A comparison of achievement level estimates from computerized adaptive testing and paper-and-pencil testing

G. Gage Kingsbury & Ronald L. Houser
Portland (OR) Public Schools

A paper presented to the Annual Meeting of the American Educational Research Association

New Orleans, LA
April 9, 1988

A comparison of achievement level estimates from computerized adaptive testing and paper-and-pencil testing

G. Gage Kingsbury & Ronald L. Houser
Portland (OR) Public Schools

Computerized adaptive testing (CAT: Weiss, 1978,1983; Weiss & Kingsbury, 1984) incorporating item response theory (IRT; Lord,1980; Lord & Novick, 1968) is being considered for use in many testing situations as an alternative or adjunct to paper-and-pencil testing. Test developers are considering CAT because of its measurement advantages over paper testing in the areas of test information, test validity, and test length (see, for example, Bejar, Weiss, & Gialluca, 1977 or English, Reckase, and Patience, 1977). In addition, testing professionals are considering the practical advantages that come from the use of CAT, including increased test security and availability, immediate scoring, and a decrease in logistical problems involved in moving test booklets and answer sheets (i. e., McBride & Moe, 1986).

With these advantages come a number of difficulties specific to computerized testing. These include computer down-time, disk crashes, proctor familiarity with computers, and availability of hardware and software. Reckase (1986) describes a CAT project in which the computer manufacturer went out of business, significantly reducing the amount of technical support available. Our own experience has indicated that major data losses are possible when test proctors fold, spindle, and mutilate floppy disks. In addition to computer problems, computerized testing requires a more general rethinking of testing procedures and score usage. To the extent that computers are in short supply, test administrators must shift from a batch processing mode (testing everyone in the class or work site at the same time), to an interactive processing mode (testing a few workers or students at a time). In the long run, these changes in behavior may prove to be the more difficult problem.

On balance, though, the measurement advantages of CAT have been demonstrated sufficiently in the past (i. e., Kingsbury & Weiss, 1981; Sympson, Weiss, & Ree, 1982; Urry, 1977) to cause researchers and practitioners to look closer into the use of adaptive testing for many testing purposes. Among the questions that must be answered when considering the use of CAT in an ongoing testing program is whether the trait level estimates obtained from the adaptive test are consistent with achievement level estimates obtained from paper-and-pencil tests (Olsen, Maynes, Slawson, & Ho, 1986; Sacher & Fletcher, 1978; Sympson, Weiss, & Ree, 1982).

The current study examines the comparability of achievement level estimates obtained from CAT and paper-and-pencil tests in the context of a grade school testing system and discusses some procedures that may be used to guide further research studies to be done with operational CAT systems. Previous

studies of this type have resulted in a single set of correlation coefficients, regression coefficients, and/or mean absolute differences comparing CAT and paper-and-pencil achievement level estimates. The current study adds a comparison of each current test with a previously administered paper-and-pencil test. This additional comparison will serve as a baseline measure to indicate the expected relationship between paper-and-pencil achievement level estimates.

Method

Testing Population. CAT tests were administered to 870 students enrolled in the third through eighth grades in an urban public school system during the 1986-1987 school year. These students took adaptive tests for a variety of reasons, including new student placement, counselor, teacher or parent request, and tracking of students in special programs. Student data was further refined by identifying and removing students with ambiguous ID numbers, inappropriate test protocols, or less than two paper-and-pencil testings within two school years. 747 students were identified for use in the final analysis sample. These students, some of whom took tests in more than one content area, took 174 language usage tests, 348 mathematics tests, and 443 reading tests.

Tests. The CAT and paper-and-pencil tests administered were from three basic skills item banks in Language Usage, Math, and Reading. These item banks consisted of 1600, 3200, and 2000 four- and five-alternative multiple choice items, respectively. The items in the three banks had previously been calibrated to three measurement scales using the one-parameter logistic IRT model (Rasch, 1960). Items were linked to the scales using the fixed parameter model described by Houser, Hathaway, and Ingebo (1983). These measurement scales were initially designed to have a mean of approximately 200 and a standard deviation of approximately 10.0. The units of these scales are called Rasch Units or RITS, for convenience.

Item subpools with 280 to 340 items were constructed for use with CAT. These subpools were designed to have the same proportion of items in each goal area as the paper-and-pencil tests described below. While the subpools were matched in this content balance to the paper-and-pencil tests, no attempt was made to match the content of individual adaptive tests. In addition, the CAT subpools were chosen to have a rectangular distribution of item difficulties. The CAT tests chose items to maximize the information at the momentary achievement level estimate. The test terminated when it reached its maximum test length (20 items), or when the standard error of measurement for the student's test score fell below five RIT points. In this situation, the average standard error of measurement for a student taking twenty items was between five and six RIT points. In this situation, then, the variable termination rule should have very little impact on the final test length. The adaptive tests were scored using a Bayesian scoring algorithm with a prior distribution with a mean of 200 and a standard deviation of 10. Since tests were administered on an individual basis, the date of the tests administrations was also obtained, to allow time of test to be used as a predictor variable in subsequent regression analyses.

Paper-and-pencil tests were drawn from the same item banks described above. In Mathematics and Language Usage, these tests were 60 items in length, while in Reading the tests were 48 items in

length. These were districtwide tests which are administered twice yearly to each student in third through eighth grade in the Portland (OR) public school system. These tests are level tests which differ in difficulty, and each of which is designed as a narrow range basic skills test. Students are assigned to a particular level based on previous test scores, locator tests, or teacher judgments. This testing system is quite similar to a two-stage adaptive test, except the routing test is replaced by an earlier set of earlier test scores. This similarity to an adaptive testing procedure might be expected to increase the similarity of adaptive and conventional test scores over the use of a single paper-and-pencil test. The paper-and-pencil tests were scored using a maximum-likelihood procedure.

Analyses. Two regression analyses were performed with the data sets described above. In each, the student's score from the conventional test administered closest in time to the adaptive test was used as the dependent variable or criterion test score. In the first analysis (CAT-P&PA), this criterion test score was predicted from the CAT score and the time lag between the two tests. In the second analysis (P&PB-P&PA), the criterion test score was predicted from the second conventional test score, and the time lag between the two tests. Assuming that each of the independent variables in each regression equation added significant variance to the prediction, the analysis would proceed as described below.

To the extent that the multiple-R in the CAT-P&PA analysis approached the multiple-R in the P&PB-P&PA analysis, it could be concluded that the adaptive test scores were as closely related to conventional test scores as the other conventional test scores were. This would be strong, but not conclusive evidence, that the CAT and conventional test scores were similar in their operational scoring characteristics.

A plot of residuals would be obtained from each analysis, to attempt to isolate patterned residuals occurring with either type of test. To the extent that patterned residuals occurred, it might indicate an incomplete or inappropriate regression model, while differences in observed patterns might indicate differences in the testing unaccounted for by the regression models.

Results

Descriptive statistics. Table 1 contains the means and standard deviations for achievement level estimates obtained from the adaptive test (CAT) and the two paper-and-pencil tests (P&PA and P&PB) in each of the three content areas. In each content area, the same general trends were seen. The mean achievement level was lowest for P&PB, next lowest for P&PA, and highest for CAT. The standard deviation of the achievement level estimates around the mean was lowest for P&PB, next lowest for P&PA, and highest for CAT. The only counter example to these trends was that the standard deviation of the Mathematics CAT achievement level estimates was slightly smaller than that for the P&PA achievement level estimates. Some part of the differences in mean achievement level may have been due to differences in the student/test interaction, while other parts of the differences may be attributable to differences in the times at which the students were tested. This will be discussed in more detail below.

In addition, Table 1 shows the mean number of items administered in each type of test, and its standard deviation. As expected, the mean test lengths for the adaptive tests were quite close to the maximum test length, 20 items. The shortest mean test length observed was 19.51 items, for the Language Usage test. In general, then, the average mean test length for the CAT tests in Reading was 42% of the P&P tests in Reading, while in the other two content areas, the mean CAT test length was 33% of the length of the P&P tests.

Table 2 shows the mean number of months of instruction that passed between the tests that the students took. To give meaning to these numbers, it is helpful to remember that there are approximately nine instructional months in the public school year. The shortest mean instructional lag time noted for any test was 2.34 months from P&PA to CAT in Reading. The longest mean lag time noted was 4.60 months from P&PA to CAT in Language Usage. In each content area, test P&PB tended to be the first test taken, P&PA the second test taken, and the CAT test the last. While this trend is consistent, it should be noted from the percentages of students taking CAT an P&PB before and after P&PA that many students took the tests in different orders.

The mean lag times shown in Table 2 do not completely describe the lags between the tests. Paper-and-pencil tests were only offered twice during the school year, and so the distribution of lag times from P&PA to P&PB tended to be discontinuous. At the same time, CAT tests were administered throughout the year, and so the distribution of lag time between P&PA and CAT tended to be more continuous.

Regression Analyses. Table 3 shows the Beta weights for the test score and instructional lag time that were included in each regression equation, as well as the probability level from the associated F-test. In addition, the multiple-R for each equation is shown. Three major trends can be seen from this table.

First, the predictor test scores were given universally high, positive beta weights in the prediction of the criterion test score. Second, the beta weights of the instructional lag between tests were smaller in magnitude than those for the predictor test score. Third, and most important, the multiple regression correlation coefficients (R) for the prediction of P&PA scores from CAT scores and the associated time lag were consistently higher than those obtained from the corresponding prediction of P&PA scores from P&PB scores.

The beta weights associated with instructional lag time were small, and tended to be even smaller (not reaching a p-level less than .05) when the predictor score came from CAT. Since these weights were significant for some of the regression equations, the full regression model was used for all further analyses.

Investigating the differences in multiple-R values was accomplished using Fisher's r-to-z transformation and a one-tailed test of significance using the transformed multiple-R values from the equation predicting P&PA from P&PB and the associated lag. The analyses indicated that the multiple-Rs for the prediction of P&PA form CAT and the associated lag were higher than those for the

prediction of P&PA from P&PB and its associated lag ($p < .01$) for Mathematics and Reading, but not for Language Usage.

Figures 1, 2 and 3 show the differences between the observed scores on P&PA and the predicted scores from the two regression equations, as a function of the observed score on P&PA, for each of the three content areas. It can be seen from these residual plots that the two regression equations from each content area performed in a remarkably similar manner. Lower scorers on the criterion test tended to be overpredicted, while higher scorers tended to be underpredicted, as in most regression situations.

Figures 4, 5, and 6 show the predicted P&PA score obtained from the CAT/lag regression equation minus the predicted P&PA score obtained from the P&PB/lag regression equation. The lack of any systematic pattern in these graphs indicates that the two regression equations were not making systematically different predictions of students' performance on the criterion test. No biasing of test scores is apparent from these figures. Correlations between the differences in prediction and the criterion scores were .038 for Language Usage ($p > .05$), .196 for Mathematics ($p < .01$), and .183 for Reading ($p < .01$). This residual correlation may be due to factors not considered in the regression models.

Discussion and Conclusions

Results of this study indicate that the CAT tests being used seem to have the same score characteristics as the paper-and-pencil tests being used. The relationship between CAT and paper-and-pencil tests was of the same magnitude (if not somewhat stronger) and form as the relationship between two paper-and-pencil tests. This study serves as an additional validation of the use of adaptive testing in this achievement context. While substantial work has already been done that shows that CAT in an achievement context is feasible (i. e., Bejar & Weiss, 1977; Reckase, 1986; Vale, 1985), some research of this type needs to be completed in each new situation to enable the findings to be generalized.

Most of the situations in which implementations of CAT are being considered will require the CAT system to work alongside a continuing paper-and-pencil testing system. In these situations, the correspondence of CAT and paper-and-pencil scores is extremely important to enable the educator to make consistent, informed decisions. In addition, legal questions are sure to arise if the two systems result in different scores for students of the same ability. If the same student can expect to receive about the same test score in the two modes of testing, the two systems should be able to function in harmony.

Results of this study indicate that scores from CAT and paper-and-pencil tests can be considered to be as interchangeable as scores from two paper-and-pencil tests would be. If additional information concerning the equivalence of operational usage of specific items in CAT and paper-and-pencil tests can be obtained, this will serve as strong evidence of the equivalence of the two types of tests for use in the school district.

A failing in studies of this type (including this one) is the universal use of the multiple regression model to compare performance in different testing modalities. For studies in which instructional time elapses

between tests, a more appropriate model would be a psychological growth model designed to match the specific testing situation. Numerous types of growth models have already been proposed in the literature (i. e., Rogosa, Brandt, & Zimowski, 1982 or Sagiv, 1979), and the use of an appropriate model should increase the validity of modality comparisons done with longitudinal testing samples.

The study described herein may serve as a minor model for doing longitudinal research with ongoing CAT systems, if only in that it may point out some unexpected complications in performing this type of study. For instance, no ordering of the tests was done intentionally in this study, however a systematic ordering was observed. Two factors may have contributed to this ordering. First, the students taking the achievement levels tests (ALT: the paper-and-pencil tests used in this study) were in grades three through eight, while the students taking the CAT tests were in grades two through twelve. Since few students were tested in grade two, the majority of students taking the CAT tests in grades not taking the ALTs tended to be high school students who had taken their last ALT sometime within the past two years. There were probably enough high school students in the sample to cause the ordering observed in the study. A second factor would also elicit the same type of ordering. This factor was that the CAT tests were just being introduced into the school system. To the extent that a student had two ALT test scores and a CAT test score, it would be quite likely that the CAT test would be the last, since it was the type of test that was made available most recently.

Research of this type needs to be closely tied to research examining the relative fit of items to the IRT models in CAT and paper-and-pencil test administration. If differences in achievement level estimates exist, they may be due to novelty effects or motivational effects that might be controllable. On the other hand, if item parameters differ with a change in test administration mode, this could differentially impact students' scores and their interpretation. In this situation, the problems inherent in using CAT and paper-and-pencil testing for the same purpose may be increased and seperate calibration procedures for items used in paper-and-pencil tests and CAT tests might be needed. An effort to establish guidelines for the validation of computerized testing is already underway (Green, Bock, Humphreys, Linn, & Reckase, 1984), and this effort needs to be supported by school districts, universities, and test publishers who are considering CAT as an adjunct to paper-and-pencil testing.

References

Bejar, I. I., Weiss, D. J., & Gialluca, K. A. (1977). An information comparison of conventional and adaptive tests in the measurement of classroom achievement (RR 77-7). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

English, R. A., Reckase, M. D., & Patience, W. M. (1977). Application of tailored testing to achievement measurement, Behavioral Research Methods & Instrumentation, 9, 158-161.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests, Journal of Educational Measurement, 21, 347-360.

Houser, R. L., Hathaway, W. E., & Ingebo, G. S. (1983). An alternate procedure to obtain ability estimates in latent trait models. Paper presented to the annual meeting of the American Educational Research Association, Montreal, Canada.

Kingsbury G. G. & Weiss, D. J. (1981). A validity comparison of adaptive and conventional strategies for mastery testing (Research Report 81-3). Minneapolis, MN: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Lord, F. M. & Novick M. R. (1968). Statistical theory of mental test scores. Reading, MA: Addison-Wesley.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.

McBride, J. R. & Moe K. C. (April, 1986). Computerized adaptive achievement testing: A prototype. Paper presented to the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Olsen, J. B., Maynes, D. D., Slawson, D., & Ho, K. (April, 1986). Comparison and equating of paper-administered, computer-administered, and computerized adaptive tests of achievement. Paper presented to the annual meeting of the American Educational Research Association, San Francisco.

Rasch, G. O. (1960). Probabilistic models for some intelligence and attainment tests. Copenhagen: Danish Institute for Educational Research.

Reckase, M. D. (March, 1986). Final report: The use of tailored testing with instructional programs (Research Report ONR 86-1). Iowa City, Iowa: The American College Testing Program, Assessment Programs Area, Test Development Division.

Rogosa, D. R., Brandt, D., & Zimowski, M. (1982). A growth curve approach to the measurement of change. Psychological Bulletin, 92, 726-748.

Sacher J. D. & Fletcher J. D. (1978). Administering paper-and pencil tests by computer, or the medium is not always the message. In D. J. Weiss (Ed.), Proceedings of the 1977 computerized adaptive testing conference. Minneapolis, MN: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Sagiv, A. (1979). General growth model for evaluation of an individual's progress in learning. Journal of Educational Psychology, 71, 866-881.

Sympson, J. B., Weiss, D. J., & Ree, M. J. (1982). Predictive validity of conventional and adaptive tests in an Air Force training environment (AFHRL-TR-81-40). BAFB, TX: Air Force Human Resources Laboratory.

Urry, V. (1977). Tailored testing: A successful application of latent trait theory. Journal of Educational Measurement, 14, 181-196.

Vale, C. D. (December, 1985). Development of a microcomputer-based adaptive testing system: Phase II -- Implementation (Research Report ONR 85-5). St. Paul, MN: Assessment Systems Corporation.

Weiss, D. J. (Ed.) (1978). Proceedings of the 1977 computerized adaptive testing conference. Minneapolis, MN: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Weiss, D. J. (Ed.) (1983).New horizons in testing: Latent trait test theory and computerized adaptive testing. New York: Academic Press.

Weiss, D. J. & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. Journal of Educational Measurement, 21, 361-375.

Table 1

Mean and Standard Deviations for Achievement Level Estimates and Test Lengths from each Test in each Content Area

	Ach. Level Estimate (RITs)			Test Length (Items)	
	Mean	SD	N	Mean	SD
Language					
CAT	210.60	16.03	174	19.51	1.72
P&PA	207.02	15.22	174	60	na
P&PB	205.49	13.02	174	60	na
Mathematics					
CAT	216.41	14.32	348	19.90	.29
P&PA	215.80	14.97	348	60	na
P&PB	212.12	12.52	348	60	na
Reading					
CAT	215.29	16.06	443	19.95	.23
P&PA	212.54	15.48	443	48	na
P&PB	208.02	13.36	443	48	na

Table 2

Means and Standard Deviations of Months of Instructional Lag Time between Tests in each Content Area, and the Percentage of Tests Preceding (% PRE) and Following (% POST) the Criterion Test (P&PA)

	Mean	SD	% PRE	% POST
Language				
P&PA to CAT	4.60	7.18	19.5	63.8
P&PA to P&PB	-3.73	6.24	69.0	21.3
Mathematics				
P&PA to CAT	2.52	6.16	39.9	51.7
P&PA to P&PB	-4.07	5.97	75.9	15.2
Reading				
P&PA to CAT	2.34	6.03	45.1	44.5
P&PA to P&PB	-3.64	5.98	67.7	15.8

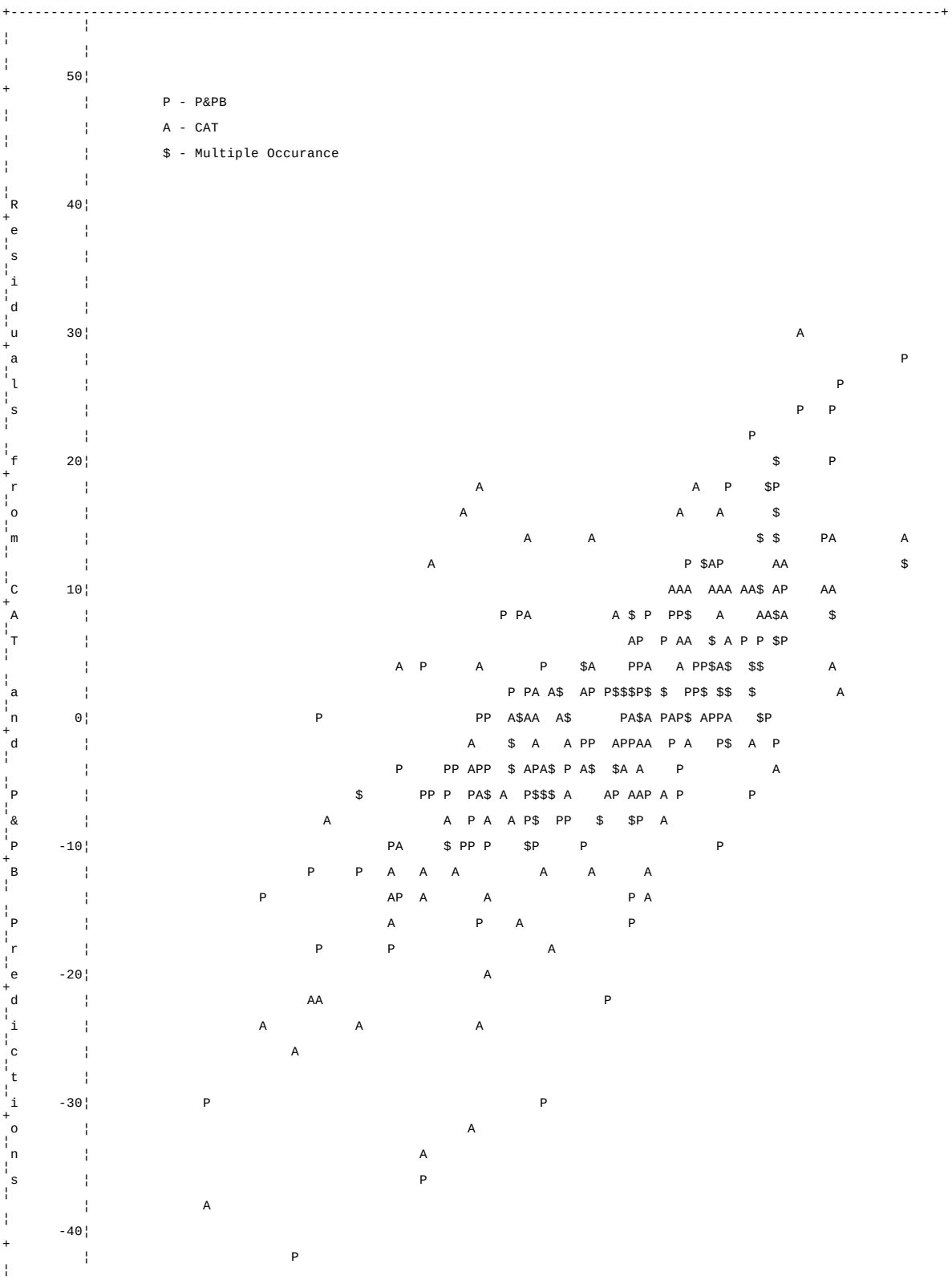
Table 3

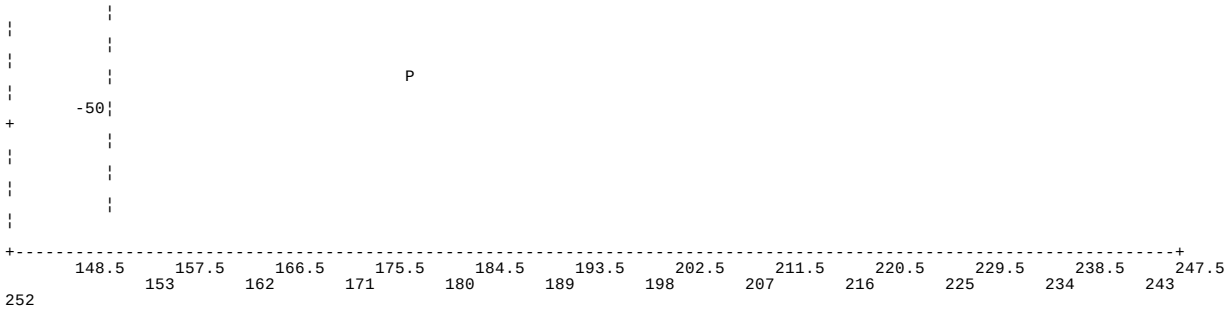
Regression Coefficients and Associated Probability Levels for each Variable Used to Predict P&PA Performance, and the Multiple-R Associated with each Prediction Equation

	Beta(Score)	p	Beta(lag)	p	R
Language					
CAT	.700	<.001	-.005	.928	.700
P&PB	.655	<.001	.193	.001	.682
Mathematics					
CAT	.643	<.001	.080	.059	.634
P&PB	.484	<.001	.181	<.001	.521
Reading					
CAT	.690	<.001	.029	.409	.690
P&PB	.595	<.001	.111	.004	.595

Figure 1

Residuals -- Language Achievement Level Estimates

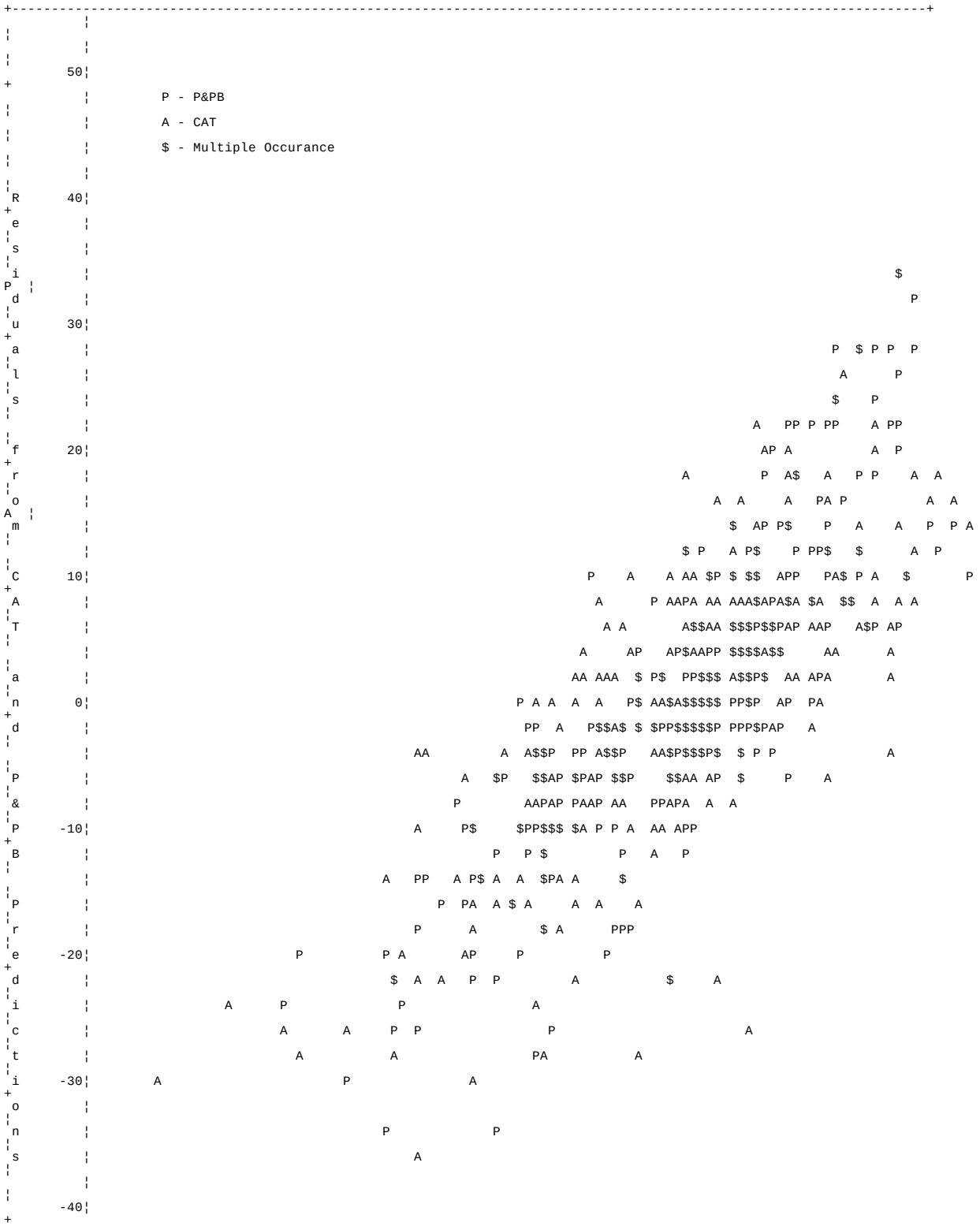


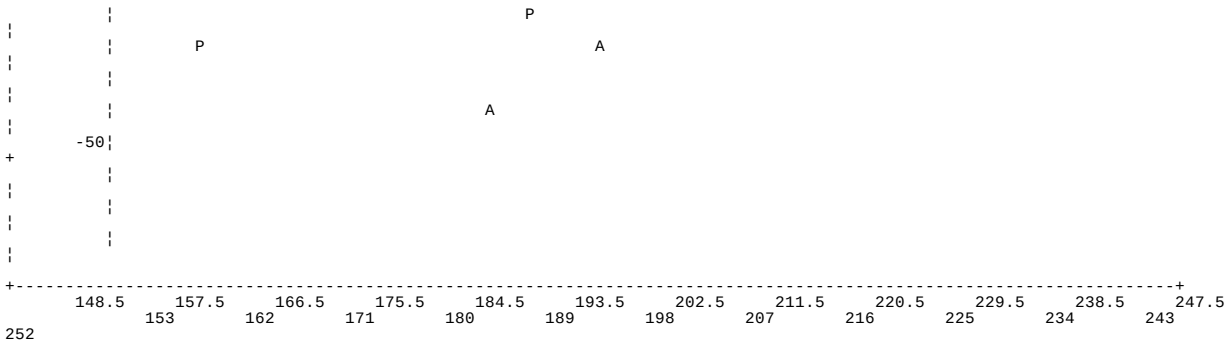


Theta Estimates: First P+P Test

Figure 2

Residuals -- Mathematics Achievement Level Estimates

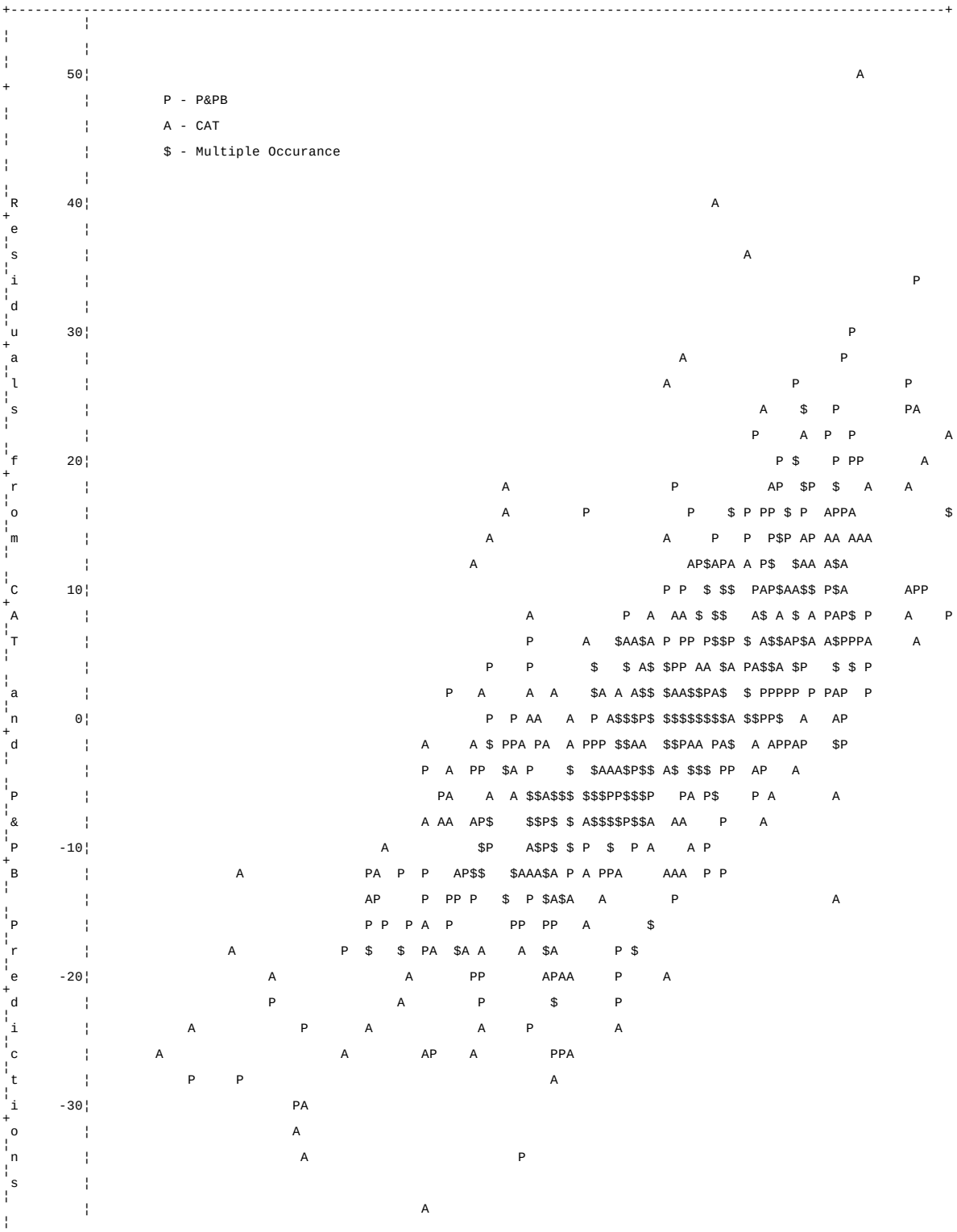


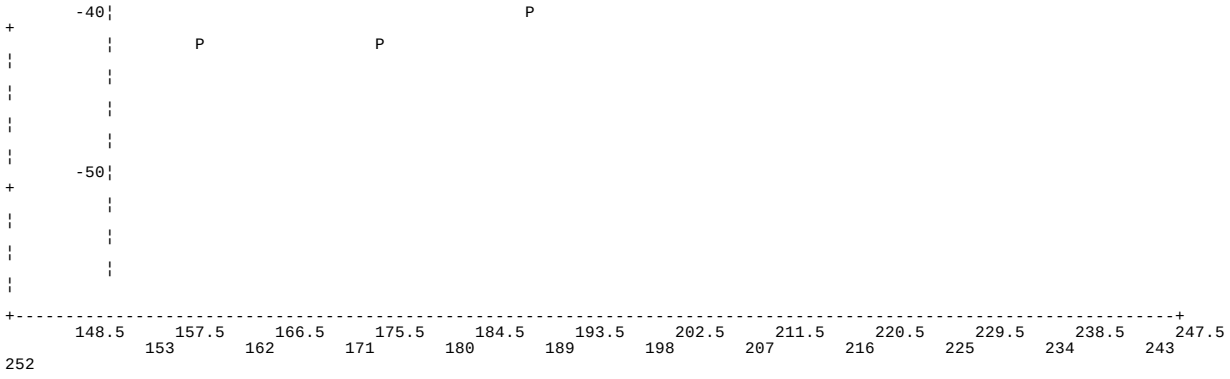


Theta Estimates: First P+P Test

Figure 3

Residuals -- Reading Achievement Level Estimates

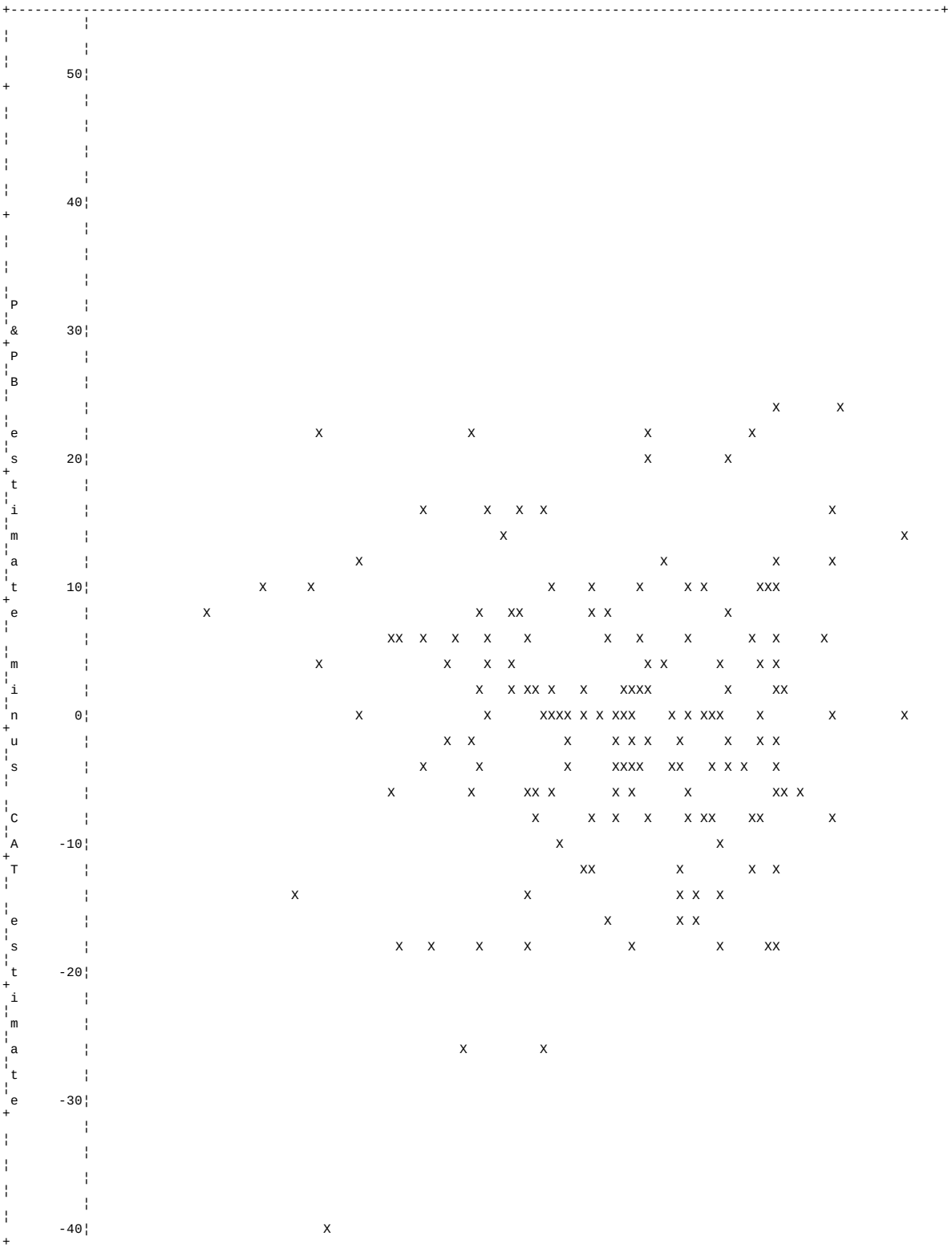


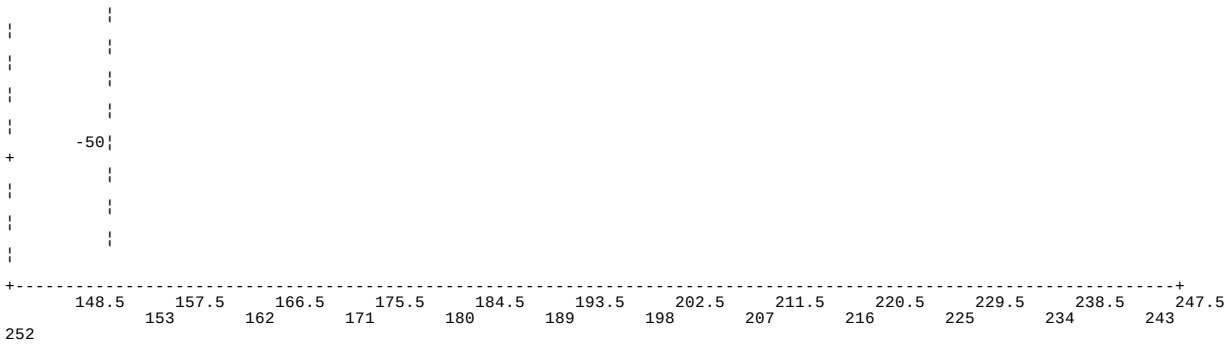


Theta Estimates: First P+P Test

Figure 4

Difference in Language Achievement Level Estimates

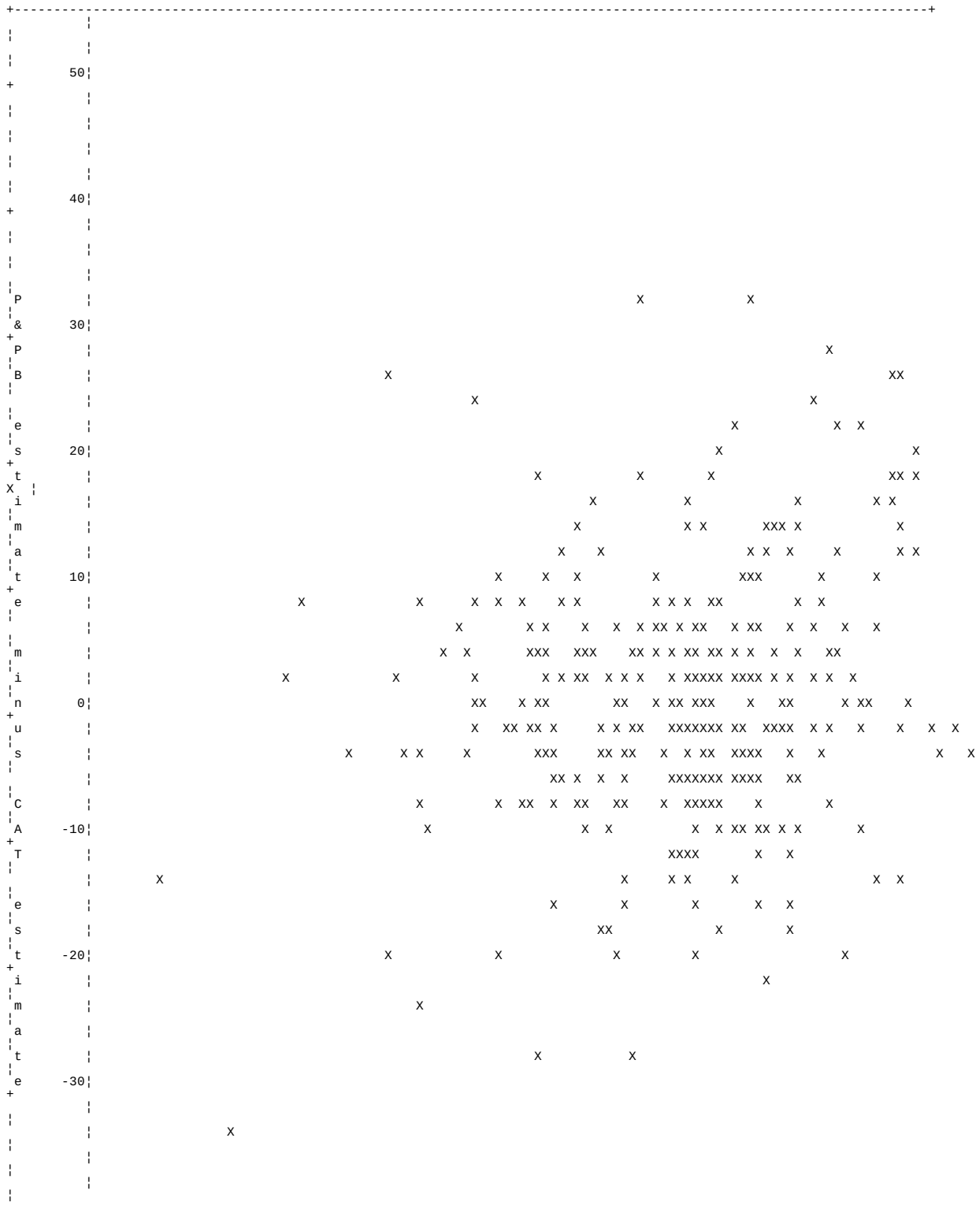


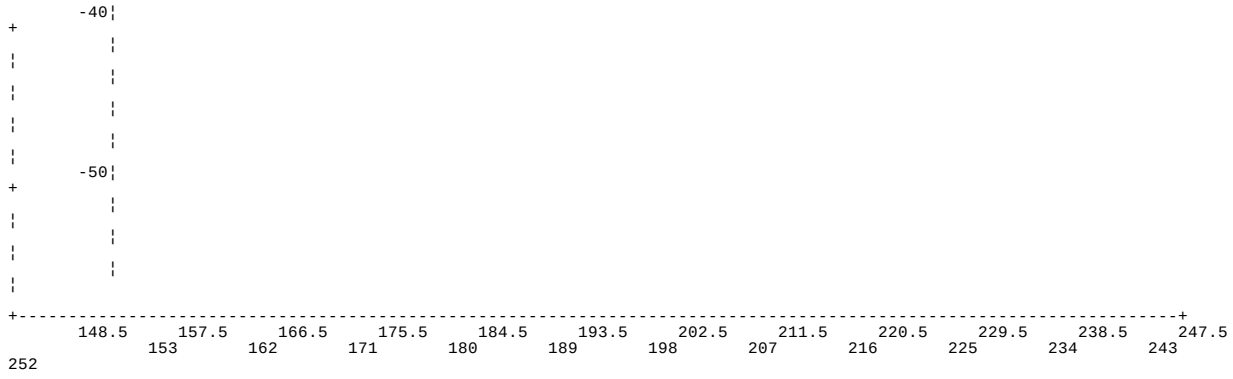


Achievement Level Estimates: P&PA

Figure 5

Difference in Mathematics Achievement Level Estimates

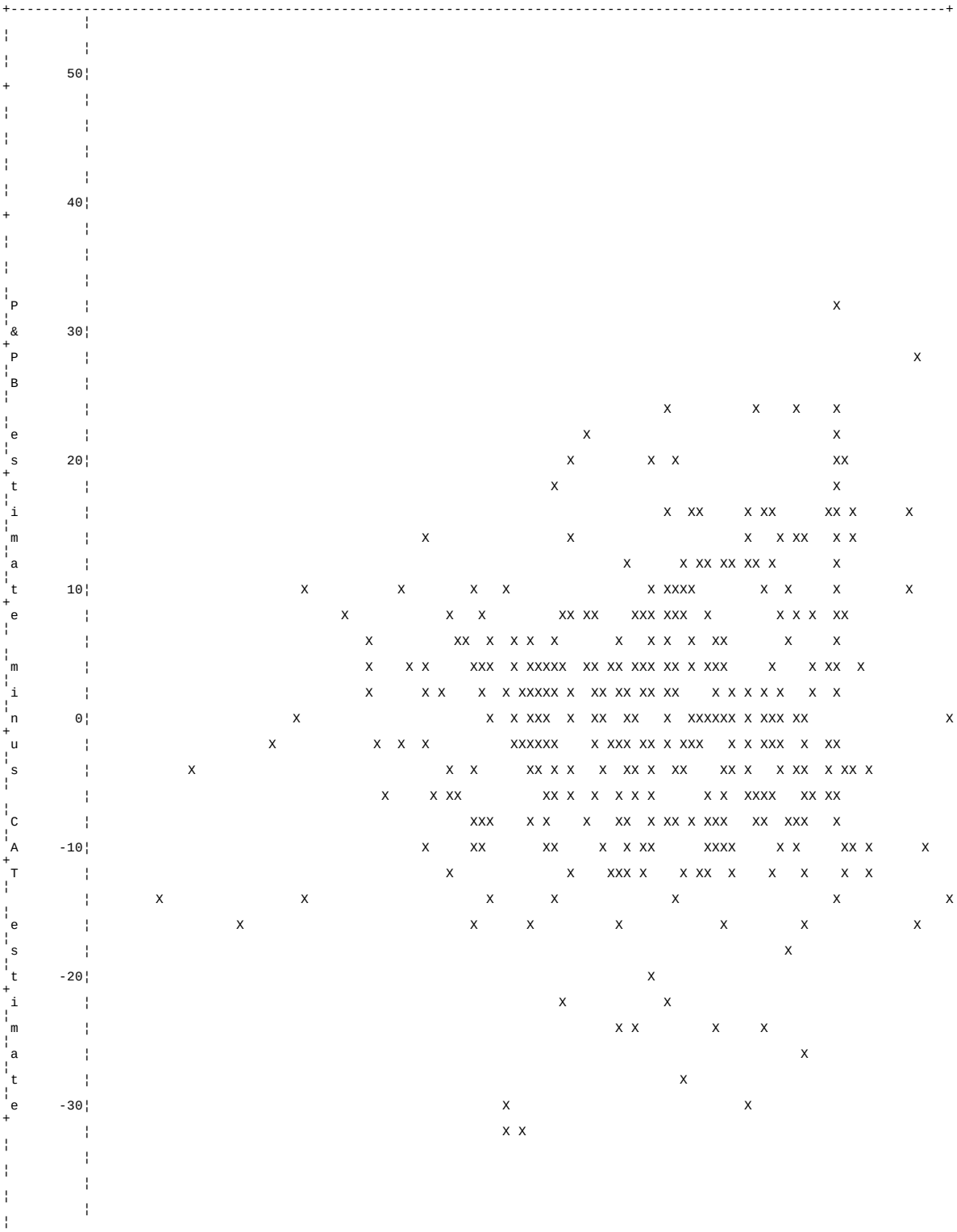


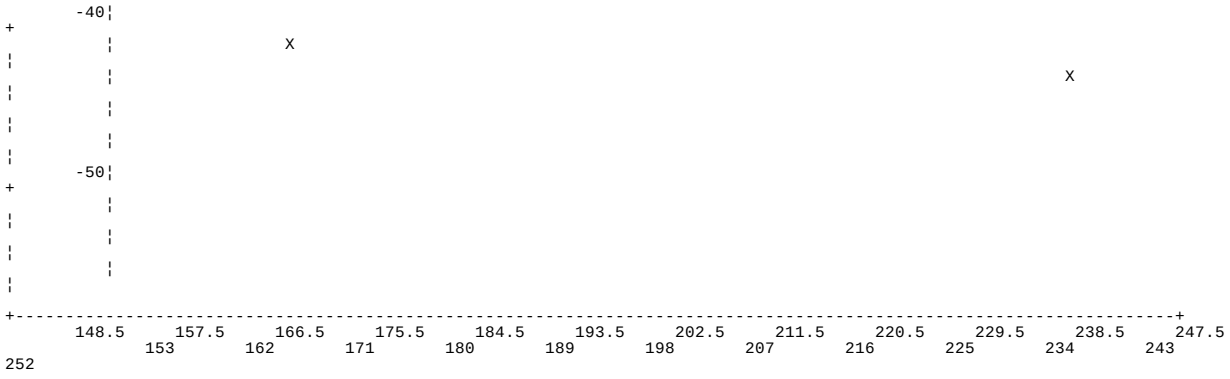


Achievement Level Estimates: P&PA

Figure 6

Difference in Reading Achievement Level Estimates





Achievement Level Estimates: P&PA