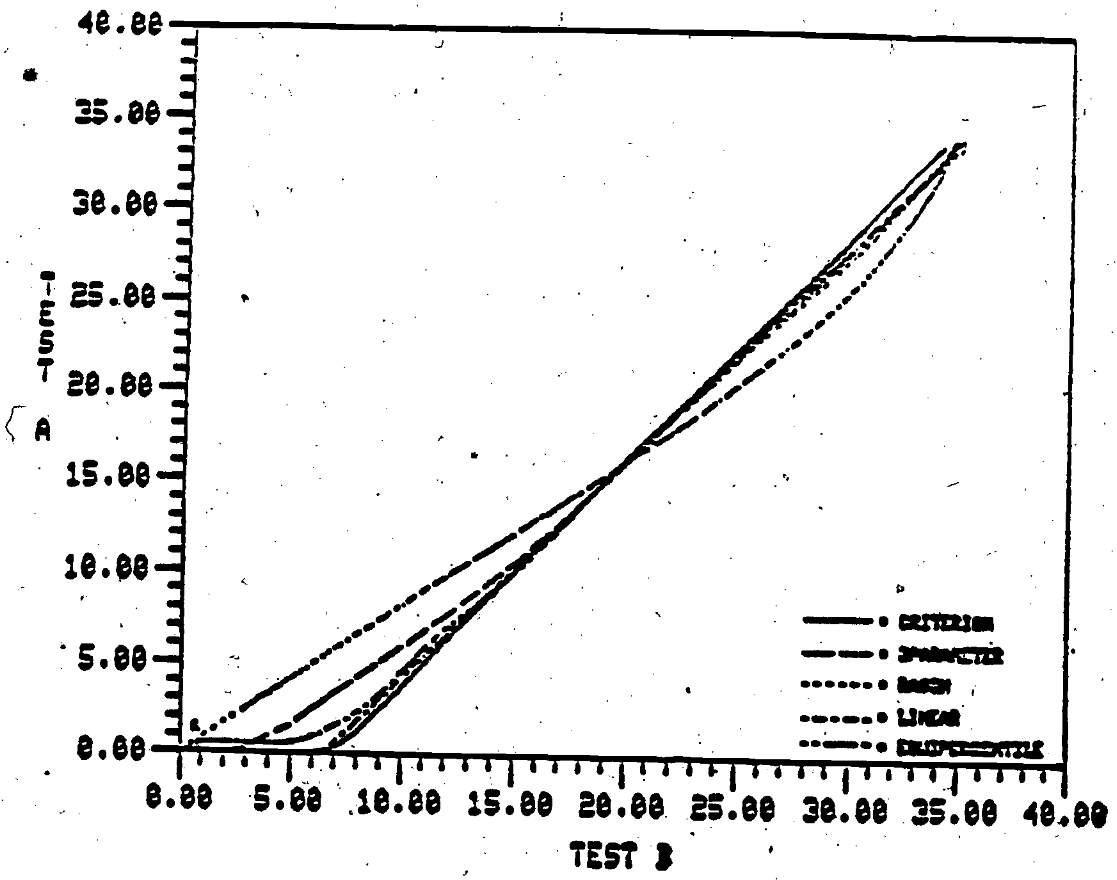
Figures follow.



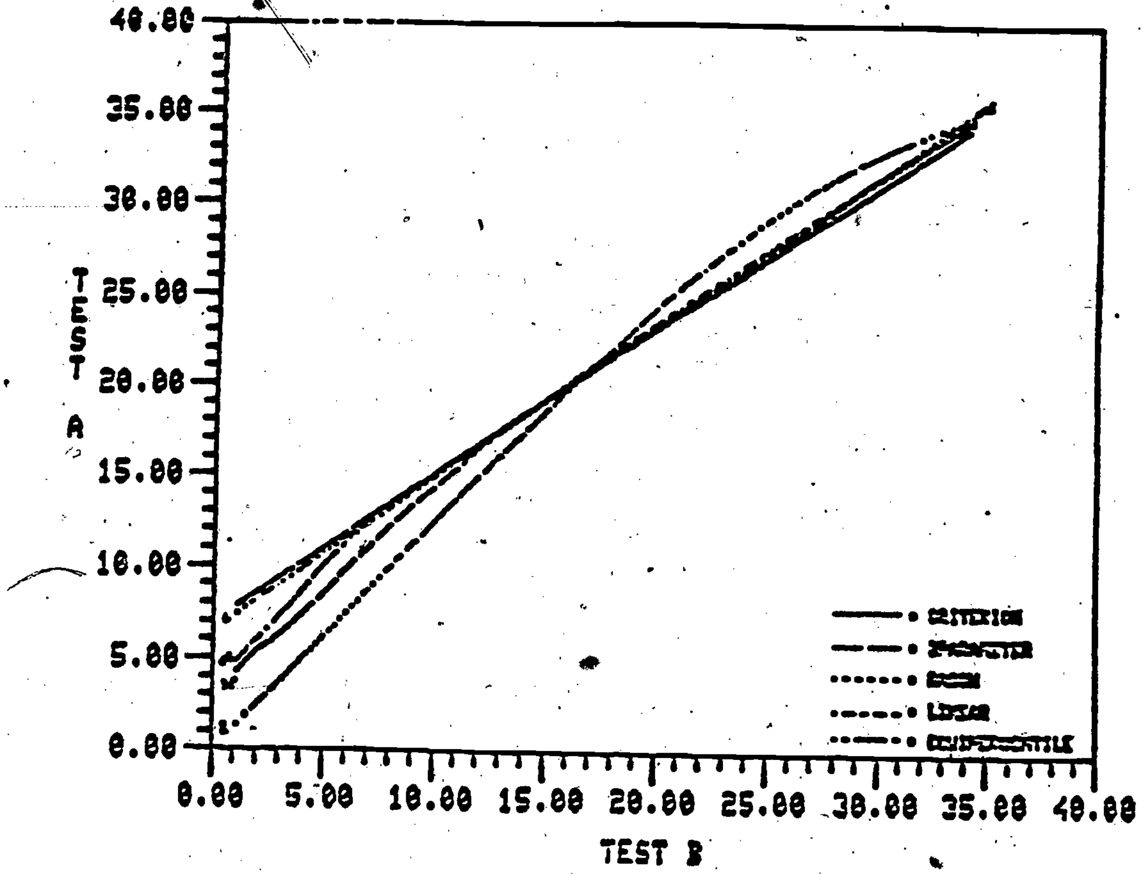
Test A: $\bar{b}$ = 0; $\bar{a}$ = .8; c = .0; $\bar{\theta}$ = 0.0
Test B: $\bar{b}$ = 0; $\bar{a}$ = .8; c = .0; $\bar{\theta}$ = 0.0



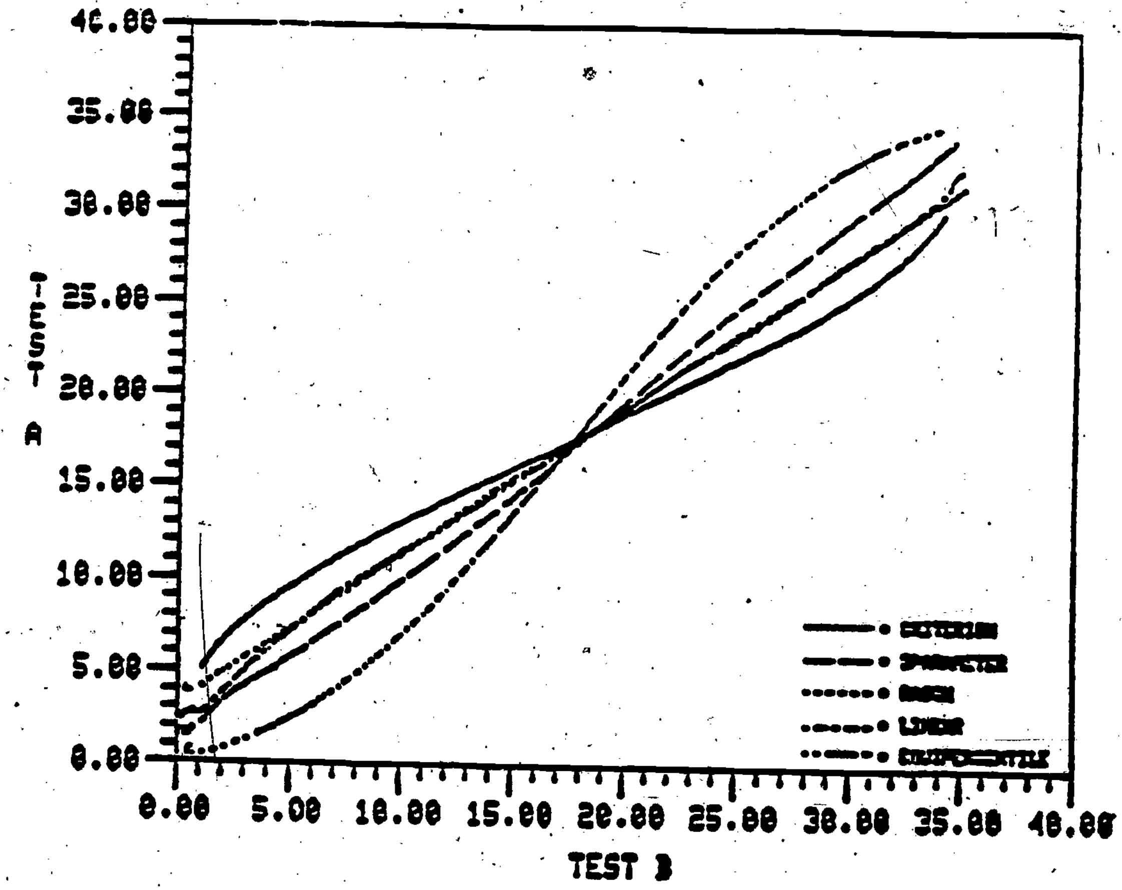
Test A: $\bar{b}$ = 0; $\bar{a}$ = .8; c = .2; $\bar{\theta}$ = 0.0
Test B: $\bar{b}$ = 0; $\bar{a}$ = .8; c = .2; $\bar{\theta}$ = 0.0



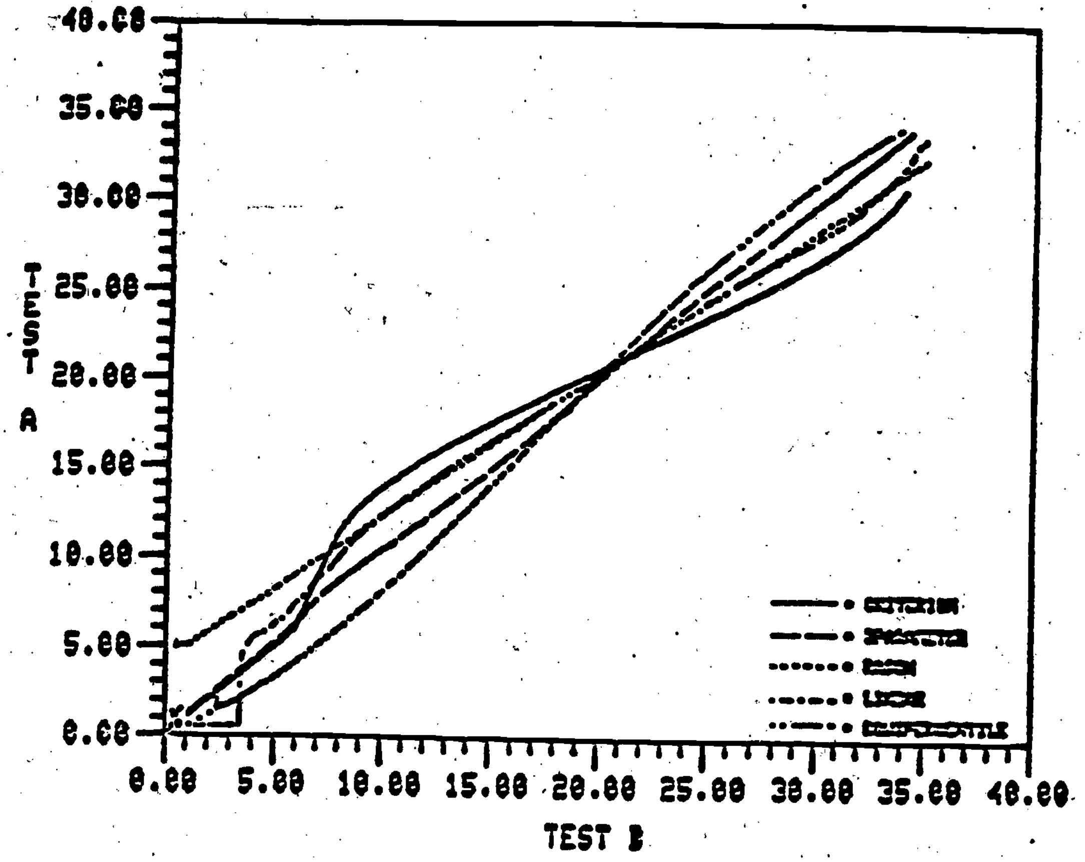
Test A; $\bar{b}$ = 0; $\bar{a}$ = .8; c = .0; $\bar{\theta}$ = 0.0
Test B: $\bar{b}$ = 0; $\bar{a}$ = .8; c = .2; $\bar{\theta}$ = 0.0



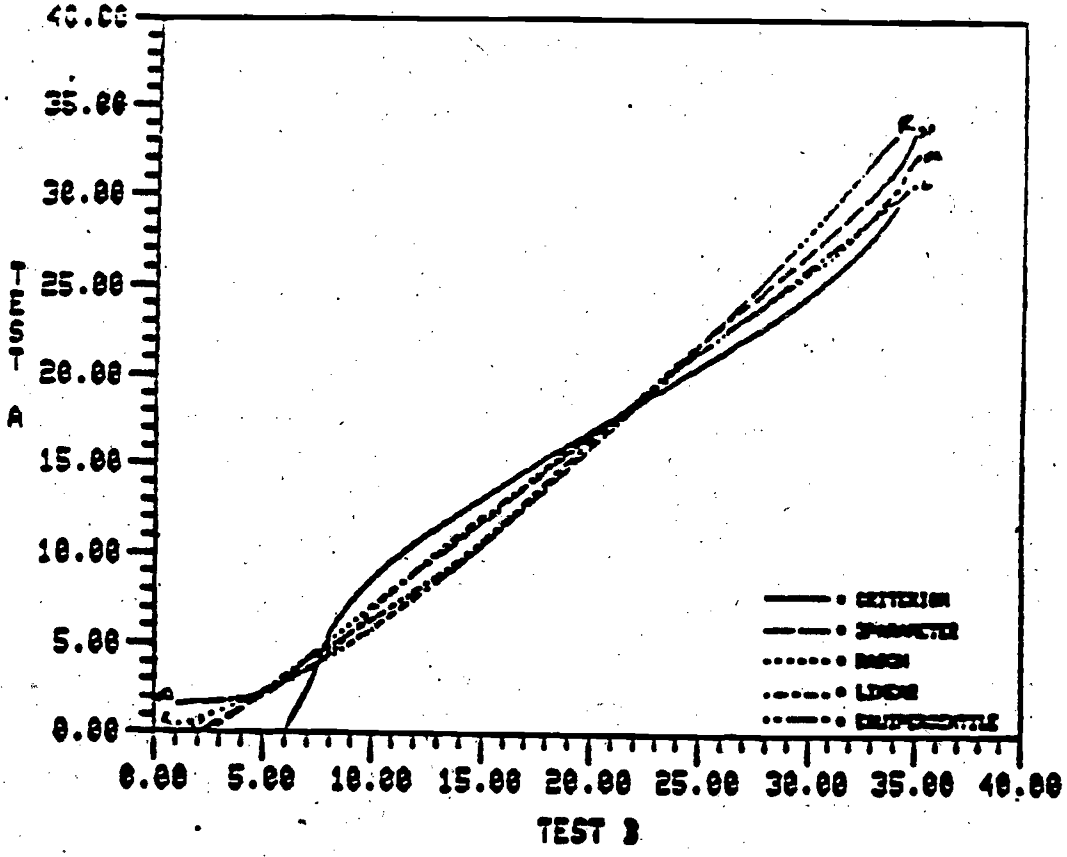
Test A: $\bar{b}$ = 0; $\bar{a}$ = .8; c = .2; $\bar{\theta}$ = 0.0
Test B: $\bar{b}$ = 0; $\bar{a}$ = .8; c = 0; $\bar{\theta}$ = 0.0



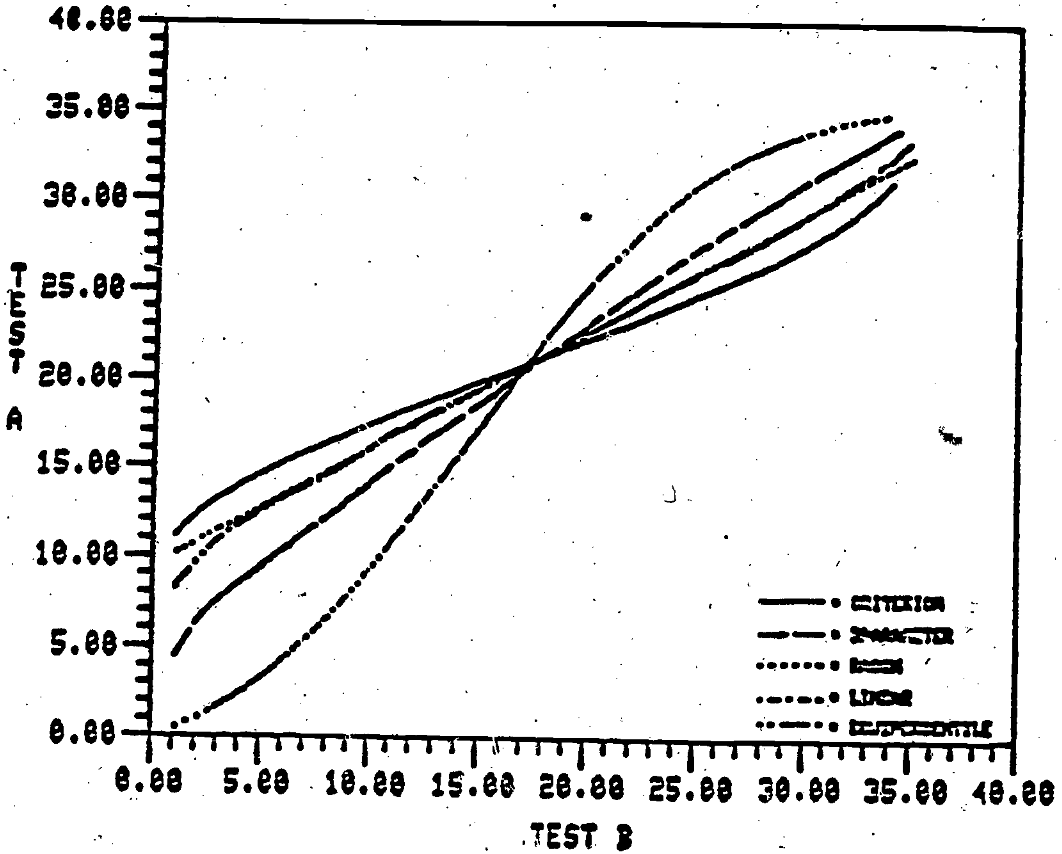
Test A: $\bar{b} = 0$; $\bar{a} = .5$; c = .0; $\bar{\theta} = 0.0$
Test B: $\bar{b} = 0$; $\bar{a} = 1.1$; c = .0; $\bar{\theta} = 0.0$



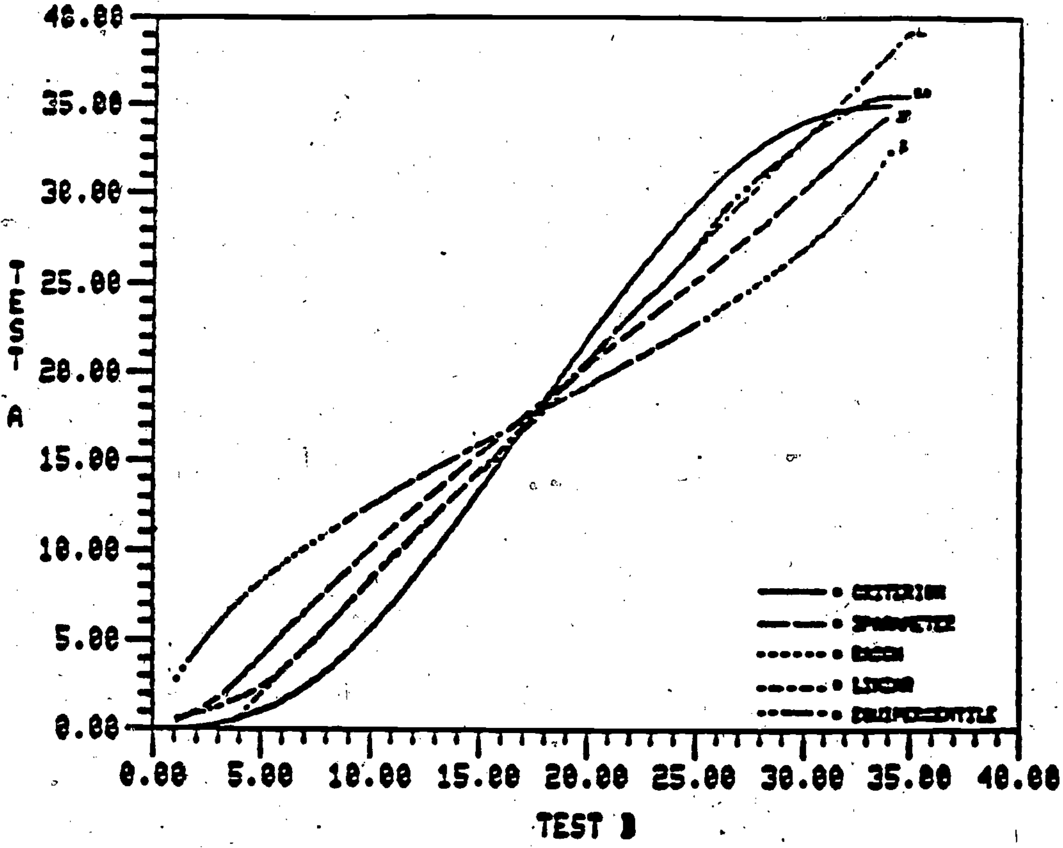
Test A: $\bar{b} = 0$; $\bar{a} = .5$; c = .2; $\bar{\theta} = 0.0$
Test B: $\bar{b} = 0$; $\bar{a} = 1.1$; c = .2; $\bar{\theta} = 0.0$



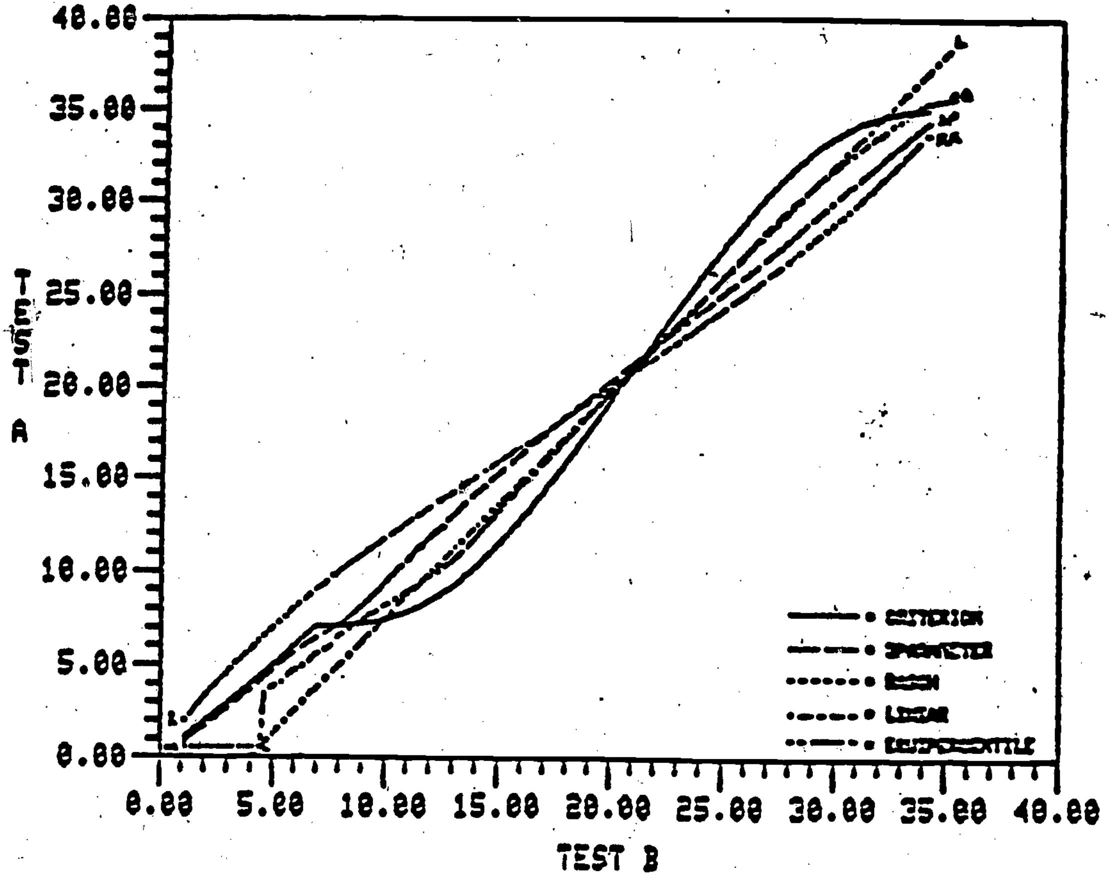
Test A: $\bar{b} = 0$; $\bar{a} = .5$; $c = .0$; $\bar{\theta} = 0.0$
Test B: $\bar{b} = 0$; $\bar{a} = 1.1$; $c = .2$; $\bar{\theta} = 0.0$



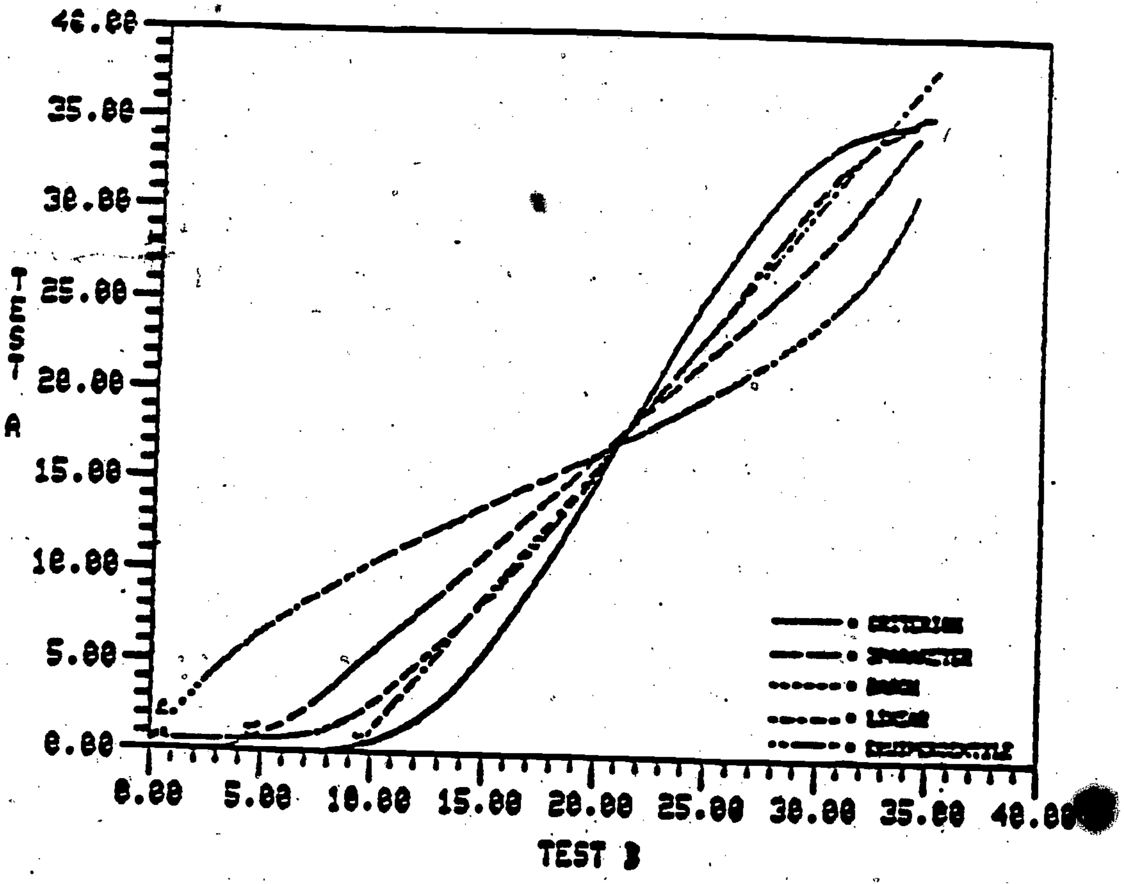
Test A: $\bar{b} = 0$; $\bar{a} = .5$; $c = .2$; $\bar{\theta} = 0.0$
Test B: $\bar{b} = 0$; $\bar{a} = 1.1$; $c = .0$; $\bar{\theta} = 0.0$



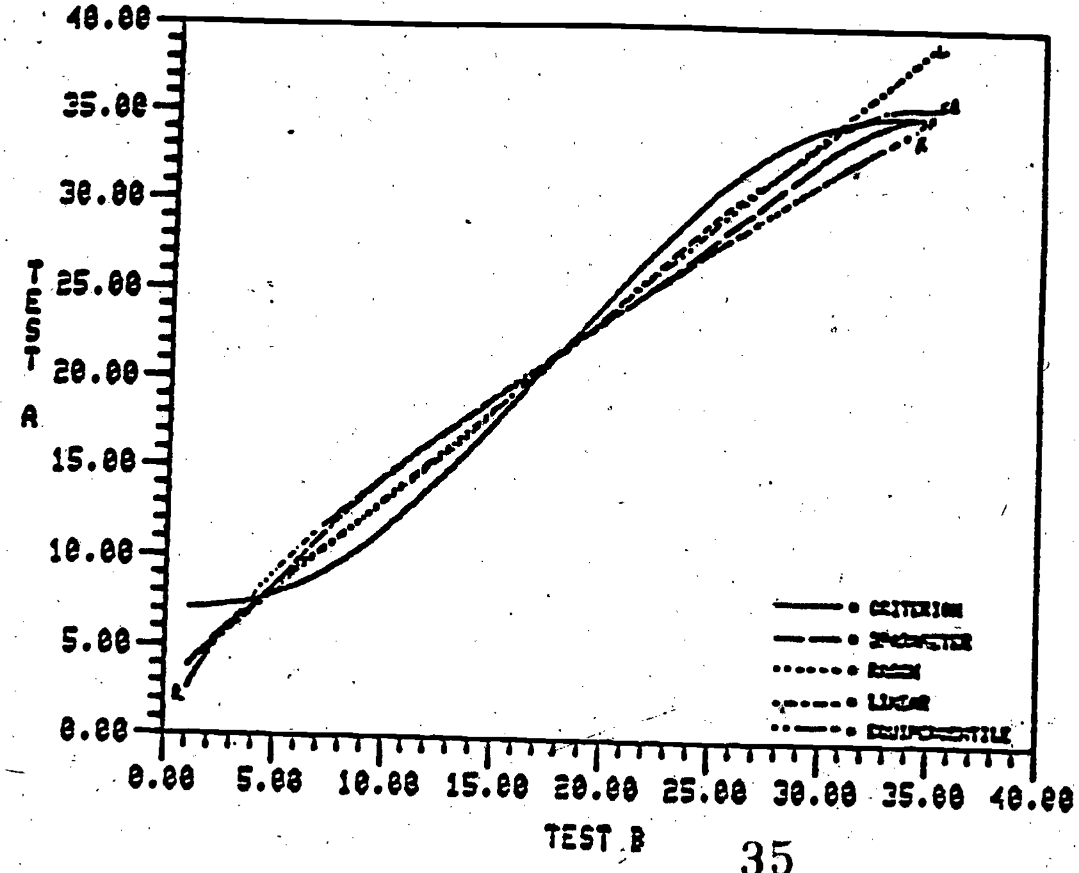
Test A: $\bar{b} = 0$; $\bar{a} = 1.1$; $c = .0$; $\bar{\theta} = 0.0$
Test B: $\bar{b} = 0$; $\bar{a} = .5$; $c = .0$; $\bar{\theta} = 0.0$



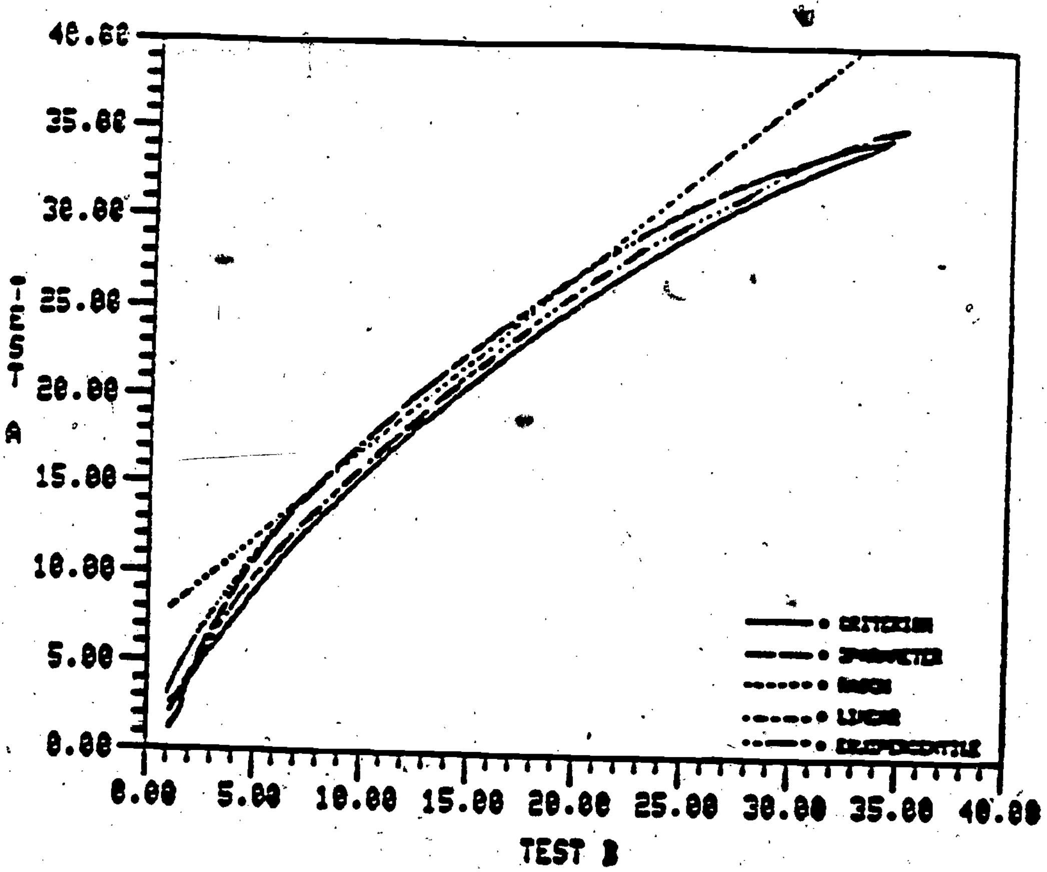
Test A: $\bar{b} = 0$; $\bar{a} = 1.1$; $c = .2$; $\bar{\theta} = 0.0$
Test B: $\bar{b} = 0$; $\bar{a} = .5$; $c = .2$; $\bar{\theta} = 0.0$



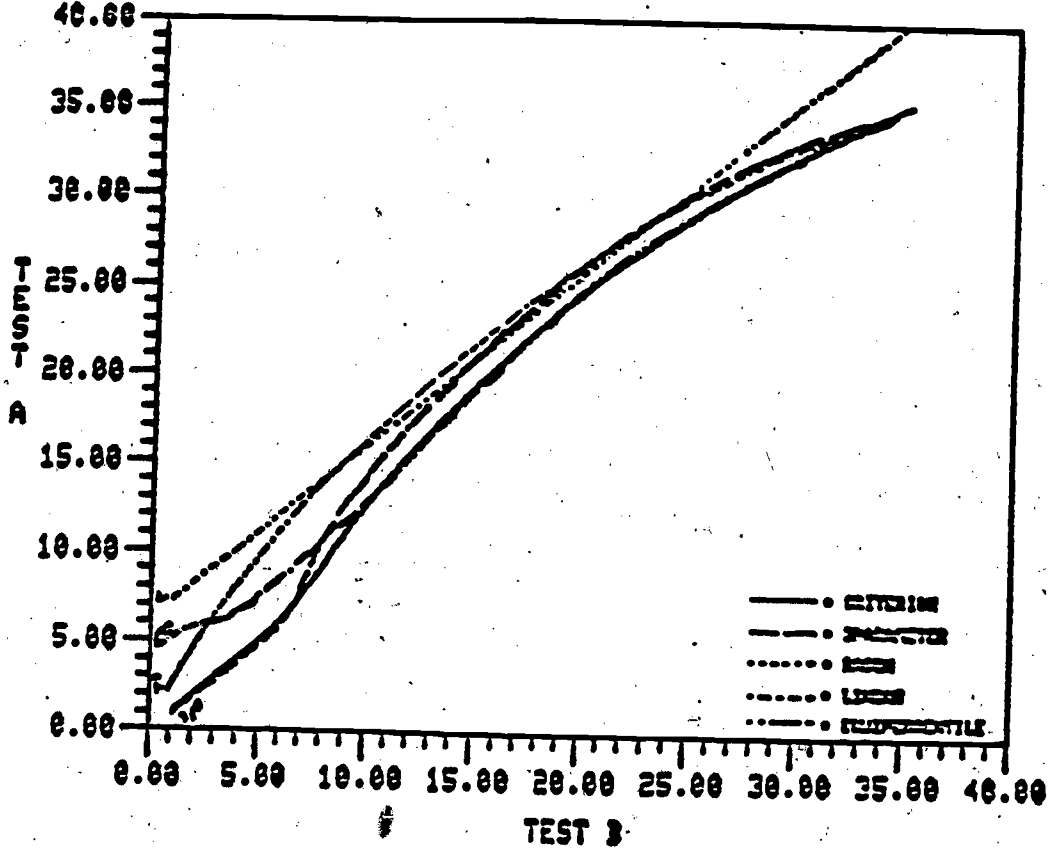
Test A: $\bar{b} = 0$; $\bar{a} = 1.1$; $c = .0$; $\bar{\theta} = 0.0$
Test B: $\bar{b} = 0$; $\bar{a} = .5$; $c = .2$; $\bar{\theta} = 0.0$



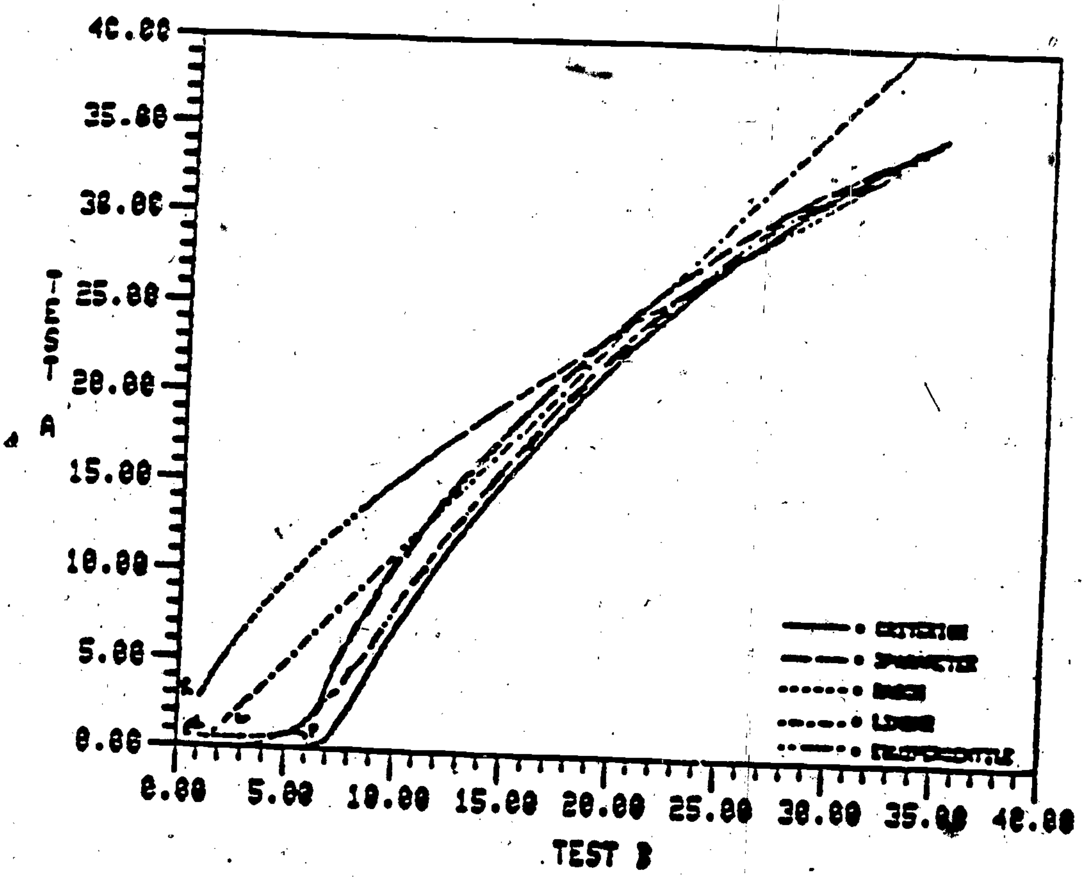
Test A: $\bar{b} = 0$; $\bar{a} = 1.1$; $c = .2$; $\bar{\theta} = 0.0$



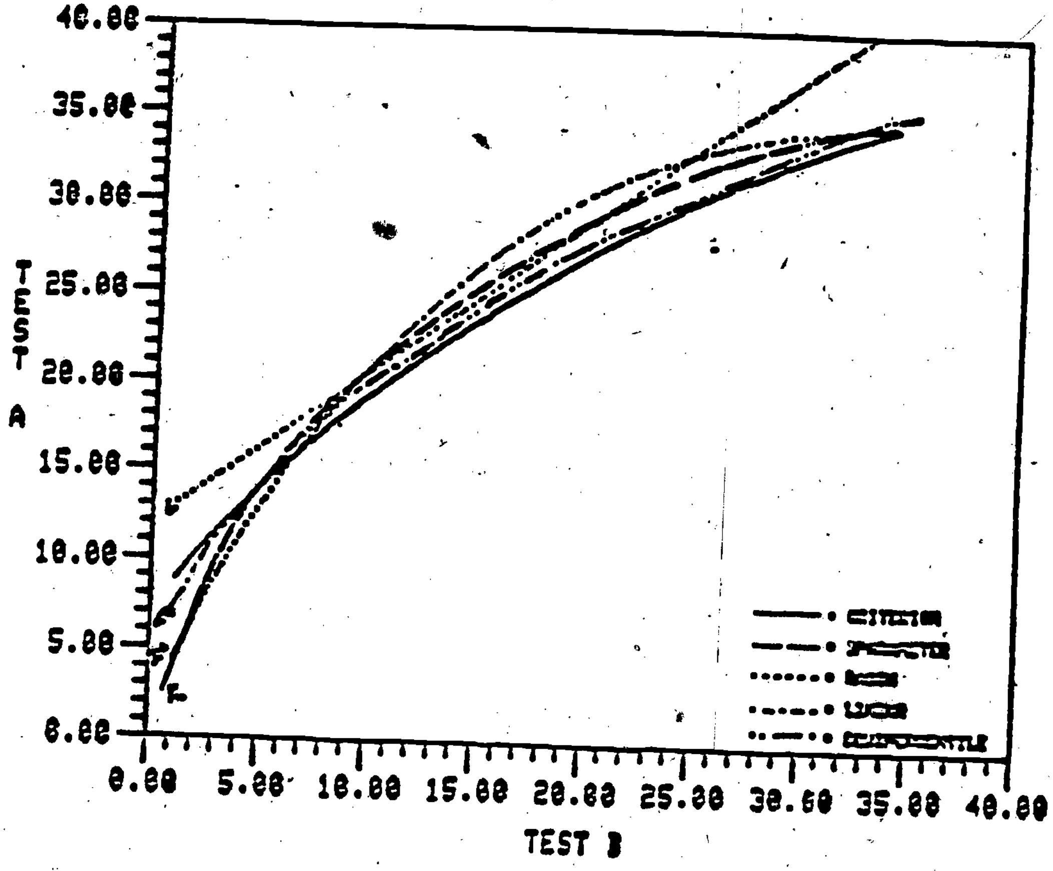
Test A: $\bar{b}$ = -.5; $\bar{a}$ = .8; c = .0; $\bar{\theta}$ = -.5
Test B: $\bar{b}$ = .5; $\bar{a}$ = .8; c = .0; $\bar{\theta}$ = .5



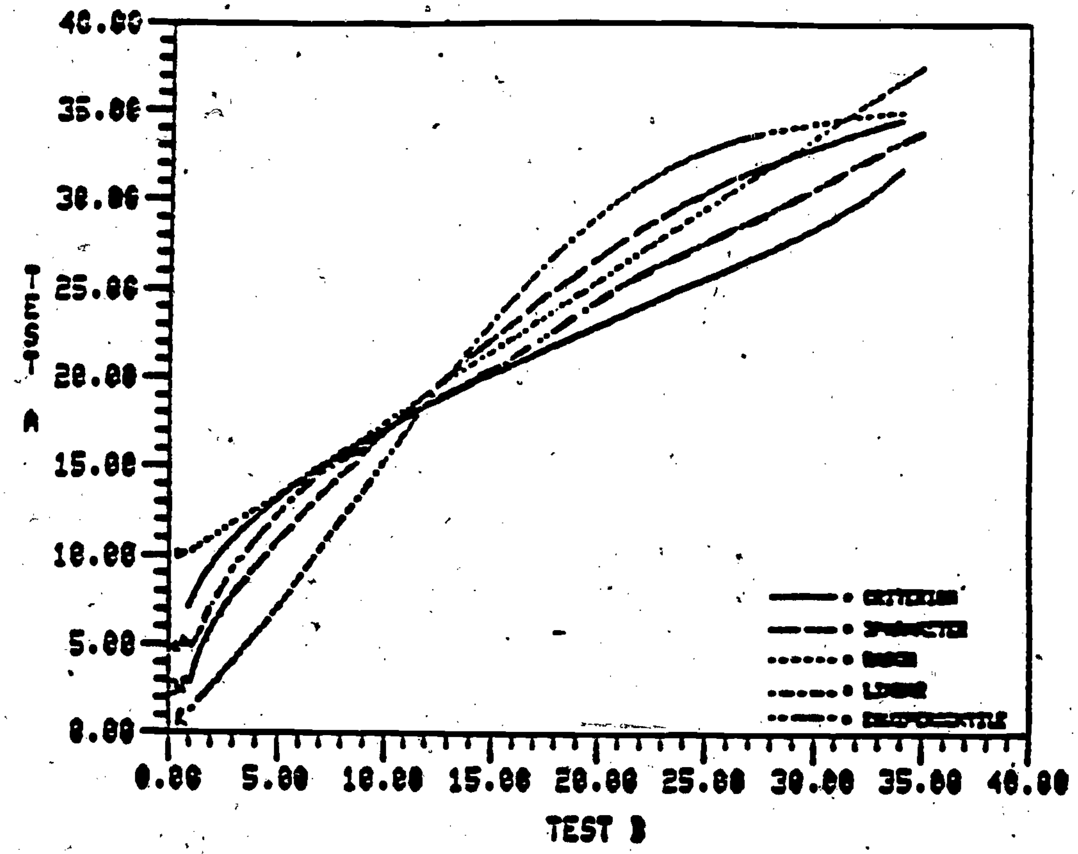
Test A: $\bar{b}$ = -.5; $\bar{a}$ = .8; c = .2; $\bar{\theta}$ = -.5
Test B: $\bar{b}$ = .5; $\bar{a}$ = .8; c = .2; $\bar{\theta}$ = .5



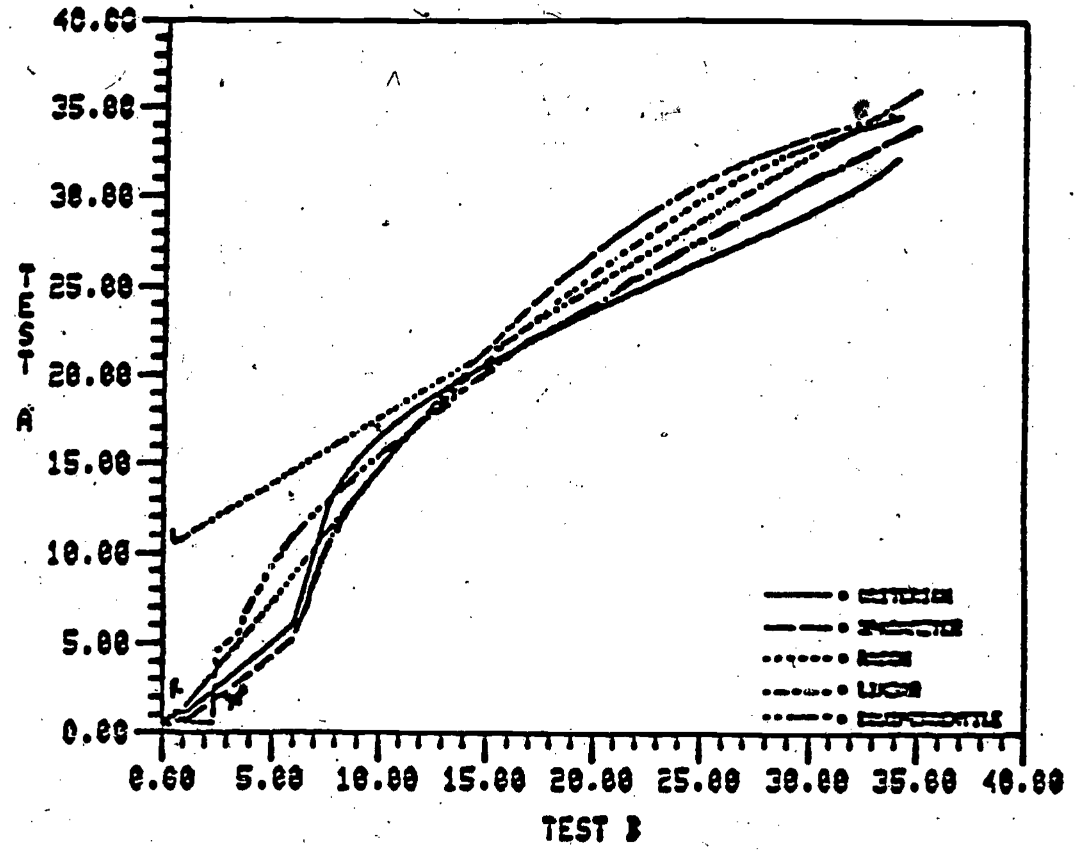
Test A: $\bar{b}$ = -.5; $\bar{a}$ = .8; c = .0; $\bar{\theta}$ = -.5
Test B: $\bar{b}$ = .5; $\bar{a}$ = .8; c = .2; $\bar{\theta}$ = .5



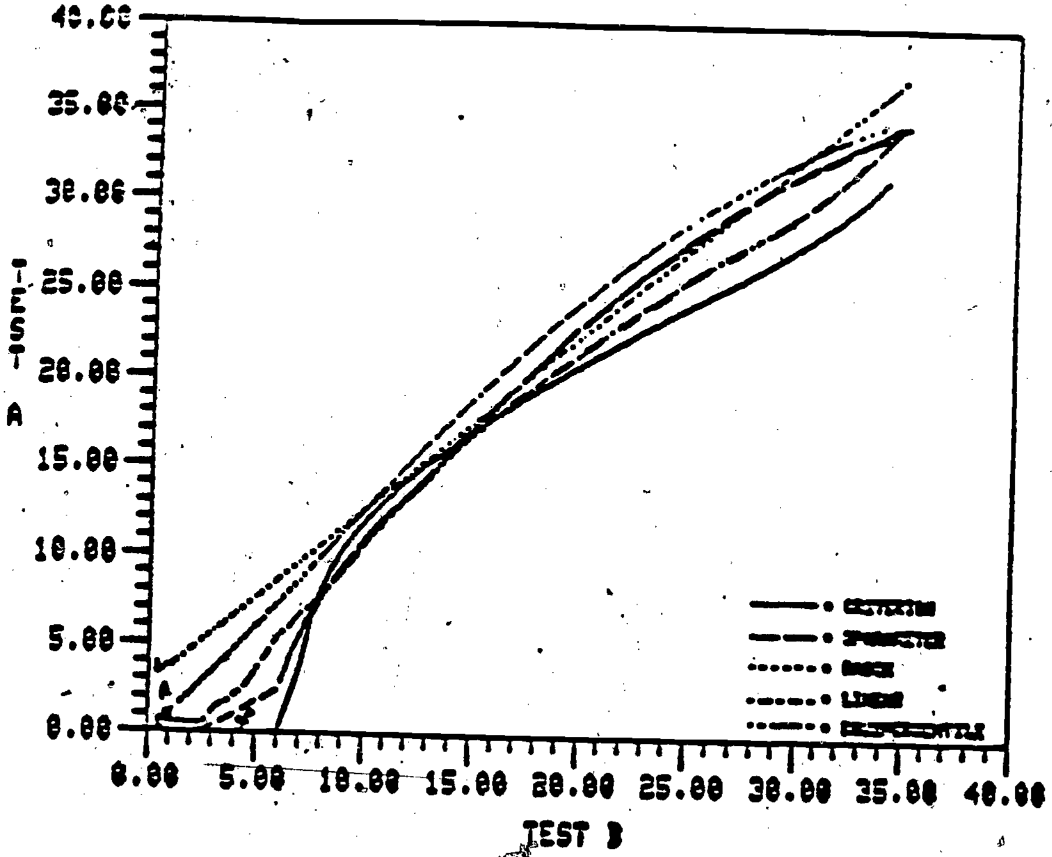
Test A: $\bar{b}$ = -.5; $\bar{a}$ = .8; c = .2; $\bar{\theta}$ = -.5
Test B: $\bar{b}$ = .5; $\bar{a}$ = .8; c = .0; $\bar{\theta}$ = .5



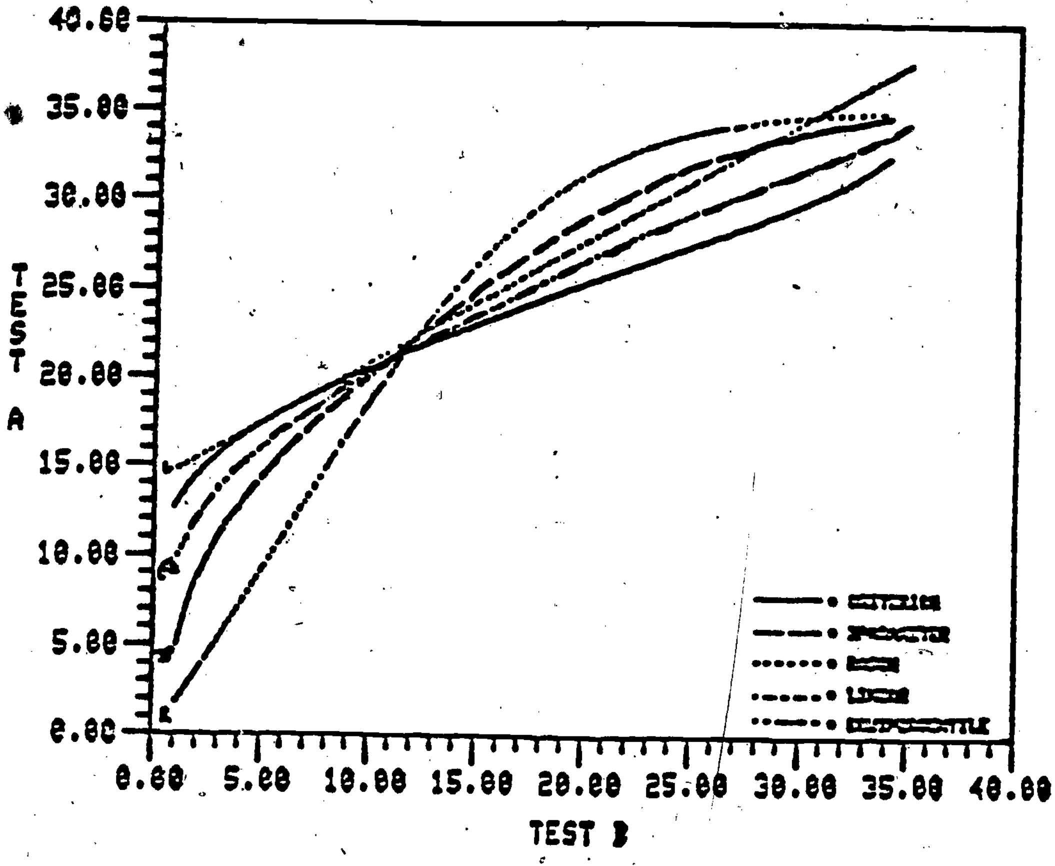
Test A: $\bar{b}$ = -.5; $\bar{a}$ = .5; c = .0; $\bar{\theta}$ = -.5
Test B: $\bar{b}$ = .5; a = 1.1; c = .0; $\bar{\theta}$ = .5



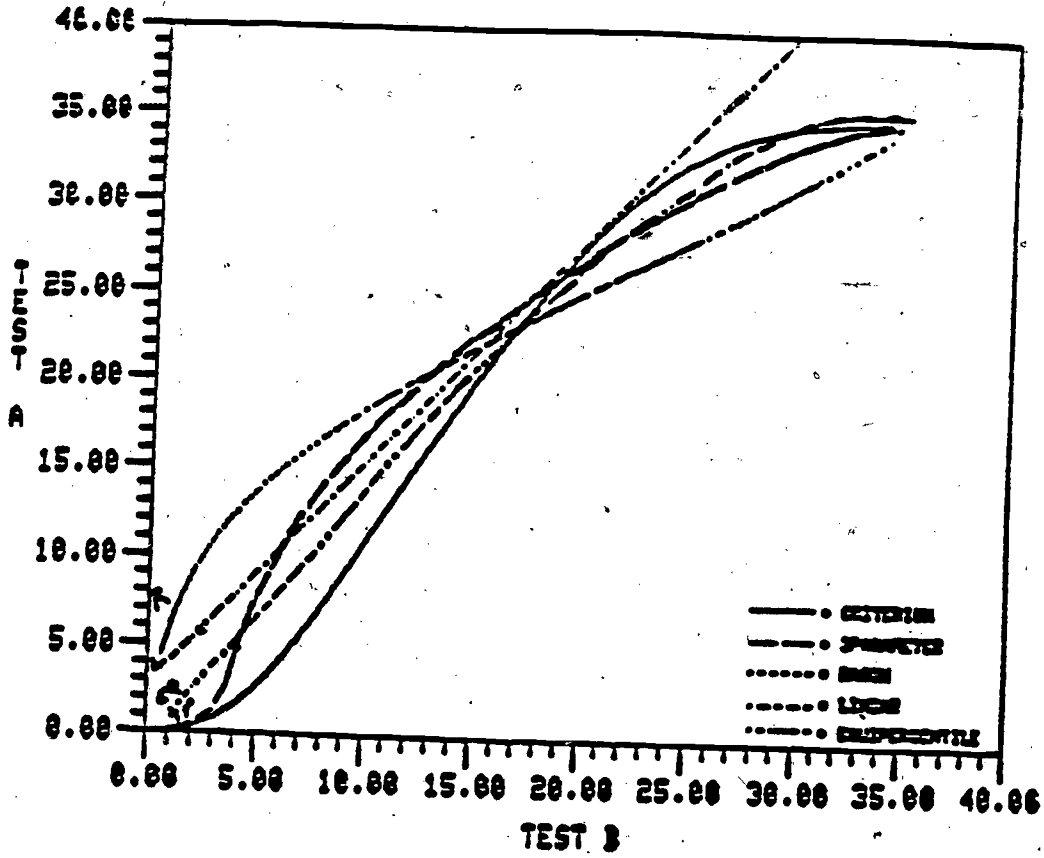
Test A: $\bar{b}$ = -.5; $\bar{a}$ = .5; c = .2; $\bar{\theta}$ = -.5
Test B: $\bar{b}$ = .5; $\bar{a}$ = 1.1; c = .2; $\bar{\theta}$ = .5



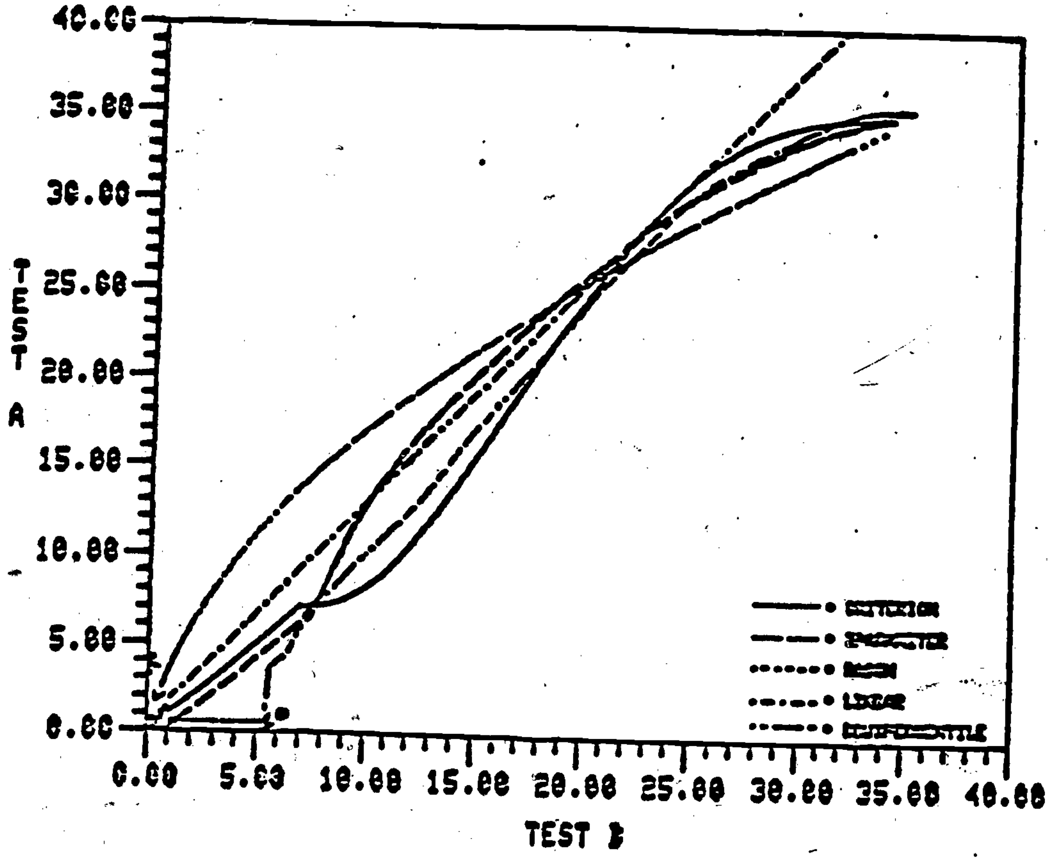
Test A: $\bar{b}$ = -.5; $\bar{a}$ = .5; c = .0; $\bar{\theta}$ = -.5
Test B: $\bar{b}$ = .5; $\bar{a}$ = 1.1; c = .2; θ = .5



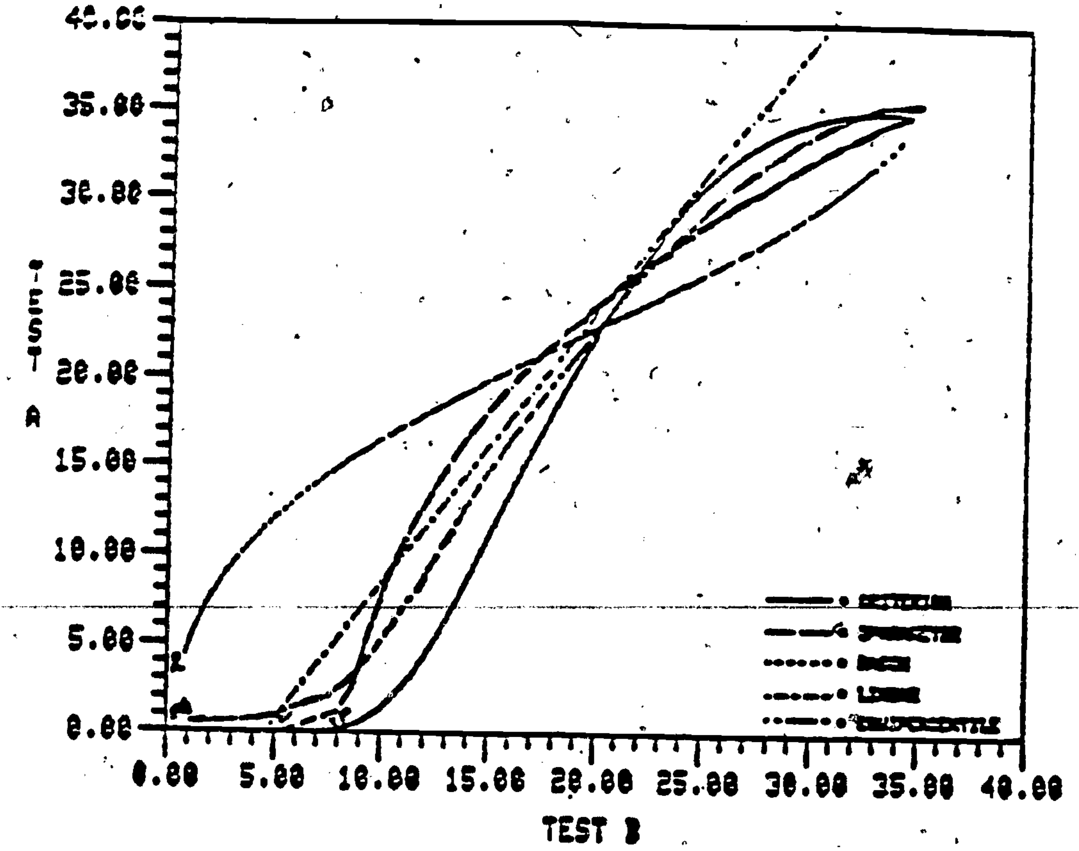
Test A: $\bar{b}$ = -.5; $\bar{a}$ = .5; c = .2; $\bar{\theta}$ = -.5
Test B: $\bar{b}$ = .5; a = 1.1; c = .0; 0 = .5



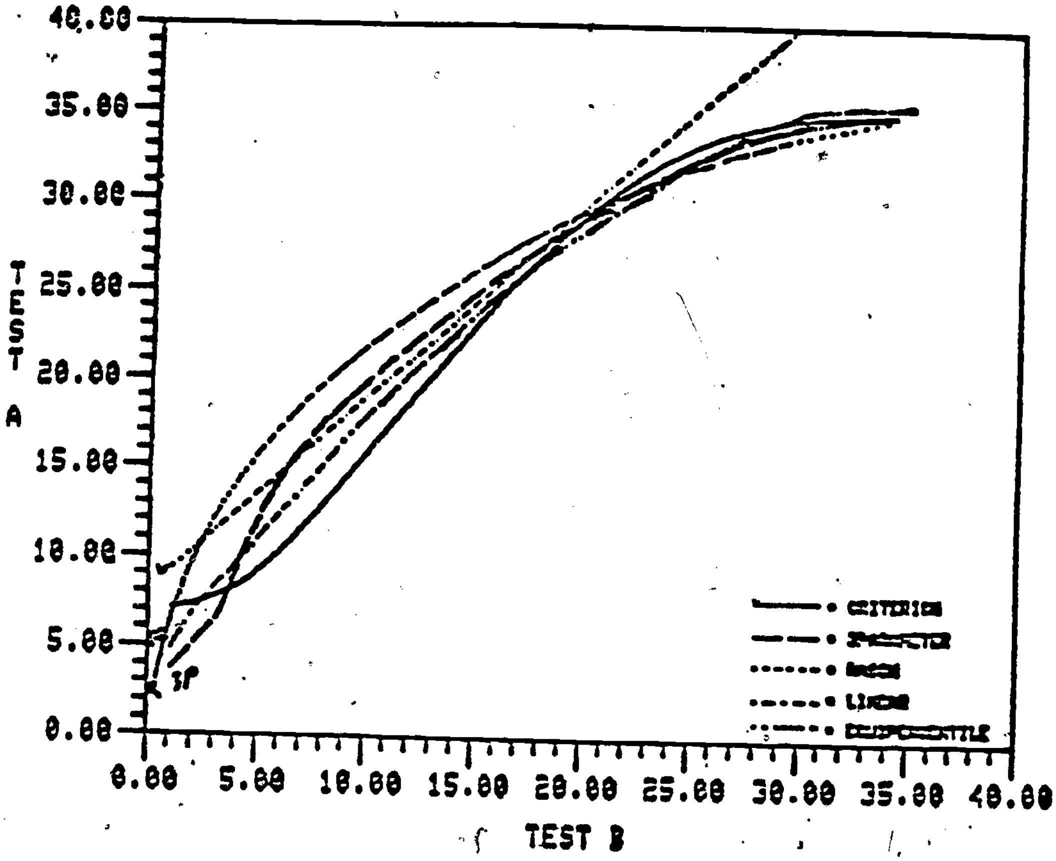
Test A: $\bar{b}$ = -.5; $\bar{a}$ = 1.1; c = .0; $\bar{\theta}$ = -.5
Test B: $\bar{b}$ = .5; a = .5; c = .0; $\bar{\theta}$ = .5



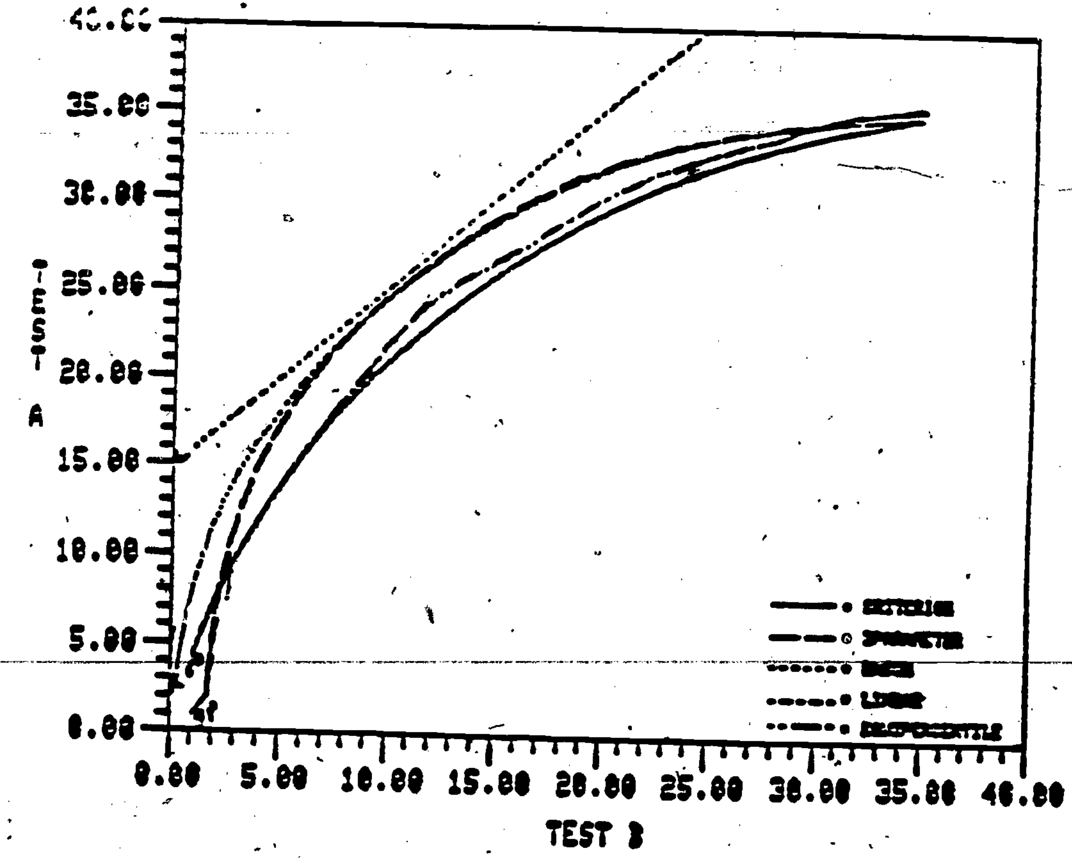
Test A: $\bar{b}$ = -.5; $\bar{a}$ = 1.1; c = .2; $\bar{\theta}$ = -.5
Test B: b = .5; a = .5; c = .2; $\bar{\theta}$ = .5



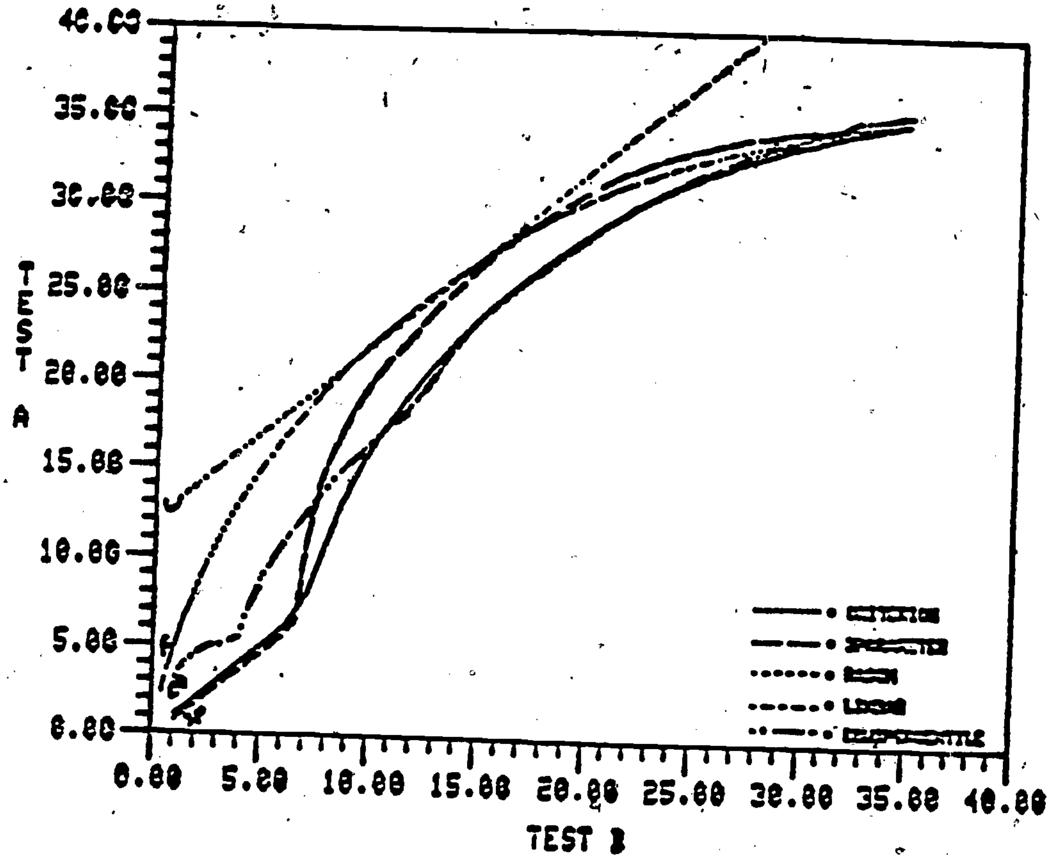
Test A: $\bar{b}$ = -.5; $\bar{a}$ = 1.1; c = .0; $\bar{\theta}$ = -.5
Test B: $\bar{b}$ = .5; a = .5; c = .2; 0 = .5



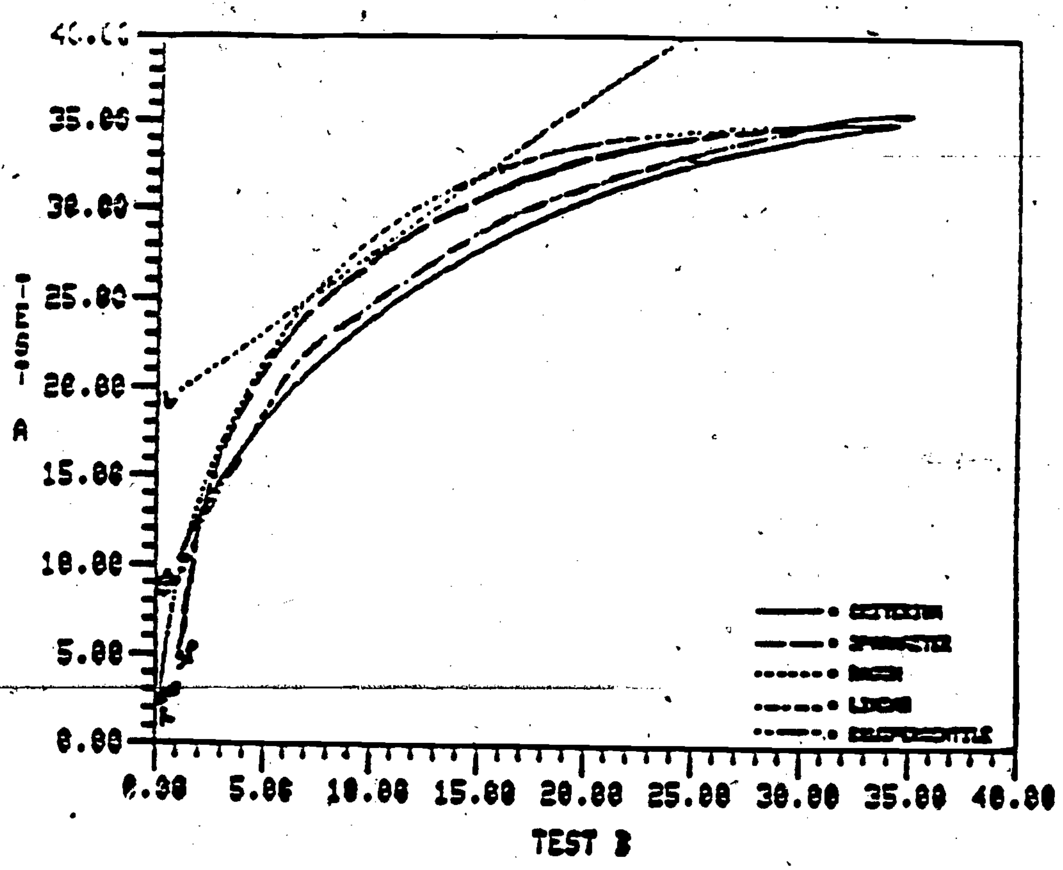
Test A: $\bar{b}$ = -.5; $\bar{a}$ = 1.1; c = .2; $\bar{\theta}$ = -.5
Test B: $\bar{b}$ = .5; $\bar{a}$ = .5; c = .0; $\bar{\theta}$ = .5



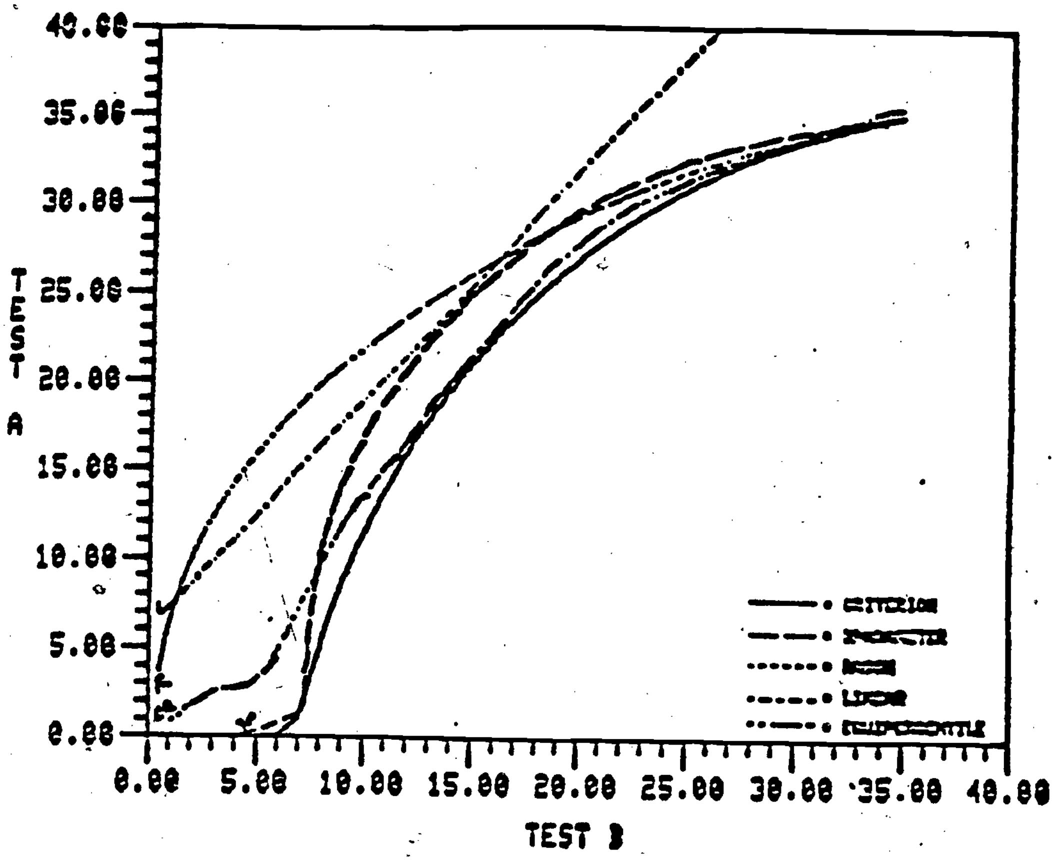
Test A: $\bar{b}$ = -1.0; $\bar{a}$ = .8; c = .0; $\bar{\theta}$ = -1.0
Test B: $\bar{b}$ = 1.0; $\bar{a}$ = .8; c = .0; $\bar{\theta}$ = 1.0



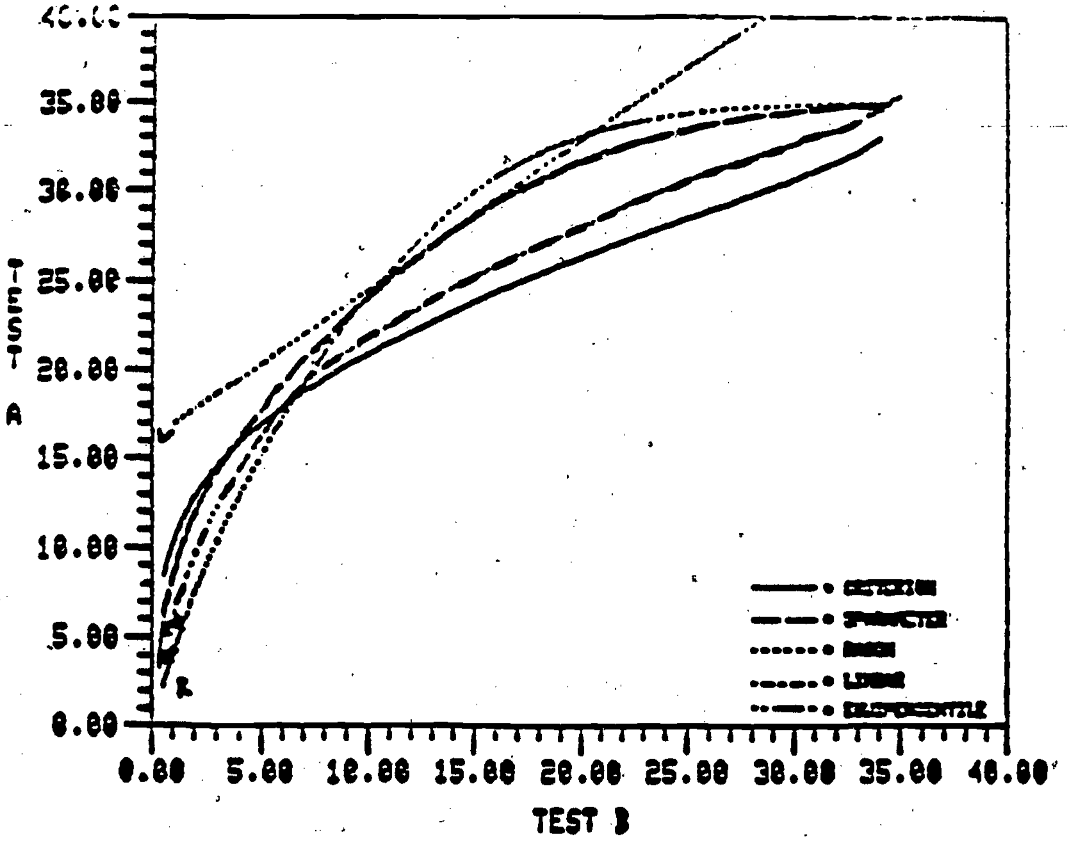
Test A: $\bar{b}$ = -1.0; $\bar{a}$ = .8; c = .2; $\bar{\theta}$ = -1.0
Test B: b = 1.0; a = .8; c = .2; 0 = 1.0



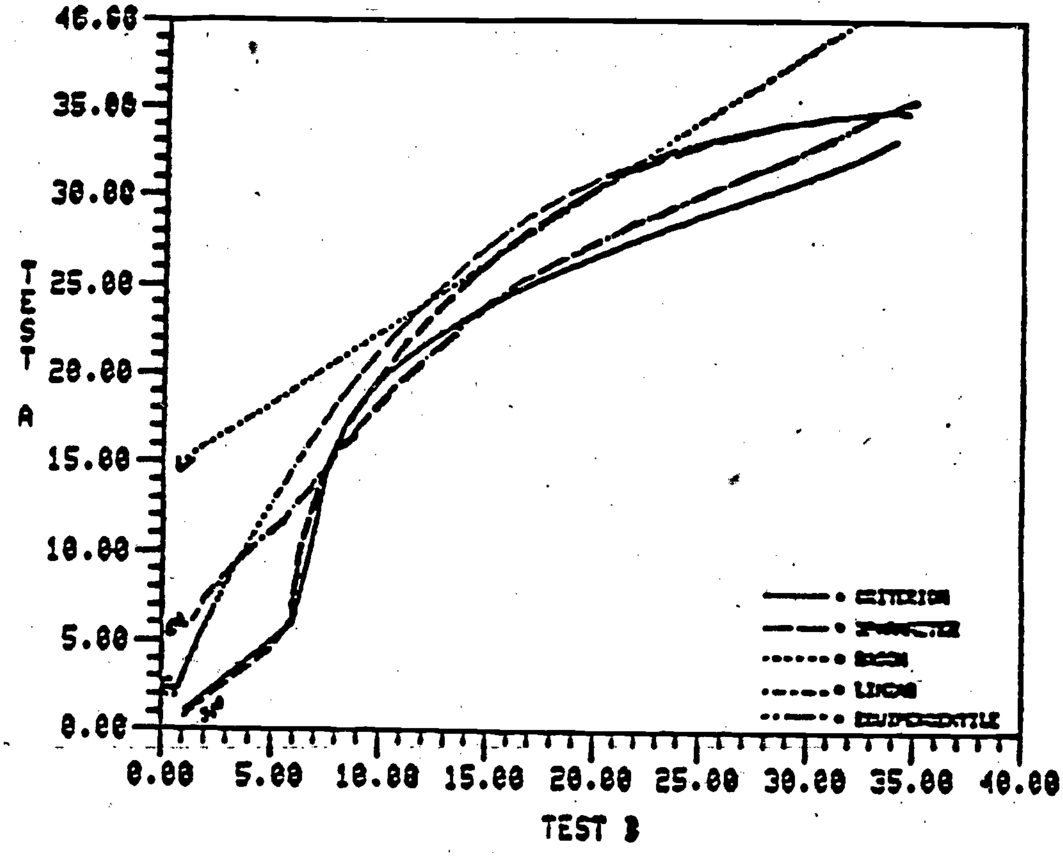
Test A: $\bar{b}$ = -1.0; $\bar{a}$ = .8; c = .2; $\bar{\theta}$ = -1.0
Test B: $\bar{b}$ = 1.0; $\bar{a}$ = .8; c = .0; $\bar{\theta}$ = 1.0



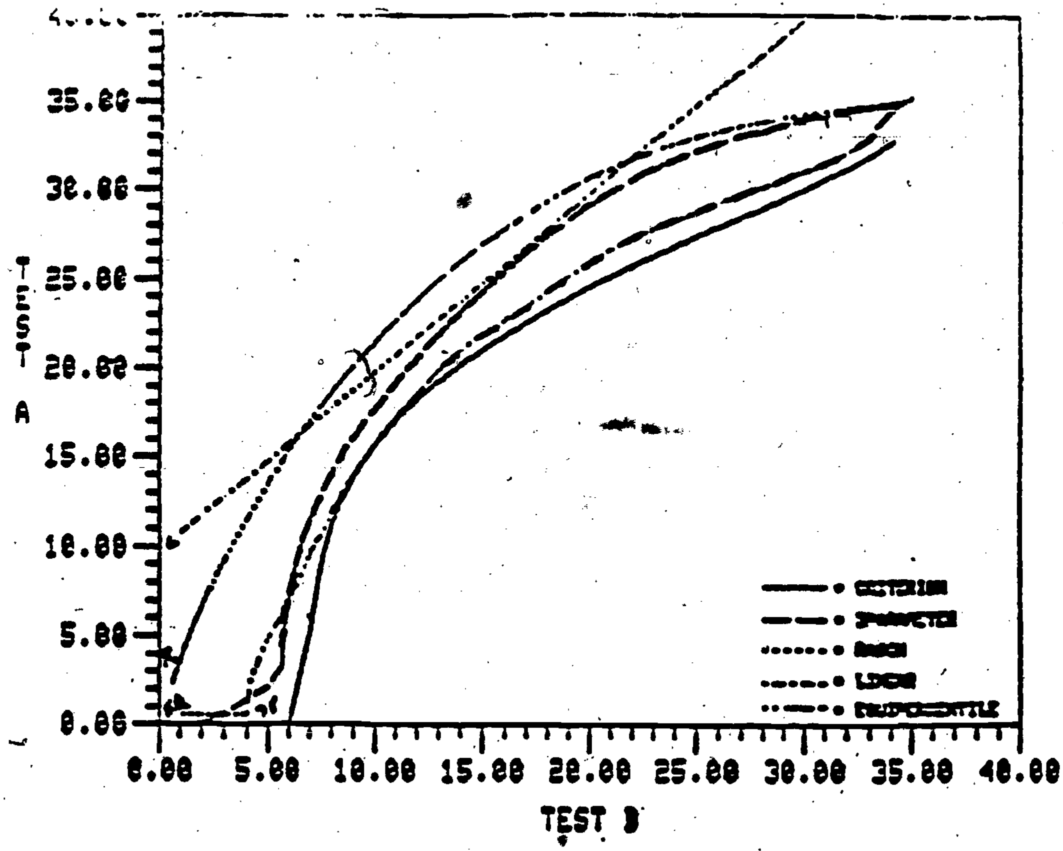
Test A: $\bar{b}$ = -1.0; $\bar{a}$ = .8; c = .0; $\bar{\theta}$ = -1.0
Test B: $\bar{b}$ = 1.0; a = .8; c = .2· 0 = 1.0



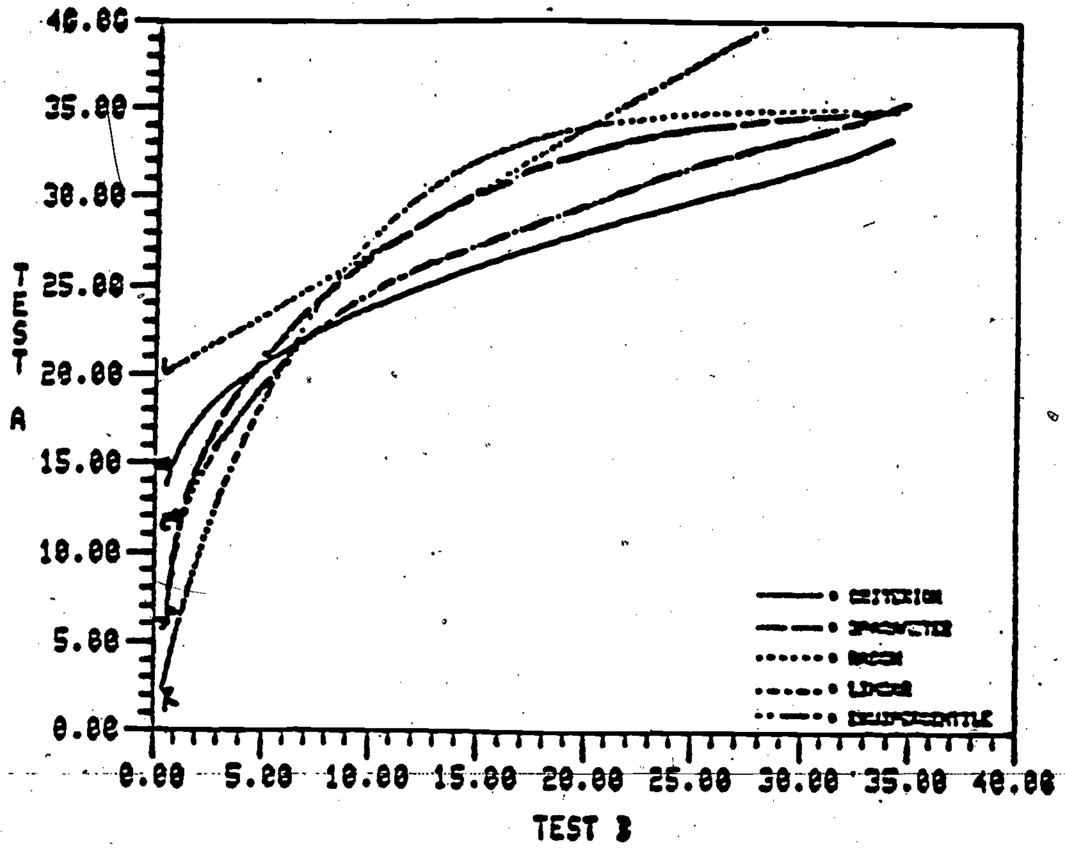
Test A: $\bar{b}$ = -1.0; $\bar{a}$ = .5; c = .0; $\bar{\theta}$ = -1.0
Test B: $\bar{b}$ = 1.0; $\bar{a}$ = 1.1; c = .0; $\bar{\theta}$ = 1.0



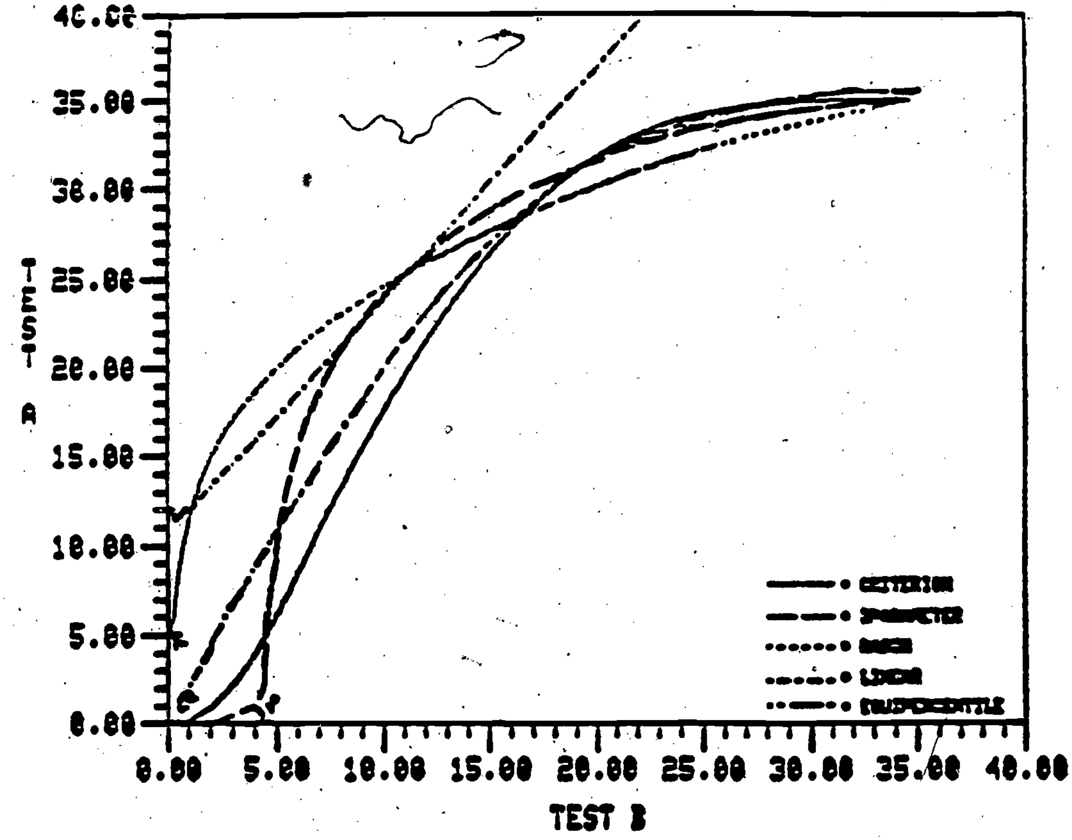
Test A: $\bar{b}$ = -1.0; $\bar{a}$ = .5; c = .2; $\bar{\theta}$ = -1.0
Test B: b = 1.0; a = 1.1; c = .2; θ = 1.0



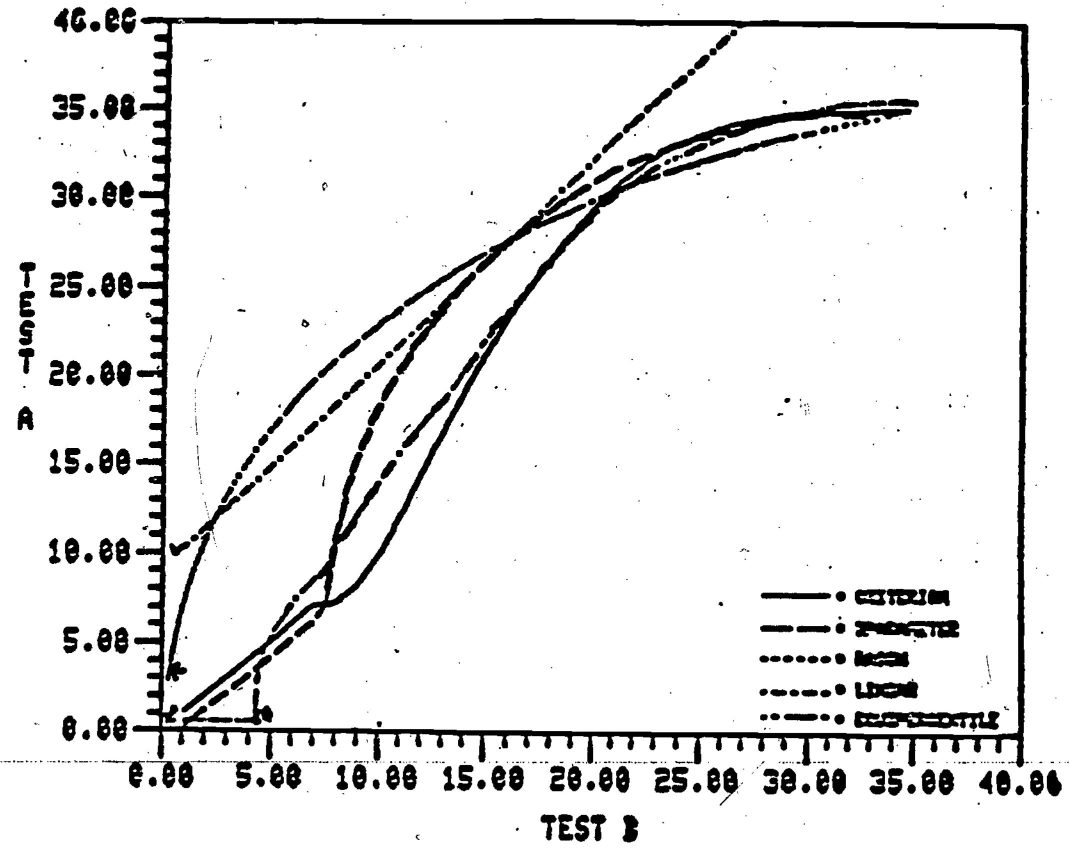
Test A: $\bar{b}$ = -1.0; $\bar{a}$ = .5; c = .0; $\bar{\theta}$ = -1.0
Test B: b = 1.0; a = 1.1; c = .2; $\bar{\theta}$ = 1.0



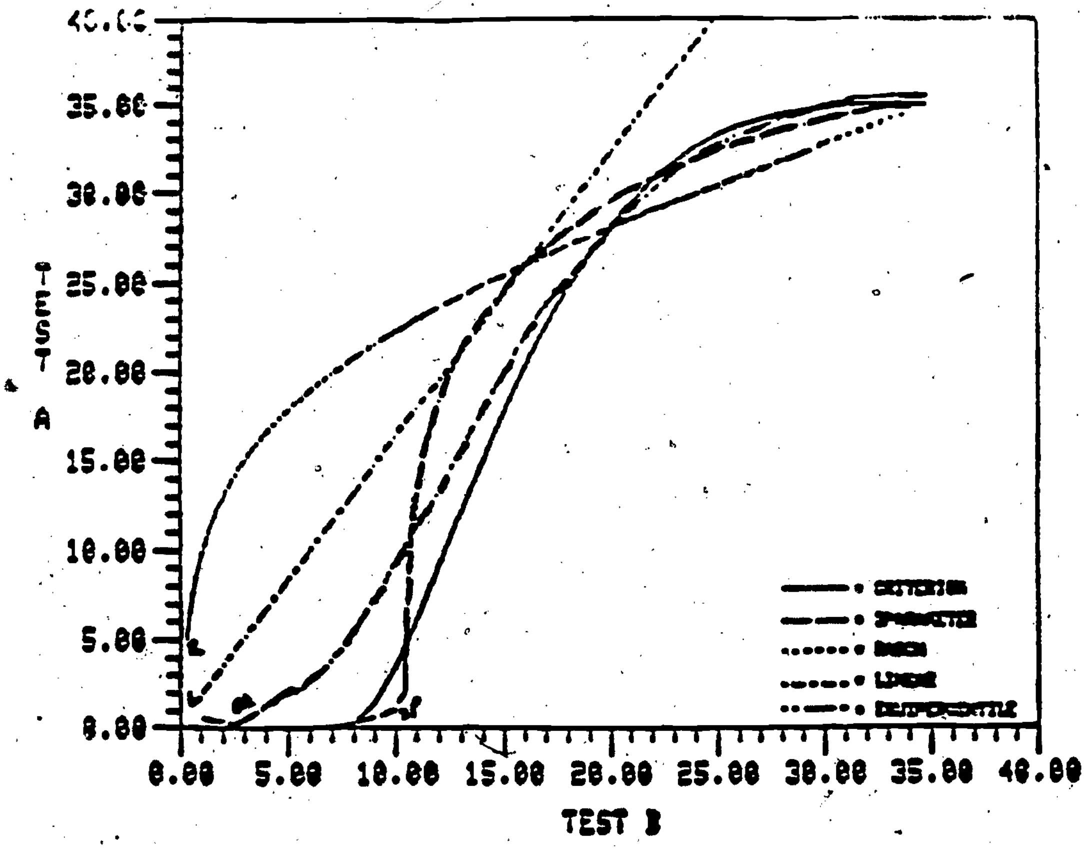
Test A: $\bar{b}$ = -1.0; $\bar{a}$ = .5; c = .2; $\bar{\theta}$ = -1.0
Test B: $\bar{b}$ = 1.0; $\bar{a}$ = 1.1; c = .0; $\bar{\theta}$ = 1.0



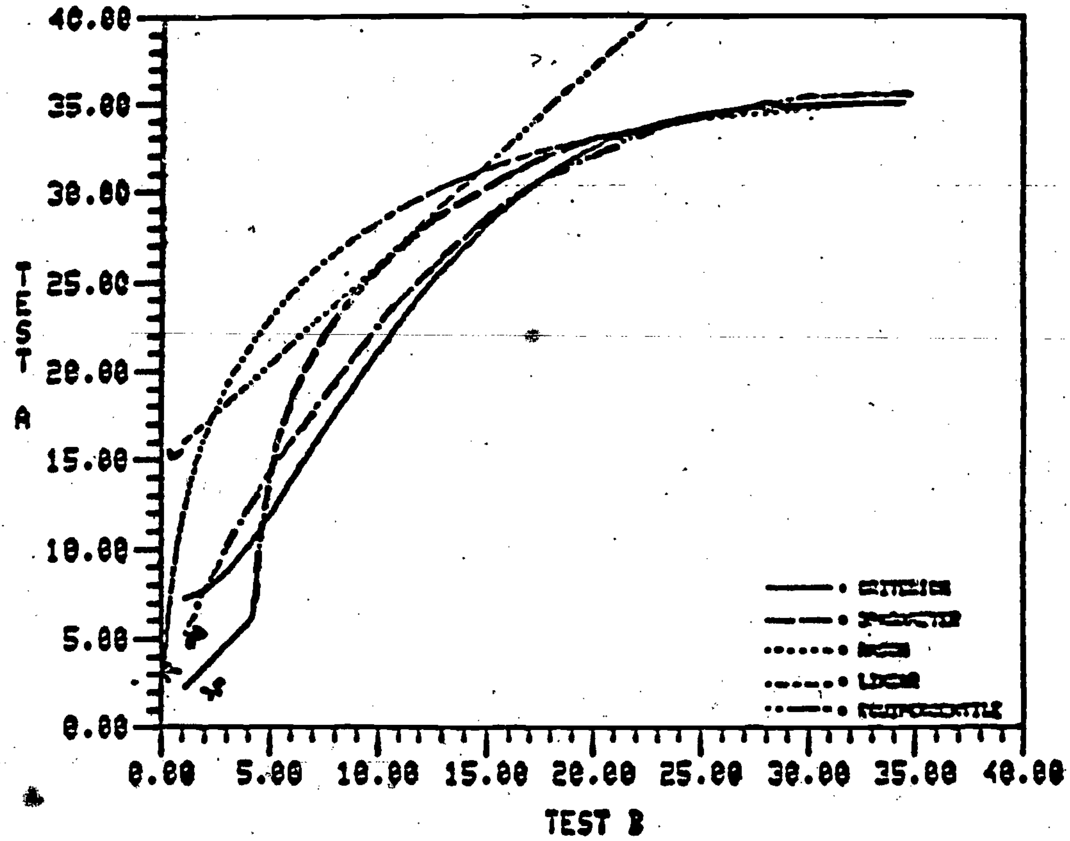
Test A: $\bar{b}$ = -1.0; $\bar{a}$ = 1.1; c = .0; $\bar{\theta}$ = -1.0
Test B: $\bar{b}$ = 1.0; $\bar{a}$ = .5; c = .0; $\bar{\theta}$ = 1.0



Test A: $\bar{b}$ = -1.0; $\bar{a}$ = 1.1; c = .2; $\bar{\theta}$ = -1.0
Test B: $\bar{b}$ = 1.0; a = .5; c = .2; $\bar{\theta}$ = 1.0



Test A: $\bar{b}$ = -1.0; $\bar{a}$ = 1.1; c = .0; $\bar{\theta}$ = -1.0
Test B: $\bar{b}$ = 1.0; a = .5; c = .2; θ = 1.0



Test A: $\bar{b}$ = -1.0; $\bar{a}$ = 1.1; c = .2; $\bar{\theta}$ = -1.0
Test B: $\bar{b}$ = 1.0; $\bar{a}$ = .5; c = .0; $\bar{\theta}$ = 1.0