DOCUMENT RESUME

ED 382 645 TM 023 079

AUTHOR Eignor, Daniel R.
TITLE Deriving Comparable Scores for Computer Adaptive and Conventional Tests: An Example Using the SAT.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-93-55
PUB DATE Nov 93
NOTE 47p.; Version of a paper presented at the Annual Meeting of the National Council on Measurement in Education (Atlanta, GA, April 13-15, 1993).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Adaptive Testing; *College Entrance Examinations; Comparative Analysis; *Computer Assisted Testing; *Equated Scores; Mathematics Tests; Scaling; *Test Items; Verbal Tests
IDENTIFIERS *Equipercentile Equating; *Linear Equating Method; Paper and Pencil Tests; Scholastic Aptitude Test

ABSTRACT

Procedures used to establish the comparability of scores derived from the College Board Admissions Testing Program (ATP) computer adaptive Scholastic Aptitude Test (SAT) prototype and the paper-and-pencil SAT are described in this report. Both the prototype, which is made up of Verbal and Mathematics computer adaptive tests (CATs), and a form of the paper-and-pencil test were administered to more than 500 examinees using a random groups counterbalanced design. Both linear and equipercentile procedures were used for equating in each of the separate testing orders (paper-and-pencil followed by CAT, or CAT, then paper-and-pencil). Data were not pooled across the orders because the groups were not randomly equivalent due to administrative problems. The linear procedure was chosen for each test (Verbal or Mathematical) for each order, and results from the two orders were averaged. The final Verbal and Mathematical CAT conversions were quite similar to the paper-and-pencil conversions, although the two conversions for Verbal and two conversions for Mathematical did differ by as much as 20 scaled score points in certain regions of the scale. Ten tables and 10 figures illustrate the analysis. (Contains 11 references.) (Author/SLD)

RESEARCH REPORT

DERIVING COMPARABLE SCORES FOR COMPUTER ADAPTIVE AND CONVENTIONAL TESTS: AN EXAMPLE USING THE SAT

Daniel R. Eignor

Educational Testing Service
Princeton, New Jersey
November 1993

Deriving Comparable Scores for Computer Adaptive and Conventional Tests: An Example Using the SAT[1,2]

Daniel R. Eignor[3]

Educational Testing Service

October 1993

---

[1]A previous version of this paper was presented at the annual meeting of NCME, Atlanta, 1993, at a Symposium entitled Practical Problems in the Development of Large Scale Computer Adaptive Tests.

[2]This study was supported by funding from the College Board/ETS Joint Planning Committee.

[3]The author would like to thank Pat Pfeister, Bobbi Goodman, Michelle Rosenthal, Mark Batleman, Mark Pomplun, Ted Blew, and Annie Nellikunnel for their help in completing the study, Paul Holland for advice on how to analyze the data, Nancy Petersen, Gary Marco, and John Mazzeo for review comments, and Linda Ferner for typing the paper.

ABSTRACT

Procedures used to establish the comparability of scores derived from the College Board Admissions Testing Program (ATP) computer adaptive SAT prototype and the paper-and-pencil SAT are described in this report. Both the prototype, which is made up of Verbal and Mathematical computer adaptive tests (CATs), and a form of the paper-and-pencil test were administered to just greater than 500 examinees using a random groups counterbalanced design. Both linear and equipercentile procedures were used for equating in each of the separate testing orders (paper-and-pencil then CAT or CAT then paper-and-pencil). Data were not pooled across the orders because the groups were not randomly equivalent due to administrative problems. The linear procedure was chosen for each test (Verbal or Mathematical) for each order, and results from the two orders were averaged. The final Verbal and Mathematical CAT conversions were quite similar to the paper-and-pencil conversions, although the two conversions for Verbal and two conversions for Mathematical did differ by as much as 20 scaled score points in certain regions of the scale.

Deriving Comparable Scores for Computer Adaptive and Conventional Tests: An Example Using the SAT

Daniel R. Eignor

INTRODUCTION

Recent psychometric and systems advances, coupled with the availability of powerful yet relatively inexpensive microcomputers, have allowed computer adaptive testing (CAT) for large scale testing programs to become a reality at Educational Testing Service (ETS) and other testing organizations. (See Stocking and Swanson, 1992, for a discussion of some of the psychometric and systems advances.) At ETS, a myriad of activities are taking place that are related to the development of operational computer adaptive versions of the Graduate Record Examinations (GRE) General Test and the National Council of State Boards of Nursing (NCSBN) Registered Nurse (RN) and Practical Nurse (PN) exams. In addition, ETSers are working on a computer adaptive Professional Assessments for Beginning Teachers examination called Praxis I: Computer-Based Academic Skills Assessments. This is a test for which no paper-and-pencil counterpart will exist.

With all of this activity taking place to draw upon, The College Board, the major client of ETS, decided to develop a computer adaptive prototype of the Admissions Testing Program (ATP) SAT. Details involving the development of the SAT CAT prototype can be found in a paper by Eignor, Stocking, Way, and Steffen (1993). One important difference, however, between the SAT CAT and the other adaptive tests being developed at ETS is that the SAT CAT prototype was never intended to be used operationally, i.e., to yield scores to be used for admissions purposes. This decision was made for two reasons: 1) the Program did not have a pool of secure items that could be devoted to the CAT and, hence, the CAT pool had to be built from items that had appeared in the past on SAT paper-and -pencil forms that have since been disclosed; and 2) even if a pool of secure items had existed for CAT purposes, no delivery mechanism was in place in the schools to deliver the SAT CAT to the many students who would want to take it during the school year.

The SAT CAT prototype, in the initial planning stage, was thought of as a means of providing colleges that administer forms of the SAT through the Institutional Admissions Testing Program (IATP) with a convenient way of obtaining SAT scores for admitted students who have these scores missing from their records. However, over the course of the development phase of the project, it was decided that the CAT should instead be introduced into selected high schools to examine the feasibility of computer delivery of tests in that setting. The present purpose of the SAT CAT prototype is to provide students with a quick, yet novel, way to get an indication of how well they would do on the present full-length

paper-and-pencil SAT. Such a use necessitated that score comparability between the paper-and-pencil SAT and the SAT CAT be established. The SAT CAT is not alone, however, in regard to the need to establish comparability of scores derived from the two modes of administration.

All testing programs that test via paper-and-pencil examinations and then want to develop computer-based versions, particularly computer adaptive versions, of these examinations face the difficult issue of establishing the comparability of scores derived from the two administrative modes. CAT and paper-and-pencil testing will, at least for some transition period, continue to occur together in these programs. Further, even if paper-and-pencil testing is eventually phased out, scores from the CAT will continue to need to be reported on the reported score scale that had existed for the paper-and-pencil examination. All of these considerations necessitate that a score comparability study be conducted.

Data collection designs for collecting data to equate test forms that are described in the current literature (see Angoff, 1984) were developed for equating parallel forms of examinations administered via the same medium, which for the most part has been paper-and-pencil. It is unclear as to the applicability of such designs to the equating of scores derived from administrations of forms in different mediums, particularly when one score is derived via an adaptive strategy while the other score is developed in a conventional or non-adaptive fashion. However, until new procedures are developed for collecting data to derive comparable scores for CAT and paper-and-pencil examinations, the traditional procedures presented in Angoff (1984) will need to be used. The comparability study described in this paper represents the first attempt at ETS to derive comparable scores on CAT and paper-and-pencil examinations. The study should be viewed in the context provided earlier; viz., that while the CAT scores need to be reported on the existing SAT scales so that students can get a good indication of how well they would do on the paper-and-pencil examination, the CAT scores will never be used operationally for admissions purposes. If the intention had been to use the CAT scores for admissions purposes, a somewhat different data collection design would undoubtedly have at least been considered and the sizes of the samples used in the comparability study would have been much larger. This matter will be discussed further in the discussion section of this paper.

The purpose of this paper is to describe the procedures used to establish the comparability of scores derived from the SAT CAT prototype and the paper-and-pencil SAT. The paper may, in addition, provide a focal point for further discussion of how the comparability of scores on CAT and paper-and-pencil examinations might be established in the future.

METHOD

Participating Schools and Students

Collecting data for the comparability study at regular national test center Admissions Testing Program (ATP) administrations of the SAT was not possible, given the large number of examinees taking the paper-and-pencil examinations at the same time at these administrations and the importance placed on the results of the paper-and-pencil testing. Hence, focus was placed on colleges that administer the SAT through the Institutional Admissions Testing Program (IATP). These colleges administer secure forms of the SAT at their campuses and frequently score the tests themselves, although ETS does maintain a central scoring service for these colleges. SAT scores from IATP administrations may be used for a variety of purposes, for admissions purposes (much in the same way scores from regular national test center administrations are used), for placement purposes, or simply to fill out a student's record.

The state of Georgia mandates that all students entering two and four year colleges and universities, even if already accepted at these colleges or universities, have SAT scores on their records. In addition, many of these schools use SAT scores for placement purposes, and test fairly large numbers of incoming freshmen for fall placement into English and Mathematics classes during the summer orientation period. Hence, institutions in Georgia were seen as an excellent source of data for the comparability study. Thus, a number of institutions in Georgia were contacted to see if they would be interested in administering the SAT CAT during the period of summer orientation when the regular paper-and-pencil SAT would be administered. Three Georgia institutions, two two-year institutions and one four-year institution, all in southern Georgia, agreed to participate in the study. The two-year institutions were Darton College and South Georgia College. The four-year institution was Valdosta State College. ETS contracted with an overall computerized testing coordinator who resides in Georgia, ten IBM 386 personal computers were rented and shipped to each of the three institutions, and the coordinator oversaw the installation/deinstallation of the equipment. All testing took place during 1992 summer orientation periods at each of the three institutions and the tests were administered by the testing coordinators at each institution. Because these periods did not coincide at the three institutions, the rented equipment could be shared across institutions.

Based on projected numbers of incoming freshmen who were to take part in the comparability study at the three Georgia institutions, it was determined that additional testing would need to take place to augment the total sample size. Invitations to participate were placed in newspapers in the Princeton, New Jersey area and in the ETS weekly newspaper. Students interested in participating, who had to be juniors or seniors in high school, were tested at the permanent ETS institutional computer-based testing center at Rider College. These students were paid $25 for taking both the paper-and-pencil SAT and the CAT, their paper-and-pencil fees were waived, and the students were given the option of having their paper-and-pencil SAT scores added to their national score records, which the student has sent

to institutions to which they are applying. The students had to make this decision after testing and before they saw their paper-and-pencil scores. Hence, it was felt that the testing for the comparability study at Rider College was done under conditions under which the students would be reasonably motivated to do well.

Incoming freshmen at the three Georgia institutions were paid $30 to participate in the comparability study. In addition, each of the institutions was offered an honorarium. Since the paper-and-pencil SAT scores for each of these students was to be used for fall placement purposes, it was felt that students would be motivated to perform well, particularly on the paper-and-pencil test. It was hoped that the students would also be motivated when they took the CAT. Because of the novel and unique nature of the CAT experience, it was felt that the students would be interested in the CAT and would attempt to perform well. (Questionnaire data, not included with this paper, bears out the fact that there was a good deal of interest in the CAT.)

Data Collection Design

Because of the number of examinees anticipated for the comparability study, it was determined early in the planning process that the data collection design to use would be a random groups counterbalanced design with both tests administered to each group (Design II in Angoff, 1984). Students were to be randomly assigned, on a within-school or testing center (i.e., Rider College) basis, to the two possible testing orders, CAT then paper-and-pencil and paper-and-pencil then CAT. For the sample sizes initially anticipated (around 400 students), the random groups counterbalanced design provides much smaller standard errors of equating then do the two other designs that could have been considered, the random groups design with one test administered to each group and the non-equivalent groups, common item design. (See Angoff, 1984, for a discussion of these designs and the standard errors or Lord, 1950, for a discussion and comparison of the standard errors.)

Practitioners who have recently conducted studies that have attempted to establish the comparability of paper-and-pencil and linear computer-based test (CBT) scores (i.e., an intact paper-and-pencil test is simply administered on a computer) via the random groups counterbalanced design have run into the problem of asymmetric practice effects (see Mazzeo and Harvey, 1988). The standard procedure for dealing with data from the random groups counterbalanced design described in Angoff (1984), which calls for pooling the summary statistics (i.e., means and standard deviations) from the two possible test orders, assumes that any practice effects that result from the testing experience are constant and symmetric. With little experience on which to base a decision and virtually nothing written on the subject of equating CATs to paper-and-pencil tests, the assumption of symmetric practice effects was seen as extremely tenuous, and, hence, plans were made to equate separately in the two orders and then form some sort of average. For both orders, scores for the CAT were to be equated to scores on the paper-and-pencil test, for which a raw to scaled score conversion table already existed. This approach of equating separately in the two orders and then averaging the two equating functions has been discussed by Holland and Thayer (1990).

However, as will be discussed later in the paper, for another more fundamental reason than asymmetric practice effects, separate equatings had to be done for the two orders and averaged in this study.

As mentioned previously, students were to be randomly assigned to the two testing orders (CAT then paper-and-pencil or paper-and-pencil then CAT) on a within-college or testing center basis. Further, the testing coordinators at the four sites were given the option of administering the CAT and paper-and-pencil tests on the same day or on different days. (A combination of procedures, where one group of students took both the paper-and-pencil test and the CAT on the same day and another group took the tests on different days, was also possible.) Figure 1 contains a description of the two designs that was sent to testing coordinators at each of the four sites. Figure 2 contains the detailed procedures sent to these coordinators for splitting the total group to be tested, either on a given day or during the entire testing session, into random subgroups.

Insert Figures 1 and 2 about here

Tests Administered

The paper-and-pencil SAT form that was administered in the study was a secure form developed for the national Admissions Testing Program (ATP) and then designated for use in the Institutional Admissions Testing Program (IATP). The form consisted of four thirty minute sections given to all examinees in the same fixed order, with the first test section being a section that contained SAT-M items. It should be noted that the variable section of the SAT is removed for IATP administrations and the section containing the Test of Standard Written English (TSWE) was specifically removed for this study. Hence, the test contained four sections rather than the usual six. The two thirty minute SAT-V sections contained 45 and 40 items, respectively, while the two thirty minute SAT-M sections contained 35 and 25 items, respectively. The total 85-item SAT-V contained the usual four item types: sentence completion, analogies, antonyms, and reading comprehension items while the total 60-item SAT-M contained the usual two item types: five-choice regular math items and four-choice quantitative comparison items. (All SAT-V items are five-choice.) Table 1 contains a breakdown of the number of items by item type for SAT-V and SAT-M and an additional breakdown of the total 60-item SAT-M by content area.

Insert Table 1 about here

The SAT-V CAT administered to examinees was a fixed length CAT of 27 items and the SAT-M CAT was a fixed length CAT of 20 items. The development of the CAT item pools and the specifics of the SAT CATs are described in a paper by Eignor, Stocking, Way,

and Steffen (1993). Table 1 contains a breakdown by item type of the number of items in the SAT-V and SAT-M CATs and the number of items in the total CAT pools. The numbers of items for the various item types on the CATs are basically proportional to the numbers of items for the item types that are contained on the full-length 85- and 60-item paper-and-pencil tests. Table 1 also contains a breakdown of the SAT-M CAT and SAT-M item pool by content area.

Unlike the paper-and-pencil tests, which were given to each examinee in the same fixed order, examinees were allowed to choose which CAT, Verbal or Math, they wanted to take first. If an examinee, for example, chose the Verbal CAT to take first, after introductory material and the tutorials, he/she was administered 27 Verbal items in up to 40 minutes, followed by a brief pause and then 20 Math items in up to 40 minutes. Examinees were not allowed to omit items on the CATs nor were they allowed to review responses to earlier items (i.e., examinees could progress only in a forward fashion).

All examinees took the two CATs or the four sections of the paper-and-pencil test on the same day. As mentioned in the previous section, examinees could either be administered all testing material (the two CATs and the four paper-and-pencil sections) on the same day or they could receive the CATs on one day and the paper-and-pencil test on another.

Scores to be Equated

For the paper-and-pencil test, scoring was straightforward. The score for each examinee on the 85-item SAT-V was created via formula scoring, using the formula R-¼W for five-choice items. The score for each examinee on the 60-item SAT-M was created via formula scoring, using the formula R-¼W for the 40 regular five-choice items and R-⅓W for the 20 four-choice quantitative comparison items. The separate scores for the two item types were then summed and rounded to the nearest integer, as was the formula score for SAT-V. Hence, rounded formula scores for the paper-and-pencil SAT-V and SAT-M were used in the equatings.

For the CATs, scoring was relatively straightforward, but involved some intermediate steps. As part of the CAT system, the paper-and-pencil test administered to examinees was imbedded as a "reference test". That is, the paper-and-pencil test, with associated three parameter logistic (3-PL) item parameter estimates, was embedded for score creation purposes; the items on the reference test were not used in any of the CATs. An examinee's final ability estimate ($\hat{\theta}$) on SAT-V, derived after administration of 27 items or however many items the examinee completed in 40 minutes (see a later section of the paper for how not reached items were treated), was then used with the 3-PL item parameter estimates on the 85-item SAT-V reference test to create an estimated true formula score for the examinee on the reference test. (See Lord, 1980, p.230, for the formula (15.6) to create estimated true formula scores.) This true formula score was then rounded to the nearest integer. Exactly the same procedure was used with the examinee's final $\hat{\theta}$ on SAT-M, derived after administration of 20 items or however many items the examinee completed in 40 minutes.

Hence, rounded estimated true formula scores on the reference test were used as the SAT-V and SAT-M CAT scores in the equating. Finally, and worth noting again, the paper-and-pencil and CAT scores used in the equatings are both scores on the same form. This does not qualify as an "equating" in the usual sense of the word in that scores on different parallel forms of the same instrument aren't being used. Rather, the scores being used are scores on the same form developed through administrations in two different modes--observed formula scores derived from administration in paper-and-pencil mode and estimated true formula scores derived from administration in adaptive mode via a computer.

RESULTS

Numbers of Examinees Tested and Deletion of Cases

The number of examinees at each of the colleges/test centers taking the paper-and-pencil SAT and the SAT CAT on the same and different days are presented in Table 2. Names of individual colleges/test centers are not identified in Table 2 and other related tables. Instead, the college/test centers are referred to as College/Centers A-D. Also presented in Table 2 are the number of examinees who took the CAT first and the number who took the paper-and-pencil test (abbreviated as P-P) first on the same or different days.

Insert Table 2 about here

Although fairly elaborate instructions were prepared for splitting the total groups of examinees to be tested into randomly equivalent (i.e., counterbalanced) subgroups (see Figure 2), it is clear from the data contained in Table 2 that the counterbalancing procedures were not closely followed. A review of the number of tests given per day at each of the four colleges/testing centers indicated that only at two of them were counterbalancing procedures closely followed each day. Hence, pooling of data from the two testing orders was clearly not possible, i.e., the groups taking the two orders were not randomly equivalent, and separate Verbal and Math equatings for each of the two orders needed to be performed.

Before any analyses could take place, examinees' CAT and paper-and-pencil records had to be matched. This was done by matching on candidates' ID numbers (the first eight digits of their social security numbers). In the process of matching, it was found that a number of examinees had not taken both the CAT and the paper-and-pencil test. In addition, for a number of examinees, the records could not be matched. Finally, the ten examinees from College/Center B and College/Center C who took the CAT and paper-and-pencil tests on different days were dropped from the data sets; clearly no attempt was made with these students to form counterbalanced groups. Table 3 contains the number of examinees remaining in the data sets after matching CAT and paper-and-pencil records and removing examinees with incomplete data or who were inappropriately tested.

Insert Table 3 about here

Incomplete CATs

While the timing of the CATs was seen as more than ample (40 minutes for 27 SAT-V items, 40 minutes for 20 SAT-M items), it was anticipated that not all examinees would complete the CATs. In lieu of a formal study, a somewhat arbitrary rule was put into place that an examinee had to complete at least 75% of each of the CATs, i.e., 21 SAT-V items and 15 SAT-M items, in order to be included in the study. For examinees completing more that 75% of one or both of the CATs but less than 100%, the final $\hat{\theta}$ used for creation of an estimated true formula score would be the $\hat{\theta}$ derived after the last item attempted.

Five examinees failed to complete the SAT-M CAT, but all of these examinees completed at least 15 items. Eleven examinees failed to complete the SAT-V CAT, but all of these examinees completed at least 21 items. Hence, no examinees were eliminated from the comparability study based on the 75% completion rule.

Summary Data by Institution and for Total Groups

Table 4 contains CAT and paper-and-pencil summary data (means, standard deviations, correlations and sample sizes) for SAT-V separately by testing order for each of the four colleges/testing centers and then for the total groups. Table 5 contains comparable data for SAT-M. The numbers in parentheses in Table 4 and 5 are the summary statistics and sample sizes after removal of outlying pairs of scores; this procedure will be described in a subsequent section of the paper.

Insert Tables 4 and 5 about here

As can be seen from the data in Tables 4 and 5, there is a good deal of variation in average performance across the four colleges /testing centers, with the weakest performers being the examinees from College/Center A and the strongest performers being the examinees from College/Center D. Outside of the somewhat lower correlations for the SAT-M CAT and paper-and-pencil test scores for examinees from College/Center A, particularly for the paper-and-pencil test taken first order, no other data in Tables 4 and 5 appears peculiar. The CAT/paper-and-pencil correlations for the total groups are particularly high, and for the SAT-M CAT taken first order, the correlation (.933) is almost as high as could possibly be expected given the reliabilities of the CAT and paper-and-pencil tests (neither of which are estimated to exceed .94).

Outlier Analysis

Although the CAT and paper-and-pencil correlations for the four testing orders (two for SAT-V and two for SAT-M) were quite high, initially ranging from .897 to .927, it was felt that the correlations might be improved upon if a bivariate outlier analysis was performed on each order, and outlying pairs of scores removed. For each of the two orders for SAT-V and for SAT-M, a bivariate plot of standardized scores was created, with standardized paper-and-pencil scores on the abscissa and standardized CAT scores on the ordinate. Figures 3 and 4 contain the two SAT-V plots while Figures 5 and 6 contain the comparable plots for SAT-M. Each point in a plot is based on the standardized paper-and-pencil and CAT scores for a particular examinee. Looking at Figures 3-6, there do appear to be some outliers, but for the most part, the shapes of the ellipses formed by the complete sets of points reflect the high correlations seen between the scores.

Insert Figures 3-6 about here

To determine which outlying sets of points to possibly exclude, a criterion suggested by Barnett and Lewis (1984, p.245) was applied; this criterion is based on a multivariate normal model. In the bivariate case, the criterion function can be written as

$$R = \frac{1}{1-r_{xy}^2} (X^2 - 2r_{xy} XY + Y^2)$$

An observation (pair of standardized scores X and Y) is considered an outlier, i.e., not a member of the same population as the other observations, at the α level of statistical significance if

$$R > -2\ln [1 - (1 - \alpha)^{1/N}] ,$$

where N is the total sample size.

For $\alpha = .05$ and the SAT-V CAT first order (N=271), R must exceed 17.1 in order for an observation to be considered an outlier. For $\alpha = .01$, R must exceed 20.4. Comparably sized cutoffs result for the other three orders. However, because of the high CAT/paper-and-pencil correlations, across all four orders the highest R seen for a particular observation was 15.1. Hence, at the $\alpha = .05$ level, none of the observations across all four orders would be considered an outlier if the Barnett and Lewis statistical criterion was used.

After further study of the bivariate plots and the R values for the observations for all four orders, it was decided that an arbitrary value of $R > 7$ would be used as the cutoff for deciding on which observations qualified as outliers. Observations with $R > 7$ are circled in Figures 3-6, with R values printed alongside the points. These score pairs were then deleted from the datasets and the summary statistics in Tables 4 and 5 were recalculated and are presented in parentheses in these tables. The data sets with these outliers removed were then used in the subsequent equatings.

Removal of these outliers clearly improved the correlations between scores for each of the four orders. Moreover, the scores removed clearly are the outliers in Figures 3-6. In sum, although the outlier analysis done for this study was based on an arbitrary criterion, various results indicate the effectiveness of the deletion process.

Total Group Means and Frequency Distributions

Total group means and standard deviations for the two orders for SAT-V and for SAT-M were extracted from Tables 4 and 5 and are presented in Table 6. Noteworthy observations about the data contained in Table 6 are: 1) for three of the four orders, there is a decrease in average performance on the test taken second when compared to average performance on the test taken first (the SAT-V paper-and-pencil test taken first order being the exception); 2) for both orders for SAT-V, the paper-and-pencil standard deviations are less than the CAT standard deviations; and 3) for both orders for SAT-M, the paper-and-pencil standard deviations are greater than the CAT standard deviations. The finding that, for three of the four orders, there is a decrease in performance on the test taken second, presumably due to a fatigue effect, runs somewhat counter to the findings in most of the studies reviewed by Mazzeo and Harvey (1988), where there appeared to be a practice effect on the test taken second, although the practice effects were frequently not symmetric. Also, the studies reviewed by Mazzeo and Harvey involved instances where linear computerized tests were equated to paper-and-pencil tests, not CATs equated to paper-and-pencil tests.

Insert Table 6 about here

Figure 7 contains grouped frequency distributions of estimated true formula scores from the SAT-V CAT and observed formula scores from the paper-and-pencil SAT-V for the two testing orders. Figure 8 contains comparable grouped frequency distributions for SAT-M. Only one examinee obtained a maximum possible score (an observed formula score of 60 on the paper-and-pencil SAT-M). Because the total sample sizes are fairly small for the four orders, score frequencies in the ungrouped frequency distributions (not presented in the paper) are often extremely small and data are sparse in certain regions. Clearly, if a curvilinear equating procedure were to be used to establish comparable scores for the CAT, these frequency distributions would need to be smoothed.

Insert Figures 7 and 8 about here

Equatings Performed and Final Unrounded Conversions

As mentioned earlier, the datasets after outlying sets of scores were removed were then used in four single-group equatings (two for SAT-V, two for SAT-M). The N's used in these equatings, which were presented in parentheses in Tables 4 and 5, are: SAT-V CAT first, N = 266; SAT-V P-P first, N = 230, SAT-M CAT first, N = 267; and SAT-M P-P first, N = 230. For each of the four single-group equatings, two procedures were used:

1. A linear procedure based on setting CAT and paper-and-pencil standard deviates equal; and

2. A curvilinear procedure based on an equipercentile equating of unsmoothed CAT and paper-and-pencil score distributions.

For each order for each test, the (raw-to-raw) linear and curvilinear equating functions were compared to see if there was any evidence of a curvilinear relationship between CAT and paper-and-pencil scores. This was done through the use of difference plots, with the linear conversion used as the criterion and differences between the curvilinear and linear conversions plotted with respect to the linear conversion. The two SAT-V plots are shown in Figure 9 and the two SAT-M plots are shown in Figure 10. In each plot, the zero difference or straight line is based on the linear conversion and the non-linear curve is based on differences between the curvilinear and linear conversions across all obtained score points. If the relationship between CAT and paper-and-pencil scores is curvilinear, this latter curve will appear to be a convex or concave curve with respect to the zero difference line or, in certain instances, an S-shaped curve.

Insert Figures 9 and 10 about here

Looking at the four plots contained in Figures 9 and 10, in no instance does there appear to be any real evidence of curvilinearity in the (raw-to-raw) relationship between CAT and paper-and-pencil scores. Hence, the linear procedure was chosen for each of the four orders and another set of linear equatings were performed, this time reading in the raw-to-scale conversion table for the paper-and-pencil SAT-V and the paper-and-pencil SAT M, so that the output would contain SAT-V and SAT-M CAT raw-to-scale conversion tables for each of the orders reflecting the results of the equating process.

Table 7 contains the unrounded raw-to-scale SAT-V conversions resulting form the two orderings, CAT taken first and paper-and-pencil taken first. Also contained in Table 7

is the unrounded paper-and-pencil raw-to-scale conversion. The two CAT conversions are very similar, differing by a maximum of 8.81 score points (on the unrounded 200 to 800 scale) at the maximum formula score of 85. Because the two conversions are so similar, a decision was made to simply average the two separate conversions in deriving the final unrounded SAT-V CAT conversion table. This unrounded average conversion is also presented in Table 7.

Insert Table 7 about here

Table 8 contains the unrounded raw-to-scale SAT-M CAT conversions resulting from the two orderings, CAT taken first and paper-and-pencil taken first. Also contained in Table 8 is the unrounded paper-and-pencil raw-to-scale conversion. Because each of the CAT conversions is higher at the top than the paper-and-pencil conversion and lower at the bottom than the paper-and-pencil conversion, there are missing points at the top and at the bottom of each of the CAT conversions. If the conversions are reasonably linear, missing conversion points can be established via linear interpolation.

Insert Table 8 about here

Unlike SAT-V, the two SAT-M CAT conversions presented in Table 8 are quite dissimilar. At a formula score of 55, the two conversions differ by 30.15 points (on the unrounded 200 to 800 scale). In addition, the paper-and-pencil first SAT-M CAT conversion is a good deal more discrepant from the original SAT-M paper-and-pencil conversion than the CAT first SAT-M CAT conversion. Since in this study, scores are being created on two different "versions" of the same test form (one score being created via the CAT process and the other from regular paper-and-pencil testing), it is reasonable to expect that the CAT conversion will fairly closely approximate the original paper-and-pencil conversion. Given this, the SAT-M CAT conversion resulting from the paper-and-pencil then CAT testing order is clearly the outlier. A decision was made not only to simply average the CAT conversions from the two orders, but also to form weighted averages where the CAT conversion from the CAT then paper-and-pencil order counted two (2:1) and three (3:1), times as much as the CAT conversion from the paper-and-pencil then CAT order. (Although the CAT conversion from the paper-and-pencil then CAT order was so discrepant, a rationale for completely discarding this conversion could not be generated.) After review of the weighted averages, it was decided that the most extreme of the weighted averages that could be justified was the 2:1 weighted average. (In addition, the CAT conversion from the 3:1 weighting provided much the same results as the 2:1 weighted average when rounded scores were used.) The 2:1 weighted average is presented in Table 8 along with the straight unweighted average. Missing conversion points for the 2:1 weighted average (for formula scores 56-60 and -16 and -17) were determined via linear interpolation using the adjacent five formula score points

at the top and at the bottom that had conversion points. Finally, the 2:1 weighted average was used to create the final rounded SAT-M CAT conversion to be used for score reporting purposes.

Doglegs and Final Rounded Conversions

Table 9 presents the final SAT-V CAT raw-to-scale conversion using both unrounded and rounded (reported) scaled scores. This final SAT-V CAT conversion was formed by simply averaging the conversions derived from the linear equatings in the two separate orders. Also presented in Table 9 are the unrounded and rounded (reported) raw-to-scale conversions for the form given in paper-and-pencil mode. As can be seen in Table 9, for higher formula scores the CAT raw-to-scale conversion is lower than the paper-and-pencil raw-to-scale conversion, sometimes as much as 20 scaled score points on the rounded scale. This is a direct outcome of the fact that the CAT estimated true formula score standard deviations were greater in both orders than the paper-and-pencil observed formula score standard deviations. (This is reflected in a slope parameter that is less than one in the linear equation derived by setting CAT and paper-and-pencil standard deviates equal.) The conversion for the form given in paper-and-pencil mode did not scale to 800, which is an ATP Program requirement, so a dogleg (see Braun and Holland, 1982) had to be fit to the top of the conversion to allow a formula score of 85 to scale to 800. (This dogleg is presented in parentheses in Table 9.) Because the SAT-V CAT conversion is lower than the paper-and-pencil conversion at the top, a dogleg encompassing more scaled score points had to be fit to the top of the CAT raw-to-scale conversion (also presented in parentheses). In both cases, the doglegs formed were established to allow a smooth progression of scores with the maximum formula score (85) reaching 800. Finally, the CAT raw-to-scale conversion presented at the far right of Table 9, under the column labeled "Reported", is the conversion embedded in the CAT system for on-screen score reporting purposes.

Insert Table 9 about here

Table 10 presents the final SAT-M CAT raw-to-scale conversion using both unrounded and rounded (reported) scaled scores. This final SAT-M conversion was formed by creating a weighted average of the conversions derived from linear equatings in the two separate orders, counting the CAT then paper-and-pencil conversion twice as much as the paper-and-pencil then CAT conversion. Also presented in Table 10 are the unrounded and rounded (reported) raw-to-scale conversions for the form given in paper-and-pencil mode. As can be seen in Table 10, for higher formula scores the CAT raw-to-scale conversion is higher than the paper-and-pencil raw-to-scale conversion, sometimes as much as 20 scaled score points on the rounded scale. This is a direct outcome of the fact that the CAT estimated true formula score standard deviations were smaller in both orders than the paper-and-pencil observed formula score standard deviations. (This is reflected in a slope parameter that is greater than one in the linear equation derived by setting CAT and paper-

and-pencil standard deviates equal.) The conversion for the form given in paper-and-pencil mode did not quite scale to 800, so a dogleg had to be fit, but only at the very top of the conversion (at a formula score of 60). Because the SAT-M CAT conversion is higher at the top than the paper-and-pencil conversion, a dogleg was not necessary. As with SAT-V, the CAT raw-to-scale conversion presented at the far right of Table 10, under the column labeled "Reported", is the conversion embedded in the CAT system for on-screen score reporting purposes.

Insert Table 10 about here

DISCUSSION

Because scores were being created on the same test form via administrations done in two different ways for the comparability study described in this paper, in an adaptive fashion via computer and in a conventional fashion via paper-and-pencil, it was anticipated that the relationship between the CAT and paper-and-pencil scores would likely be linear and that the resulting Verbal and Math CAT raw-to-scale conversions would be quite similar to the Verbal and Math paper-and-pencil raw-to-scale conversions. While all equating relationships in the study appeared to be quite linear, the final CAT conversion tables were not as similar to the paper-and-pencil conversion tables as expected. In addition, a somewhat different outcome occurred with the final SAT-V CAT conversion than occurred with the final SAT-M CAT conversion. In the case of SAT-V, the final CAT conversion was lower than the paper-and-pencil conversion at the top of the scale while for the SAT-M CAT, just the opposite occurred--the CAT conversion was higher than the paper-and-pencil conversion at the top of the scale. As mentioned earlier, this was the direct result of the differences observed in the CAT estimated true formula score and paper-and-pencil observed formula score standard deviations. In the case of SAT-V, the CAT standard deviations were greater than the paper-and-pencil standard deviations in both orders, while for SAT-M the CAT standard deviations were less than the paper-and-pencil standard deviation in both orders. It would seem, at this point, that these results are somehow related to unexplained differences in the SAT-V and SAT-M CAT test taking experiences. One possible explanation now being explored has to do with differences in percentages of examinees completing the SAT-V CAT and paper-and-pencil tests versus differences in percentages of examinees completing the SAT-M CAT and paper-and-pencil tests and the relationship of these differences to ability level.

One clear outcome of this study is that the random groups counterbalanced equating design should probably be avoided in comparability studies of this sort. Even though fairly elaborate directions for counterbalancing were created, these directions were not followed and the groups taking the tests in the two orders could not be considered randomly equivalent. However, even if the counterbalancing directions had been followed, results from the Mazzeo and Harvey (1988) review indicate that the effects of having taken a particular sort of test first in a random groups counterbalanced design, like a CAT, are likely

not to be the same as the effects of having taken another sort of test,like a paper-and-pencil test, first in this design. It would appear that, in the process of equating a CAT, or any computer-based test, to a paper-and-pencil test, it is a bad idea to set up a design where examinees take the tests to be equated sequentially. It should be noted that this observation holds equally well for the common item, non-equivalent groups equating design, if the anchor test is an externally administered block of items given in either paper-and-pencil or computer format. That is, suppose one is attempting to equate scores on a CAT to scores on a paper-and-pencil test, using the common item, non-equivalent groups design, and the common items are a set of external items (external to the CAT and the paper-and-pencil form) given in paper-and-pencil format after the CAT or the paper-and-pencil test. In this case, it is highly likely that the experience of taking the CAT first will influence performance on the external set of common items in a way that is different from the experience of having taken the paper-and-pencil form first. In short, a design is needed where groups of examinees take the tests to be equated in either one mode or the other. The random groups design is such a design, but the standard errors of equating associated with this design necessitate much larger sample sizes than do the two designs just discussed.

In sum, although the CAT raw-to-scale conversions in this study differed more from the paper-and-pencil raw-to-scale conversions than had been anticipated, the conversions were viewed as acceptable, given the purposes for which the SAT CAT prototype was developed. This is not to say that the equatings, or more precisely, the size and nature of the samples used in the equatings, would have been viewed as completely adequate if the scores from the SAT CAT were to be used for actual admissions purposes. It is clear that if a computer adaptive version of the SAT is ever constructed from a pool of secure SAT items, and the resulting scores are to be used for actual admissions purposes, a greater level of attention will need to be paid to equating and data collection activities. Based on the results of this study, should this activity take place in the future, it is recommended that a random groups with one test administered to each group equating design be used to establish the comparability of scores on the CAT and the paper-and-pencil test.

References

Angoff, W. H. (1984). Scales, norms, and equivalent scores. Princeton, NJ: Educational Testing Service.

Barnett, V., and Lewis, T. (1984). Outliers in statistical data. New York: Wiley.

Braun, H. I., and Holland, P.W. (1982). Observed score test equating: A mathematical analysis of some ETS equating procedures. In P.W. Holland and D. B. Rubin (Eds.), Test equating. New York: Academic Press.

Eignor, D. R., Stocking, M. L., Way, W. D., and Steffen, M. (1993, April). Case studies in computer adaptive test design through simulation. Paper presented at the annual meeting of NCME, Atlanta.

Holland, P. W., and Thayer, D. T. (1990, April). Kernel equating and the counterbalanced design. Paper presented at the annual meeting of AERA, Boston.

Lord, F. M. (1950). Notes on comparable scales for test scores (RB-50-48). Princeton, NJ: Educational Testing Service.

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

Mazzeo, J., and Harvey, A. L. (1988). The equivalence of scores from automated and conventional educational and psychological tests: A review of the literature (College Board Report No. 88-8). New York: College Entrance Examination Board.

Stocking, M. L., and Swanson, L. (1992). A method for severely constrained item selection in adaptive testing (RR-92-37). Princeton, NJ: Educational Testing Service.

Table 1

Numbers of Items by Item Type and Content Area
in the SAT-V and SAT-M Paper-and-Pencil Test,
CAT, and CAT Item Pools

SAT-V

	Reading Comprehension Items	Antonym Items	Analogy Items	Sentence Completion Items	Total Number of Items
Paper-and-Pencil Test	25[1]	25	20	15	85
CAT	8[2]	8	6	5	27
Verbal CAT Pool	91[3]	74	51	87	303

SAT-M

	Regular 5-Choice Items	4-Choice Quantitative Comparison Items	Arithmetic Items	Algebra Items	Geometry Items	Miscellaneous Items	Total Number of Items
Paper-and-Pencil Test	40	20	18-19	17	16-17	7-9	60
CAT	13	7	5-6	6	6	2-3	20
Math CAT Pool	128	107	70	65	66	34	235

[1]Based on 5 or 6 passages with 3 to 5 items per passage

[2]Based on 3 passages, with 2 passages having 3 items each and 1 passage having 2 items

[3]Based on 27 passages, having from 3 to 6 items per passage, of which either 2 or 3 items are chosen for a CAT

Table 2

Number of Examinees by College/Testing Center
Taking the Paper-and-Pencil SAT and SAT CAT
on Same and Different Days

College/Center	Paper-and-Pencil and CAT on Same Day		Paper-and-Pencil and CAT on Different Days		Total
	CAT First	P-P First	CAT First	P-P First	
A	21	12	23	29	85
B	102	83	0	7	192
C	71	64	0	3	138
D	61	55	-	-	116

Examinees	Overall Totals
Both Orders	521
CAT Taken First	278
Paper-and-Pencil Taken First	243

Table 3

Number of Examinees After Matching and Removal
Because of Incomplete Records/Inappropriate Testing

Examinees	Overall Totals
Both Orders	506
CAT Taken First	271
Paper-and-Pencil Taken First	235

Table 4

SAT-V CAT and Paper-and-Pencil Summary Data
by Testing Order for Each of the Four Colleges/Testing
Centers and for the Total Groups

CAT Taken First

		College/Center A	College/Center B	College/Center C[3]	College/Center D	Totals
CAT	$\bar{X}$[1]	27.61 (27.83)[2]	29.68 (29.29)	31.41	46.97 (46.71)	33.70 (33.48)
	SD	13.72 (13.82)	17.00 (16.94)	18.43	16.65 (16.59)	18.27 (18.22)
P-P	$\bar{X}$	28.88 (28.63)	28.12 (27.76)	30.45	45.74 (45.59)	32.79 (32.54)
	SD	12.92 (12.98)	16.01 (15.53)	17.64	16.38 (16.60)	17.50 (17.41)
	r	.819 (.842)	.888 (.903)	.928	.874 (.903)	.907 (.919)
	N	41 (40)	100 (98)	69	61 (59)	271 (266)

Paper-and-Pencil Taken First

		College/Center A	College/Center B	College/Center C[3]	College/Center D	Totals
P-P	$\bar{X}$	23.08 (23.30)	27.74 (27.62)	31.36	40.13 (40.06)	30.87 (30.83)
	SD	13.05 (13.16)	14.38 (14.55)	18.00	18.09 (18.08)	17.06 (17.08)
CAT	$\bar{X}$	21.24 (21.86)	30.05 (29.96)	31.31	39.07 (38.87)	31.08 (31.08)
	SD	15.04 (14.74)	17.18 (16.89)	18.46	16.30 (16.54)	17.81 (17.65)
	r	.871 (.878)	.874 (.896)	.930	.886 (.923)	.897 (.913)
	N	38 (37)	78 (76)	64	55 (53)	235 (230)

[1]Means are estimated true formula score or observed formula score means on the 85-item SAT-V.

[2]Data in parentheses were derived after removal of outlying pairs of scores.

[3]There were no outlying pairs of scores for this college/center.

Table 5

SAT-M CAT and Paper-and-Pencil Summary Data
by Testing Order for Each of the Four Colleges/Testing
Centers and for the Total Groups

CAT Taken First

		College/Center A[3]	College/Center B	College/Center C	College/Center D[3]	Totals
CAT	$\bar{X}$ [1]	14.42	17.52 (17.62)[2]	21.49 (20.81)	34.05	21.78 (21.68)
	SD	10.31	12.98 (13.09)	15.69 (15.38)	12.83	15.00 (14.96)
P-P	$\bar{X}$	14.63	16.81 (16.90)	20.59 (19.88)	34.57	21.44 (21.34)
	SD	11.28	13.55 (13.56)	16.67 (16.30)	12.57	15.67 (15.58)
	r	.861	.894 (.907)	.923 (.929)	.927	.927 (.933)
	N	41	100 (98)	69 (67)	61	271 (267)

Paper-and-Pencil Taken First

		College/Center A[3]	College/Center B	College/Center C[3]	College/Center D	Totals
P-P	$\bar{X}$	11.11	15.39 (15.25)	21.55	29.86 (29.44)	19.76 (19.53)
	SD	9.08	13.34 (13.49)	14.55	13.26 (13.38)	14.58 (14.57)
CAT	$\bar{X}$	9.63	14.10 (14.47)	19.66	27.60 (26.48)	18.05 (17.83)
	SD	9.04	13.00 (12.95)	13.42	13.19 (12.63)	14.00 (13.64)
	r	.784	.888 (.916)	.928	.882 (.902)	.913 (.922)
	N	38	78 (76)	64	55 (52)	235 (230)

[1]Means are estimated true formula score or observed formula score means on the 60-item SAT-M.

[2]Data in parentheses were derived after removal of outlying pairs of scores.

[3]There were no outlying pairs of scores for the particular testing order at this college/center.

Table 6

SAT-V and SAT-M CAT and Paper-and-Pencil
Means and Standard Deviations for Total Groups

SAT-V

CAT Taken First			Paper-and-Pencil Taken First		
	CAT	P-P		P-P	CAT
$\bar{X}$	33.48	32.54	$\bar{X}$	30.83	31.08
SD	18.22	17.41	SD	17.08	17.65

SAT-M

CAT Taken First			Paper-and-Pencil Taken First		
	CAT	P-P		P-P	CAT
$\bar{X}$	21.68	21.34	$\bar{X}$	19.53	17.83
SD	14.96	15.58	SD	14.57	13.64

Table 7

SAT-V PAPER-AND-PENCIL AND CAT
UNROUNDED RAW-TO-SCALE CONVERSIONS

Formula Score	Paper-and-Pencil Conversion	CAT Conversions		
		CAT Taken First	P-P Taken First	Average
85	765.7534	740.9539	749.7674	745.3607
84	760.3660	734.2651	742.8217	738.5434
83	749.7511	726.9687	736.1918	731.5802
82	742.5784	719.7467	728.8554	724.3010
81	735.7382	713.1658	721.4636	717.3147
80	728.1016	706.7055	714.6908	710.6981
79	720.4621	697.5467	708.1604	702.8536
78	713.5629	688.6647	699.4463	694.0555
77	706.8445	681.5607	690.3917	685.9762
76	697.2582	674.1265	682.8253	678.4759
75	687.9710	667.1247	675.3950	671.2599
74	680.6924	660.0639	668.2028	664.1334
73	672.8395	652.2745	661.0837	656.6791
72	665.6194	644.8733	653.3250	649.0992
71	658.1815	638.0097	645.7124	641.8611
70	649.7708	631.8787	638.7169	635.2978
69	642.3246	625.5679	632.4079	628.9879
68	635.2870	618.0866	626.1085	622.0976
67	629.2920	610.6547	618.6324	614.6436
66	622.1838	604.1462	611.0088	607.5775
65	613.6360	597.9631	604.4122	601.1877
64	606.7781	592.1326	598.1261	595.1293
63	600.0319	585.8224	592.2280	589.0252
62	594.0447	579.0276	585.8428	582.4352
61	587.6807	572.8717	578.9665	575.9191
60	580.4023	567.0207	572.7375	569.8791
59	573.8099	561.7002	566.8336	564.2669
58	567.6985	555.9431	561.4245	558.6838
57	562.1671	549.2384	555.5343	552.3864
56	556.1771	542.9823	548.7727	545.8825
55	549.1584	537.0428	542.4610	539.7519
54	542.6275	531.7301	536.4538	534.0920
53	536.4192	526.2432	531.1428	528.6930
52	530.9346	519.8786	525.5526	522.7156
51	525.1453	513.5102	518.9741	516.2421
50	518.2675	507.4842	512.5621	510.0231
49	511.6707	502.0592	505.5193	504.2893
48	505.4827	496.8017	501.1484	498.9750
47	500.0655	490.9074	495.7895	493.3535
46	494.5087	484.6111	489.6298	487.1204
45	487.8736	478.5947	483.2700	480.9324
44	481.3245	473.1156	477.2888	475.2022
43	475.3060	467.9507	471.8882	469.9194
42	469.9502	462.2953	466.6260	464.4607
41	464.4627	455.7916	460.5764	458.1840
40	457.6849	448.3464	453.9627	451.6545
39	450.8045	443.4144	447.5882	445.5013
38	444.4807	437.9065	441.7502	439.8254
37	438.7470	432.6777	435.0896	434.0837
36	432.7595	425.3554	429.8611	427.6083
35	425.7150	418.7739	423.1031	420.9385
34	418.8243	412.4304	416.5226	414.4765
33	412.1846	406.5104	410.2429	408.3767
32	406.0062	400.2831	404.1609	402.2235
31	399.4589	393.0181	397.5216	395.2698
30	391.6911	385.9670	390.1072	388.0371
29	384.3955	379.2527	383.1117	381.1822
28	377.4457	373.0158	376.4753	374.7455
27	371.0754	366.6421	370.2583	368.4502
26	364.2633	359.4762	363.5719	361.5241
25	356.3665	352.2551	355.9626	354.1089
24	349.0457	345.2573	348.8792	347.0733
23	341.7375	338.5396	341.8077	340.1737
22	334.9682	331.9003	335.2525	333.5764
21	327.7870	324.8468	328.3304	326.6386

Table 7 (cont.)

SAT-V PAPER-AND-PENCIL AND CAT
UNROUNDED RAW-TO-SCALE CONVERSIONS

Formula Score	Paper-and-Pencil Conversion	CAT Conversions		
		CAT Taken First	P-P Taken First	Average
20	320.3594	317.9054	321.1535	319.5295
19	313.1123	310.8431	314.1205	312.4820
18	305.3943	303.5691	306.7183	305.1437
17	298.0765	296.5616	299.5689	298.0653
16	290.6796	289.4860	292.4277	290.9568
15	283.2298	282.4110	285.2317	283.8213
14	276.1484	275.6530	278.2806	276.9668
13	269.1728	268.9785	271.4991	270.2388
12	261.8190	261.9526	264.5096	263.2311
11	254.2757	254.7492	257.2799	256.0146
10	246.9199	247.7097	250.0877	248.8987
9	239.6942	240.7823	243.0400	241.9161
8	232.4461	233.8708	236.0370	234.9539
7	224.7575	226.6117	228.8157	227.7137
6	217.1527	219.3250	221.4130	220.3694
5	209.4242	211.8774	214.0041	212.9907
4	202.0323	204.8039	206.6521	205.7280
3	194.0106	197.3762	199.2838	198.3300
2	186.2092	189.8303	191.5903	190.7103
1	179.6673	182.9955	184.3915	183.6935
0	172.2797	176.3676	177.8539	177.1108
-1	164.7299	169.2443	170.6710	169.9577
-2	156.9613	161.9530	163.3262	162.6396
-3	151.1258	155.1356	156.0959	155.6157
-4	143.7771	148.1543	150.2738	149.7141
-5	136.4287	142.1336	143.1633	142.6485
-6	129.0803	135.1129	136.0529	135.5829
-7	121.7319	128.0922	128.9425	128.5174
-8	114.3835	121.0715	121.8321	121.4518
-9	107.0351	114.0509	114.7217	114.3863
-10	99.6867	107.0302	107.6114	107.3208
-11	92.3383	100.0095	100.5010	100.2552
-12	84.9900	92.9888	93.3906	93.1897
-13	77.6416	85.9683	86.2802	86.1242
-14	70.2932	78.9476	79.1699	79.0587
-15	62.9448	71.9269	72.0595	71.9932
-16	55.5964	64.9062	64.9491	64.9277
-17	48.2480	57.8856	57.8387	57.8621
-18	40.8996	50.8649	50.7283	50.7966
-19	33.5512	43.8442	43.6179	43.7310
-20	26.2028	36.8235	36.5075	36.6655
-21	18.8544	29.8029	29.3971	29.6000

Table 8

SAT-M PAPER-AND-PENCIL AND CAT
UNROUNDED RAW-TO-SCALE CONVERSIONS

Formula Score	Paper-and-Pencil Conversion	CAT Conversions			
		CAT Taken First	P-P Taken First	1:1 Average	2:1 Average
60	790.9530				814.0579
59	781.3002				803.9611
58	773.3467	782.8907			793.8643
57	762.8606	774.3267			783.7675
56	753.1326	763.7168			773.6707
55	743.6692	753.5226	783.6765	768.5996	763.5739
54	734.1952	743.6553	774.7589	759.2071	754.0232
53	724.5774	733.7814	764.0029	748.8922	743.8553
52	714.7512	723.7463	753.5248	738.6356	733.6725
51	704.7181	713.4856	743.4011	728.4434	723.4575
50	694.4930	703.0034	733.2631	718.1332	713.0899
49	684.1221	692.3228	722.9508	707.6368	702.5322
48	673.6617	681.4985	712.4019	696.9502	691.7996
47	663.1666	670.5932	701.6222	686.1077	680.9362
46	652.6920	659.6688	690.6413	675.1551	669.9930
45	642.2720	648.7794	679.5194	664.1494	659.0261
44	631.9491	637.9668	668.3236	653.1452	648.0858
43	621.7254	627.2604	657.1202	642.1903	637.2137
42	611.6123	616.6672	645.9621	631.3146	626.4321
41	601.6140	606.1960	634.8964	620.5462	615.7628
40	591.7290	595.8482	623.9429	609.8956	605.2131
39	581.9527	585.6203	613.1119	599.3661	594.7842
38	572.2845	575.5098	602.4105	588.9601	584.4767
37	562.7122	565.5077	591.8381	578.6729	574.2845
36	553.2318	555.6065	581.3960	568.5012	564.2030
35	543.8397	545.8019	571.0765	558.4392	554.2268
34	534.5276	536.0861	560.8653	548.4757	544.3458
33	525.2904	526.4525	550.7576	538.6050	534.5542
32	516.1234	516.8957	540.7476	528.8216	524.8463
31	507.0256	507.4140	530.8265	519.1202	515.2182
30	497.9884	497.9986	520.9884	509.4935	505.6619
29	489.0166	488.6565	511.2296	499.9431	496.1809
28	480.1094	479.3837	501.5443	490.4640	486.7706
27	471.2582	470.1730	491.9311	481.0521	477.4257
26	462.4745	461.0353	482.3917	471.7135	468.1541
25	453.7578	451.9713	472.9188	462.4451	458.9538
24	445.1133	442.9859	463.5198	453.2528	449.8305
23	436.5426	434.0815	454.1970	444.1392	440.7866
22	428.0495	425.2627	444.9570	435.1099	431.8275
21	419.6389	416.5349	435.8049	426.1699	422.9582
20	411.3148	407.9013	426.7419	417.3216	414.1815
19	403.0786	399.3633	417.7735	408.5684	405.5000
18	394.9312	390.9213	408.9040	399.9126	396.9155
17	386.8722	382.5756	400.1347	391.3552	388.4287
16	378.9028	374.3237	391.4662	382.8950	380.0379
15	371.0172	366.1580	382.8989	374.5285	371.7383
14	363.2081	358.0698	374.4302	366.2500	363.5233

Table 8 (cont.)

SAT-M PAPER-AND-PENCIL AND CAT
UNROUNDED RAW-TO-SCALE CONVERSIONS

Formula Score	Paper-and-Pencil Conversion	CAT Conversions			
		CAT Taken First	P-P Taken First	1:1 Average	2:1 Average
13	355.4675	350.0485	366.0521	358.0503	355.3830
12	347.7849	342.0808	357.7554	349.9181	347.3057
11	340.1481	334.1508	349.5285	341.8397	339.2767
10	332.5420	326.2419	341.3573	333.7996	331.2804
9	324.9519	318.3369	333.2244	325.7807	323.2994
8	317.3624	310.4184	325.1121	317.7652	315.3163
7	309.7580	302.4668	317.0011	309.7339	307.3116
6	302.1210	294.4640	308.8711	301.6675	299.2664
5	294.4354	286.3922	300.7011	293.5466	291.1618
4	286.6882	278.2381	292.4725	285.3553	282.9829
3	278.8661	269.9882	284.1695	277.0789	274.7153
2	270.9567	261.6321	275.7766	268.7043	266.3469
1	262.9495	253.1669	267.2796	260.2232	257.8711
0	254.8428	244.5885	258.6704	251.6294	249.2824
-1	246.6305	235.9050	249.9444	242.9247	240.5848
-2	238.3170	227.1441	241.1013	234.1227	231.7965
-3	229.9184	218.3520	232.1549	225.2535	222.9530
-4	221.4794	209.6339	223.1476	216.3907	214.1385
-5	213.0339	201.7735	214.1239	207.9487	205.8903
-6	204.7787	193.3801	205.2777	199.3289	197.3459
-7	198.1509	183.7171	198.0721	190.8946	188.5021
-8	188.5125	174.4080	187.8261	181.1171	178.8807
-9	179.5749	165.0988	178.2752	171.6870	169.4910
-10	170.6372	155.7897	168.7242	162.2569	160.1012
-11	161.6994	146.4806	159.1731	152.8269	150.7114
-12	152.7617	137.1715	149.6222	143.3968	141.3217
-13	143.8241	127.8624	140.0713	133.9668	131.9320
-14	134.8864	118.5532	130.5203	124.5367	122.5422
-15	125.9487	109.2442	120.9693	115.1067	113.1525
-16	117.0110		111.4183		103.7628
-17	108.0734				94.3731

Table 9

COMPARISONS OF SAT-VERBAL PAPER-AND-PENCIL AND CAT CONVERSIONS

FORMULA SCORE	PAPER-AND-PENCIL SCALED SCORE		CAT SCALED SCORE		DIFFERENCE IN REPORTED SCORES[1]
	UNROUNDED	REPORTED	UNROUNDED	REPORTED	
85	765.7534 (795.1)	800	745.3607 (795.1)	800	0
84	760.3660 (775.1)	780	738.5434 (775.1)	780	0
83	749.7511 (755.1)	760	731.5802 (755.1)	760	0
82	742.5784 (745.1)	750	724.3010 (735.1)	740	10
81	735.7382	740	717.3147	720	20
80	728.1016	730	710.6981	71	20
79	720.4621	720	702.8536	700	20
78	713.5629	710	694.0555	690	20
77	706.8445	710	685.9762	690	20
76	697.2582	700	678.4759	680	20
75	687.9710	690	671.2599	670	20
74	680.6924	680	664.1334	660	20
73	672.8395	670	656.6791	660	10
72	665.6194	670	649.0992	650	20
71	658.1815	660	641.8611	640	20
70	649.7708	650	635.2978	640	10
69	642.3246	640	628.9879	630	10
68	635.2870	640	622.0976	620	20
67	629.2920	630	614.6436	610	20
66	622.1838	620	607.5775	610	10
65	613.6360	610	601.1877	600	10
64	606.7781	610	595.1293	600	10
63	600.0319	600	589.0252	590	10
62	594.0447	590	582.4352	580	10
61	587.6907	590	575.9191	580	10
60	580.4023	580	569.8791	570	10
59	573.9099	570	564.2669	560	10
58	567.6985	570	558.6838	560	10
57	562.1671	560	552.3864	550	10
56	556.1771	560	545.8825	550	10
55	549.1594	550	539.7519	540	10
54	542.6275	540	534.0920	530	10
53	536.4192	540	528.6930	530	10
52	530.9346	530	522.7156	520	10
51	525.1453	530	516.2421	520	10
50	518.2675	520	510.0231	510	10
49	511.6707	510	504.2893	500	10
48	505.4827	510	498.9750	500	10
47	500.0655	500	493.3535	490	10
46	494.5087	490	487.1204	490	0
45	487.8736	490	480.9324	480	10
44	481.3245	480	475.2022	480	0
43	475.3060	480	469.9194	470	10
42	469.9502	470	464.4607	460	10
41	464.4627	460	458.1840	460	0
40	457.6849	460	451.6545	450	10
39	450.8045	450	445.5013	450	0
38	444.4807	440	439.8254	440	0
37	438.7470	440	434.0837	430	10
36	432.7595	430	427.6083	430	0
35	425.7150	430	420.9385	420	10
34	418.8243	420	414.4765	410	10
33	412.1846	410	408.3767	410	0
32	406.0062	410	402.2235	400	10
31	399.4599	400	395.2698	400	0
30	391.6911	390	388.0371	390	0
29	384.3955	380	381.1822	380	0
28	377.4457	380	374.7455	370	10
27	371.0754	370	368.4502	370	0
26	364.2633	360	361.5241	360	0
25	356.3665	360	354.1089	350	10
24	349.0457	350	347.0733	350	0
23	341.7375	340	340.1737	340	0
22	334.9682	330	333.5764	330	0
21	327.7970	330	326.6386	330	0

[1]Paper-and-pencil reported - CAT reported

Table 9 (cont.)

COMPARISONS OF SAT-VERBAL
PAPER-AND-PENCIL AND CAT CONVERSIONS

FORMULA SCORE	PAPER-AND-PENCIL SCALED SCORE		CAT SCALED SCORE		DIFFERENCE IN REPORTED SCORES[1]
	UNROUNDED	REPORTED	UNROUNDED	REPORTED	
20	320.3594	320	319.5295	320	0
19	313.1123	310	312.4820	310	0
18	305.3943	310	305.1437	310	0
17	298.0766	300	298.0653	300	0
16	290.6796	290	290.9568	290	0
15	283.2298	280	283.8213	280	0
14	276.1484	280	276.9668	280	0
13	269.1728	270	270.2388	270	0
12	261.8190	260	263.2311	260	0
11	254.2757	250	256.0146	260	-10
10	246.9199	250	248.8987	250	0
9	239.6942	240	241.9161	240	0
8	232.4461	230	234.9539	230	0
7	224.7575	220	227.7137	230	-10
6	217.1527	220	220.3694	220	0
5	209.4242	210	212.9907	210	0
4	202.0323	200	205.7280	210	-10
3	194.0106	200	198.3300	200	0
2	186.2092	200	190.7103	200	0
1	179.6673	200	183.6935	200	0
0	172.2797	200	177.1108	200	0
-1	164.7299	200	169.9577	200	0
-2	156.9613	200	162.6396	200	0
-3	151.1256	200	155.6157	200	0
-4	143.7771	200	149.7141	200	0
-5	136.4287	200	142.6485	200	0
-6	129.0803	200	135.5829	200	0
-7	121.7319	200	128.5174	200	0
-8	114.3835	200	121.4518	200	0
-9	107.0351	200	114.3863	200	0
-10	99.6867	200	107.3208	200	0
-11	92.3383	200	100.2552	200	0
-12	84.9900	200	93.1897	200	0
-13	77.6416	200	86.1242	200	0
-14	70.2932	200	79.0587	200	0
-15	62.9448	200	71.9932	200	0
-16	55.5964	200	64.9277	200	0
-17	48.2480	200	57.8621	200	0
-18	40.8996	200	50.7966	200	0
-19	33.5512	200	43.7310	200	0
-20	26.2028	200	36.6655	200	0
-21	18.8544	200	29.6000	200	0

[1]Paper-and-pencil reported - CAT reported

Table 10

COMPARISONS OF SAT-MATH
PAPER-AND-PENCIL AND CAT CONVERSIONS

	PAPER-AND-PENCIL SCALED SCORE		CAT SCALED SCORE		
FORMULA SCORE	UNROUNDED	REPORTED	UNROUNDED	REPORTED	DIFFERENCE IN REPORTED SCORES[1]
60	790.9530 (795.1)	800	814.0579	800	0
59	781.3002	780	803.9611	800	-20
58	773.3467	770	793.8643	790	-20
57	762.8606	760	783.7675	780	-20
56	753.1326	750	773.6707	770	-20
55	743.6692	740	763.5739	760	-20
54	734.1952	730	754.0232	750	-20
53	724.5774	720	743.8553	740	-20
52	714.7512	710	733.6725	730	-20
51	704.7181	700	723.4575	720	-20
50	694.4930	690	713.0899	710	-20
49	684.1221	680	702.5322	700	-20
48	673.6617	670	691.7996	690	-20
47	663.1666	660	680.9362	680	-20
46	652.6920	650	669.9930	670	-20
45	642.2720	640	659.0261	660	-20
44	631.9491	630	648.0858	650	-20
43	621.7254	620	637.2137	640	-20
42	611.6123	610	626.4321	630	-20
41	601.6140	600	615.7628	620	-20
40	591.7290	590	605.2131	610	-20
39	581.9527	580	594.7842	590	-10
38	572.2845	570	584.4767	580	-10
37	562.7122	560	574.2845	570	-10
36	553.2318	550	564.2030	560	-10
35	543.8397	540	554.2268	550	-10
34	534.5276	530	544.3458	540	-10
33	525.2904	530	534.5542	530	0
32	516.1234	520	524.8463	520	0
31	507.0256	510	515.2182	520	-10
30	497.9884	500	505.6619	510	-10
29	489.0166	490	496.1809	500	-10
28	480.1094	480	486.7706	490	-10
27	471.2582	470	477.4257	480	-10
26	462.4745	460	468.1541	470	-10
25	453.7578	450	458.9538	460	-10
24	445.1133	450	449.8305	450	0
23	436.5426	440	440.7866	440	0
22	428.0495	430	431.8275	430	0
21	419.6389	420	422.9582	420	0
20	411.3148	410	414.1815	410	0
19	403.0786	400	405.5000	410	-10
18	394.9312	390	396.9155	400	-10
17	386.8722	390	388.4287	390	0
16	378.9028	380	380.0379	380	0
15	371.0172	370	371.7383	370	0
14	363.2081	360	363.5233	360	0
13	355.4675	360	355.3830	360	0
12	347.7849	350	347.3057	350	0
11	340.1481	340	339.2767	340	0
10	332.5420	330	331.2804	330	0
9	324.9519	320	323.2994	320	0
8	317.3624	320	315.3163	320	0
7	309.7580	310	307.3116	310	0
6	302.1210	300	299.2664	300	0
5	294.4354	290	291.1618	290	0
4	286.6882	290	282.9829	280	10
3	278.8661	280	274.7153	270	10
2	270.9567	270	266.3469	270	0
1	262.9495	260	257.8711	260	0
0	254.8428	250	249.2824	250	0
-1	246.6305	250	240.5848	240	10
-2	238.3170	240	231.7965	230	10
-3	229.9184	230	222.9530	220	10
-4	221.4794	220	214.1385	210	10
-5	213.0339	210	205.8903	210	0

[1]Paper-and-pencil reported - CAT reported

Table 10 (cont.)

COMPARISONS OF SAT-MATH
PAPER-AND-PENCIL AND CAT CONVERSIONS

	PAPER-AND-PENCIL SCALED SCORE		CAT SCALED SCORE		
FORMULA SCORE	UNROUNDED	REPORTED	UNROUNDED	REPORTED	DIFFERENCE IN REPORTED SCORES[1]
-6	204.7787	200	197.3459	200	0
-7	198.1509	200	188.5021	200	0
-8	188.5125	200	178.8807	200	0
-9	179.5749	200	169.4910	200	0
-10	170.6372	200	160.1012	200	0
-11	161.6994	200	150.7114	200	0
-12	152.7617	200	141.3217	200	0
-13	143.8241	200	131.9320	200	0
-14	134.8864	200	122.5422	200	0
-15	125.9487	200	113.1525	200	0
-16	117.0110	200	103.7628	200	0
-17	108.0734	200	94.3731	200	0

[1]Paper-and-pencil reported - CAT reported

IATP Computer Adaptive SAT Pilot

THE RESEARCH DESIGN

The study is designed to establish comparable reported score scales for the paper-and-pencil and the computerized adaptive versions of the IATP SAT. It is very important that the study be conducted according to one of the designs outlined below and that students complete both tests within two (2) weeks. We ask that you choose one of the two designs and then test all students using that design.

Design I: Counterbalanced Design

This design requires that half of the students testing on a given day take the paper-and-pencil version first, while the second half testing on that day take the computerized version first. The students testing on a specific day should be divided into two groups of equal size in a random fashion. (We will supply you with specific procedures for splitting the total group testing on a specific day into subgroups in a random fashion at a later date. This is critical to the success of the study.) Both tests would be administered on the same day with possibly a lunch break in between.

Design II: Modified Counterbalanced Design

This design has the advantage of allowing you to test all students at the same time with the paper-and-pencil test. Test center personnel, with the knowledge of who will be tested beforehand, should randomly split the total group to be tested into two equal sized groups, Group A and Group B. (We will supply you with specific procedures for splitting your total group into subgroups in a random fashion at a later date. This is critical to the success of the study.)

Group A students will be tested with the computer version of the IATP SAT for as many days as needed to complete the computer based testing. However, computer testing may not occur more than two weeks prior to the paper-and-pencil test administration. Group A and B will then be brought together to take the paper-and-pencil test.

At the conclusion of the paper-and-pencil test, Group A completes the questionnaire about their experiences with the computerized test. Group B participants may then begin computer based testing after completing the paper-and-pencil test. Testing should continue for as many days as are necessary to test the entire group (but no longer than two weeks after the paper-and-pencil test). Students in Group B complete the questionnaire immediately after taking the computer based test.

Figure 1: Designs for conducting the SAT CAT pilot/comparability study .

IATP Computer Adaptive SAT Pilot

Research Design/Random Assignment Guidelines

PROCEDURES FOR SPLITTING TOTAL GROUP TO BE TESTED INTO SUBGROUPS

DESIGN I: COUNTERBALANCED DESIGN

This design requires that half of the examinees testing on a given day take the paper-and-pencil version first, while the second half testing on that day take the computerized version first. The examinees testing on a specific day should be divided into two groups of equal size in a random fashion.

Procedure

TWO COMPUTERIZED SESSIONS

Condition A: If you are planning on running only two sessions of computerized testing on a given day (one session before the paper-and-pencil test, one session after; ALL EXAMINEES TAKE THE PAPER-AND-PENCIL TEST TOGETHER) and:

NO ROSTER

A1. You are allowing examinees to choose the day they want to test, possibly by phone (i.e., you do not have beforehand an intact roster of examinees to be tested on a given day), then as you are contacted by the examinees, alternate assignment to testing orders. Assign the first examinee who contacts you to the computer then paper-and-pencil order, the second examinee who contacts you to the paper-and-pencil then computer order, the third examinee to the computer then paper-and-pencil order, etc., until all slots are filled. If you are using five computers, this means that your total group for that given day will consist of 10 examinees, with 5 receiving the computer then paper-and-pencil order and 5 the paper-and-pencil then computer order. All examinees testing on the given day should complete the questionnaire after they have taken both tests.

ROSTER

A2. You have beforehand an intact roster of examinees to be tested on a given day (this may be the case if you are testing local high school students). Alphabetize the roster and assign the first examinee listed in the alphabetized roster to the computer then paper-and-pencil order, the second listed to the paper-and-pencil then computer order, the third listed to the computer then paper-and-pencil order, etc. If you are using five computers, this means your total group for that given day will consist of 10 examinees, with

Figure 2: Procedures for splitting SAT CAT total groups of examinees into randomly equivalent subgroups.

5 receiving the computer then paper-and-pencil order and 5 receiving the paper-and-pencil then computer order. The 5 receiving the computer then paper-and-pencil order will be in positions 1, 3, 5, 7, and 9 on your roster for that day while the 5 examinees receiving the paper-and-pencil then computer order will be in positions 2, 4, 6, 8, and 10 on your roster. All examinees testing on the given day should complete the questionnaire after they have taken both tests.

MULTIPLE COMPUTERIZED SESSIONS

Condition B: If you are planning on running multiple sessions of computerized testing on a given day, then you must schedule the paper-and-pencil testing in the middle of the day (ALL EXAMINEES TAKE THE PAPER-AND-PENCIL TEST TOGETHER) and an equal number of computerized testing sessions before and after the paper-and-pencil testing. If:

NO ROSTER

B1. You are allowing the examinees to choose the day they want to test, possibly by phone (i.e., you do not have beforehand an intact roster of examinees to be tested on a given day), then as you are contacted by examinees, alternate assignments to testing orders. That is, assign the first examinee who contacts you to the computer then paper-and-pencil order. This examinee is free to choose which of the sessions of computerized testing before the paper-and-pencil testing on that day he/she wants to attend. Assign the second examinee who contacts you to the paper-and-pencil then computer order. This examinee is also free to choose which of the sessions of computerized testing after the paper-and-pencil testing on that day he/she wants to attend. The third examinee who contacts you would be assigned to the computer then paper-and-pencil order, etc. This examinee and later examinees are free to choose which of the appropriate sessions of computerized testing before or after paper-and-pencil testing on that day they want, provided that slots are open. Examinees who contact you later in the process will have to be assigned to a session of computerized testing.

After you have completed assigning all examinees to test orders and computerized testing sessions, you should make sure there are an equal number of examinees assigned to sessions of computerized testing before paper-and-pencil testing as there are examinees assigned to sessions of computerized testing after paper-and-pencil testing. If there are not, assign the last examinee who contacted you to

Figure 2 (cont.): Procedures for splitting SAT CAT total groups of examinees into randomly equivalent subgroups.

another day of testing.

All examinees testing on the given day should complete the questionnaire after they have taken both tests.

ROSTER B2. You have beforehand an intact roster of examinees testing on a given day (this may be the case if you are bringing examinees to your institution to be tested on a specific day). Alphabetize the roster and assign the first examinee listed in the alphabetical roster to the computer then paper-and-pencil order. Assign the second examinee listed in the alphabetical roster to the paper-and-pencil then computer order and the third examinee to the computer then paper-and-pencil order, etc. As you assign examinees to sessions of computerized testing, it would be a good idea (but it isn't necessary) to fill the sessions closest to the paper-and-pencil testing first, to minimize the number of examinees who will have a waiting period between testing sessions.

After you have completed assigning all examinees to testing orders and computerized testing sessions, you should make sure there are an equal number of examinees assigned to sessions of computerized testing before paper-and-pencil testing as there are examinees assigned to sessions of computerized testing after paper-and-pencil testing. (If the totals differ by one, it means that you had an odd number of examinees on your roster. This is okay for testing purposes (i.e., go ahead and test everyone), but we will be unable to use the data from the last examinee assigned in the comparability study. Please record the name of this examinee and provide it to us. We will provide scores for that examinee.)

All examinees testing on the given day should complete the questionnaire after they have taken both tests.

DESIGN II: MODIFIED COUNTERBALANCED DESIGN

This design has the advantage of allowing you to test all examinees at the same time with the paper-and-pencil test. With knowledge of the total group to be tested beforehand, this total group should be randomly split into two equally sized

Figure 2 (cont.): Procedures for splitting SAT CAT total groups of examinees into randomly equivalent subgroups.

groups, Group A and Group B. Group A examinees will take the computerized test before the paper-and-pencil test while Group B examinees will take the paper-and-pencil test before the computerized test.

Procedure

The paper-and-pencil testing session needs to be scheduled in the middle of your testing period, so you have an equal number of days before and after this session for computerized testing. You may run one or multiple sessions of computerized testing on those days.

If the total roster of examinees to be tested is not alphabetized, then alphabetize it. The first examinee on the alphabetized roster should be assigned to Group A, the second examinee to Group B, the third examinee to Group A, etc. You should end up with an equal number of examinees in Groups A and B. If you do not, and are off by one examinee (i.e., your total group roster had an odd number of examinees), go ahead and test everyone, but keep a record of the name of the last examinee assigned and provide it to us. We will not be able to use the data from that examinee in the comparability study, but we will provide scores for that examinee.

After you have split the total group into Groups A and B, you may assign or allow examinees to select the sessions when they take the computerized test. All Group A examinees must, however, take the computerized test before they take the paper-and-pencil test. All Group B examinees must take the computerized test after they take paper-and-pencil test. Group A examinees should complete the questionnaire immediately following the paper-and-pencil test. Group B examinees should complete the questionnaire immediately following the computerized test.

Figure 2 (cont.): Procedures for splitting SAT CAT total groups of examinees into randomly equivalent subgroups.

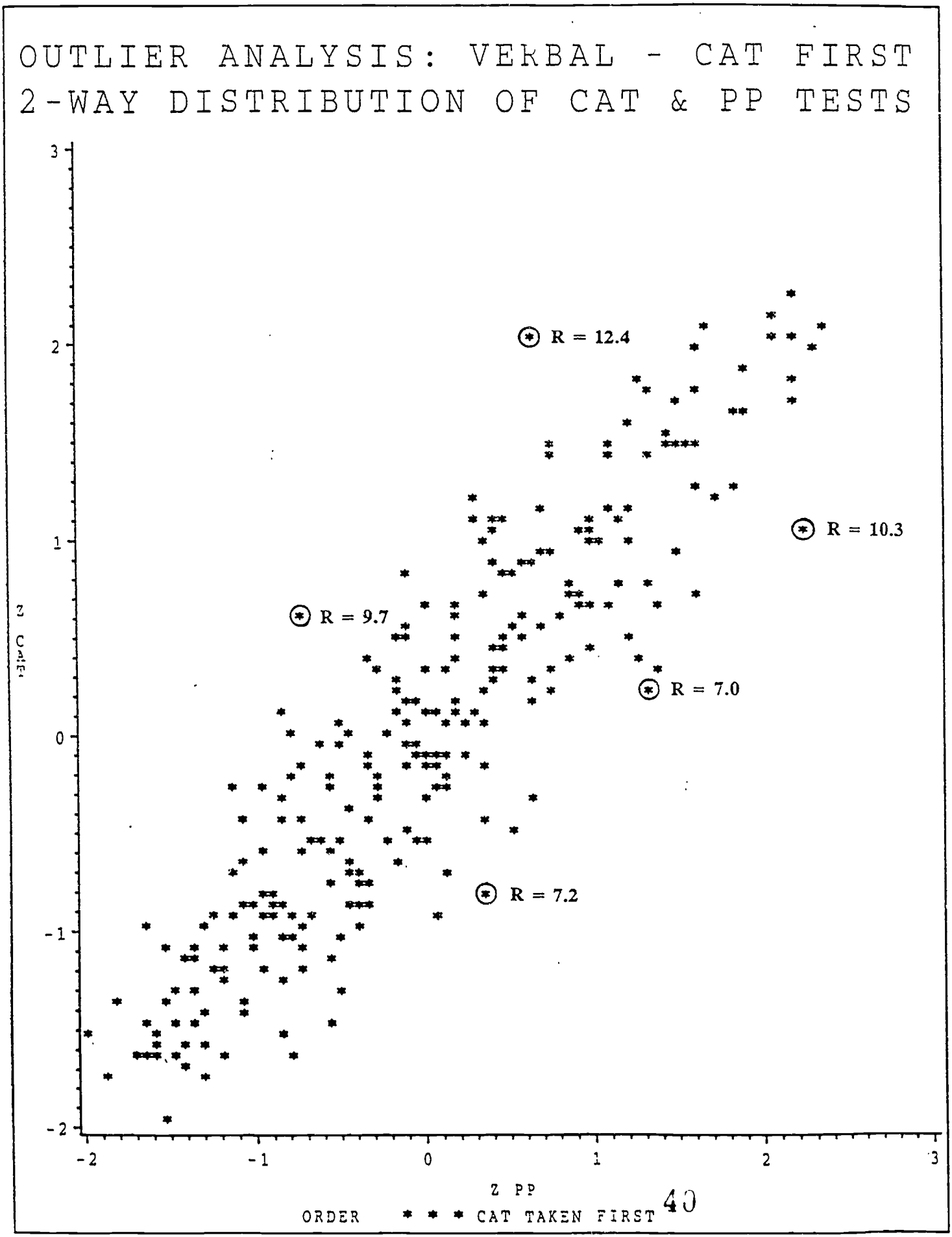

OUTLIER ANALYSIS: VERBAL - CAT FIRST
2-WAY DISTRIBUTION OF CAT & PP TESTS
R = 12.4
R = 10.3
R = 9.7
R = 7.0
R = 7.2
Z CAT
Z PP
ORDER * * * CAT TAKEN FIRST

Figure 3

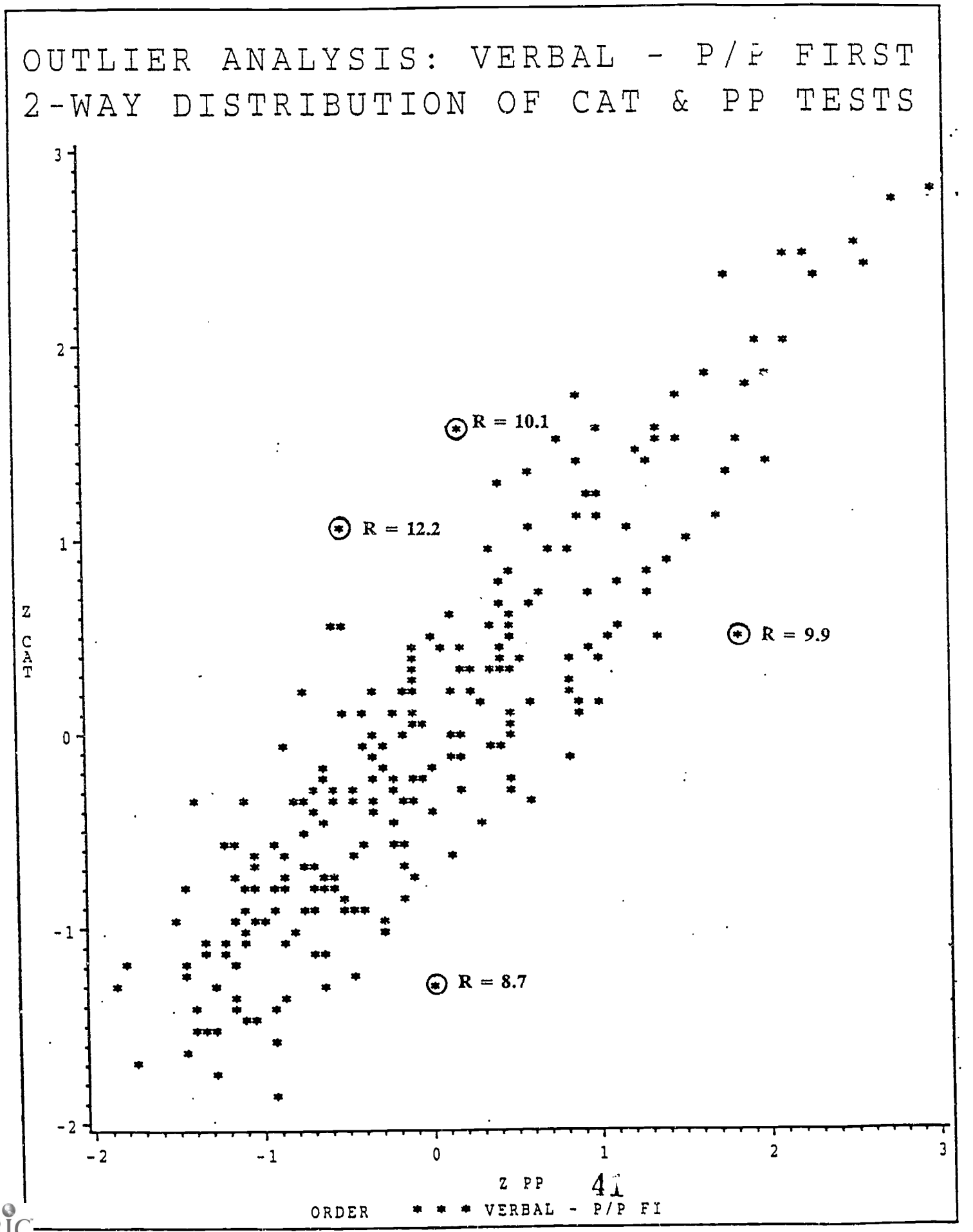

OUTLIER ANALYSIS: VERBAL - P/P FIRST
2-WAY DISTRIBUTION OF CAT & PP TESTS
R = 10.1
R = 12.2
R = 9.9
R = 8.7
Z CAT
Z PP
ORDER * * * VERBAL - P/P FI

Figure 4

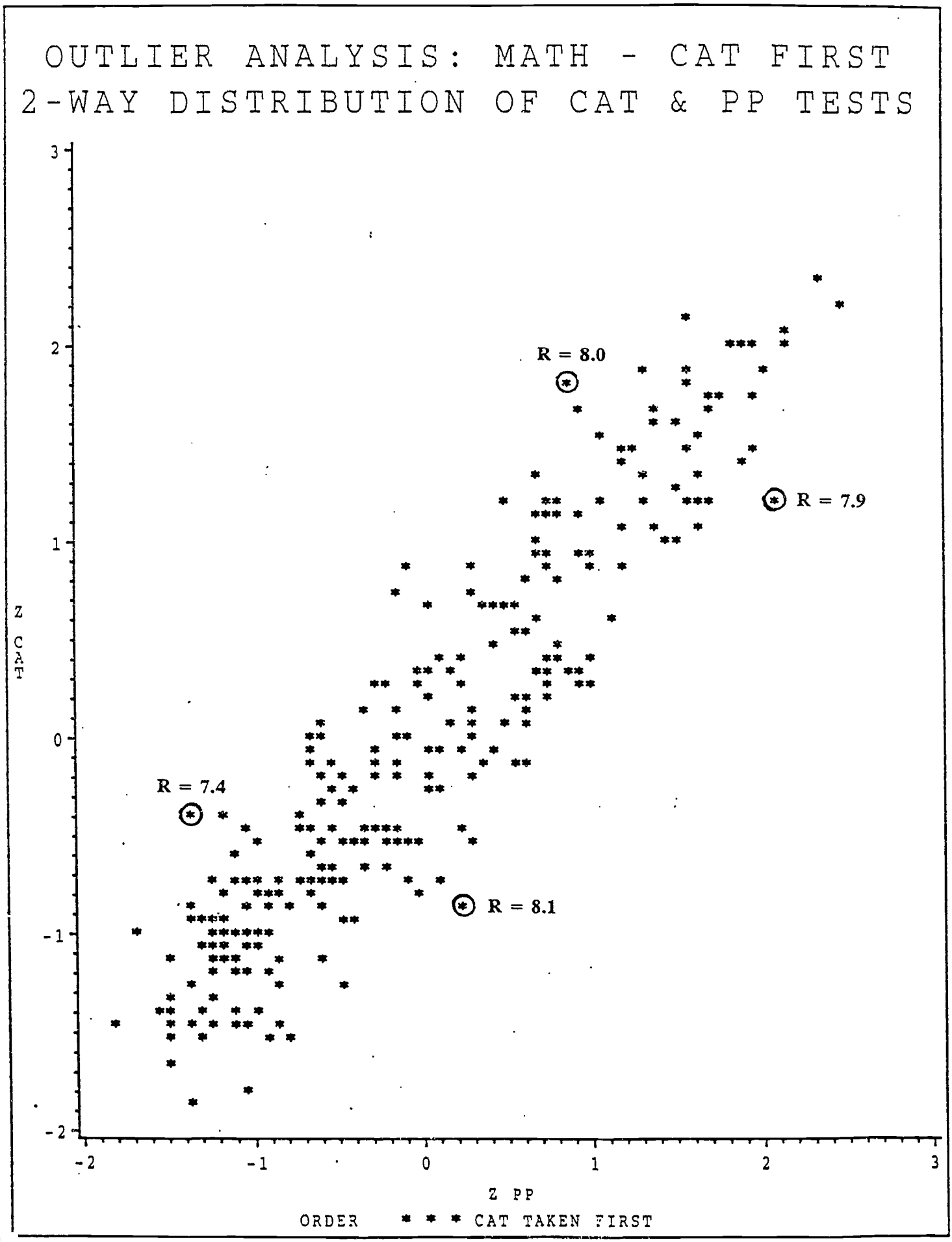

OUTLIER ANALYSIS: MATH - CAT FIRST
2-WAY DISTRIBUTION OF CAT & PP TESTS
R = 8.0
R = 7.9
R = 7.4
R = 8.1
Z CAT
Z PP
3
2
1
0
-1
-2
-2
-1
0
1
2
3
ORDER * * * CAT TAKEN FIRST

Figure 5

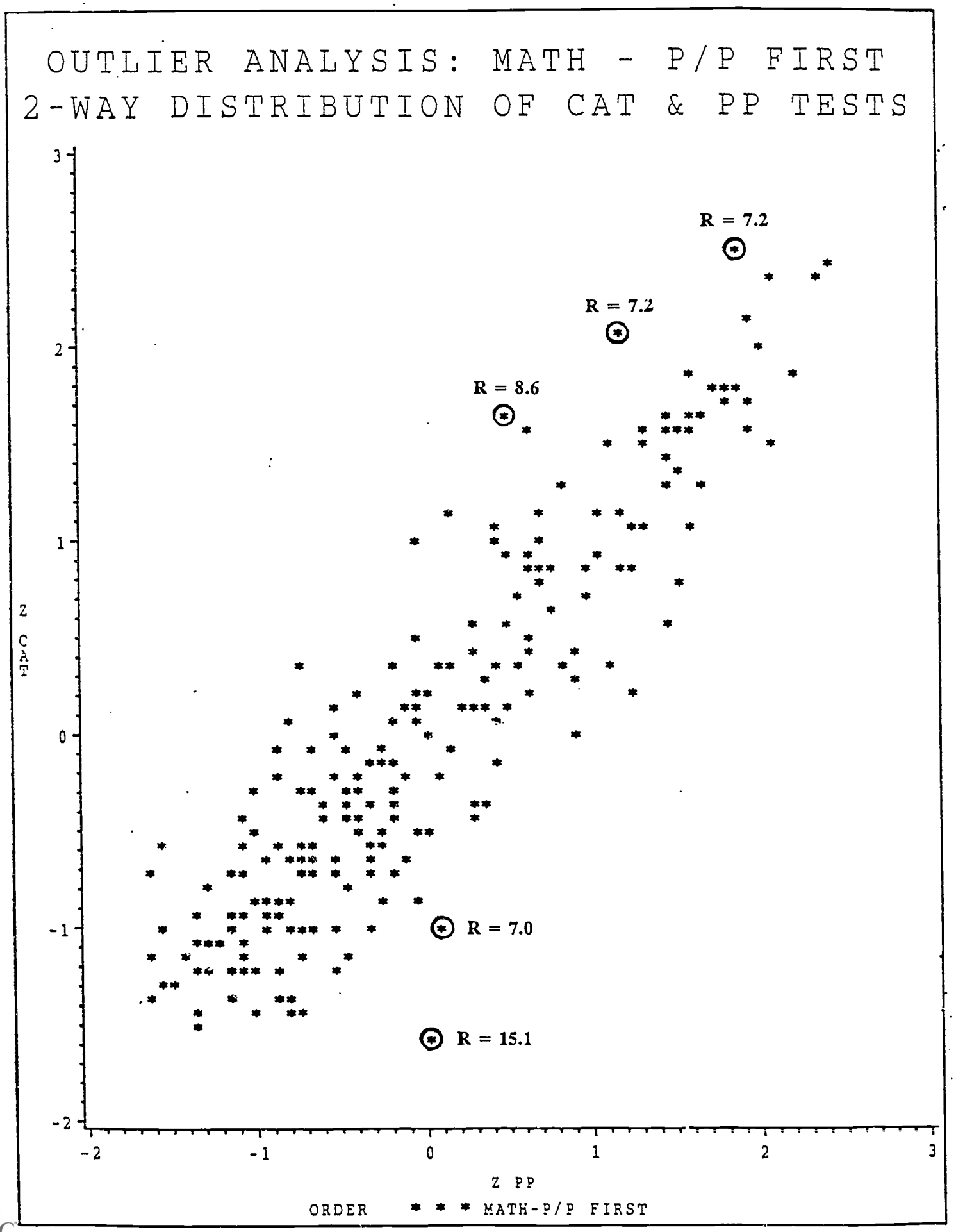
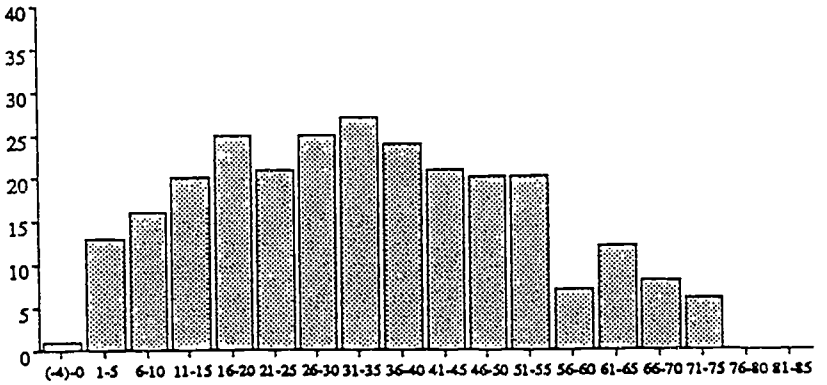
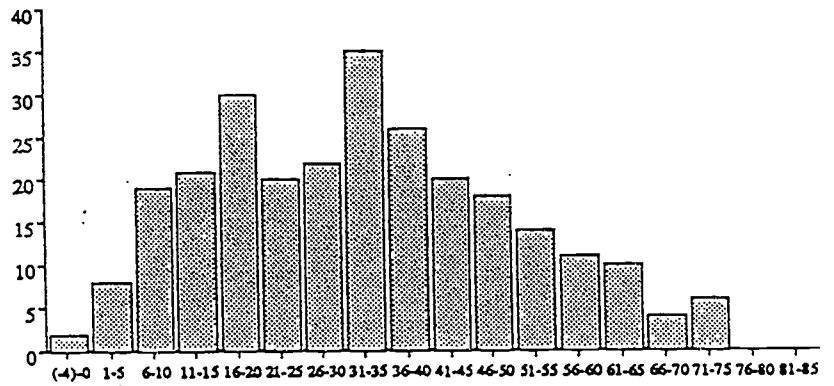

OUTLIER ANALYSIS: MATH - P/P FIRST
2-WAY DISTRIBUTION OF CAT & PP TESTS
R = 7.2
R = 7.2
R = 8.6
R = 7.0
R = 15.1
Z CAT
Z PP
ORDER * * * MATH-P/P FIRST

Figure 6

CAT Taken First

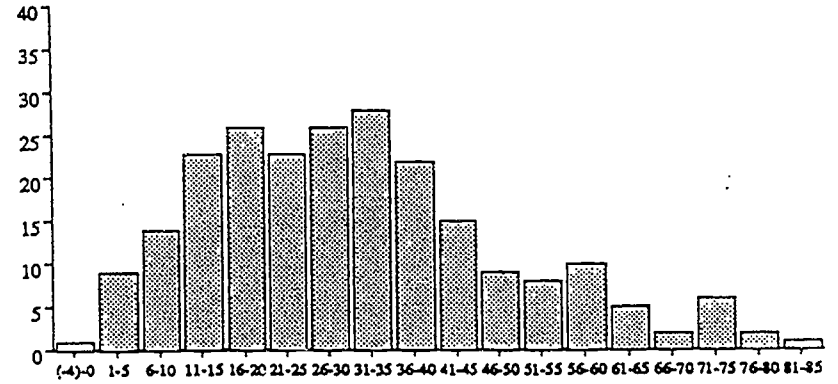


CAT

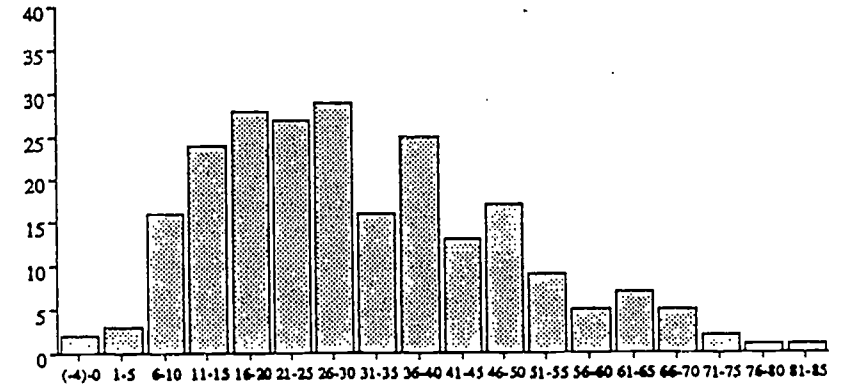


Paper-and-Pencil

Paper-and-Pencil Taken First



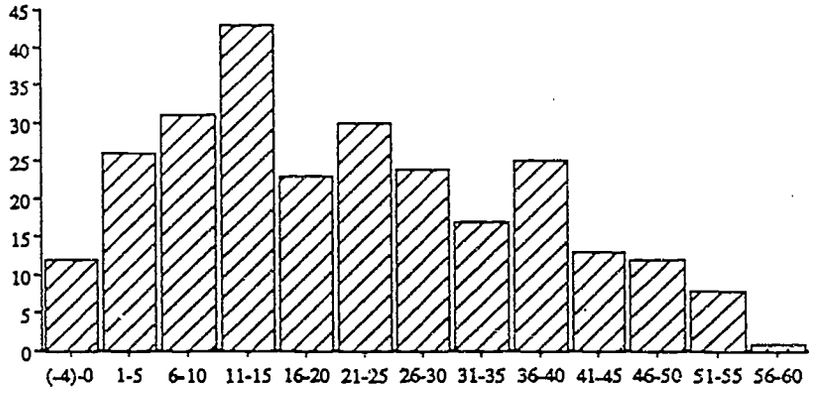
CAT

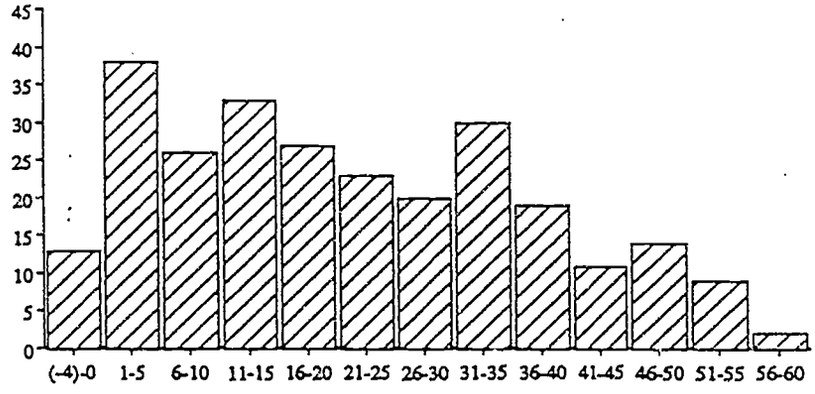


Paper-and-Pencil

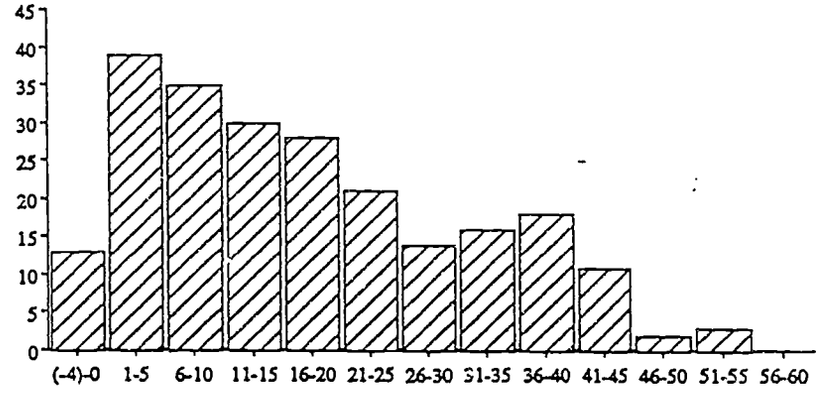
Figure 7: Grouped frequency distributions of SAT-V formula scores.

CAT Taken First





Paper-and-Pencil Taken First



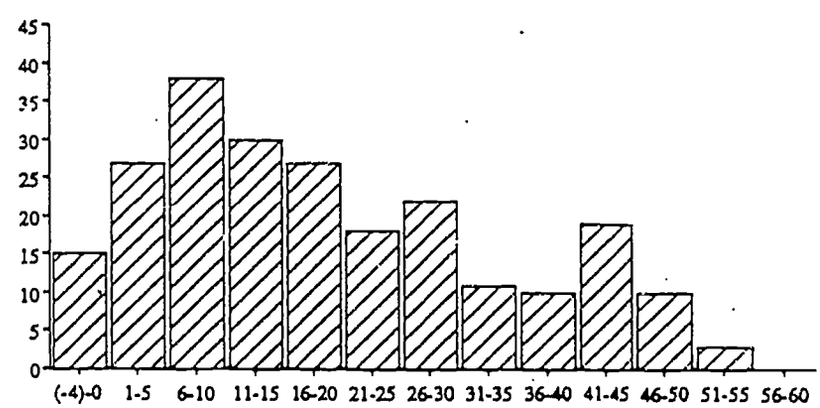
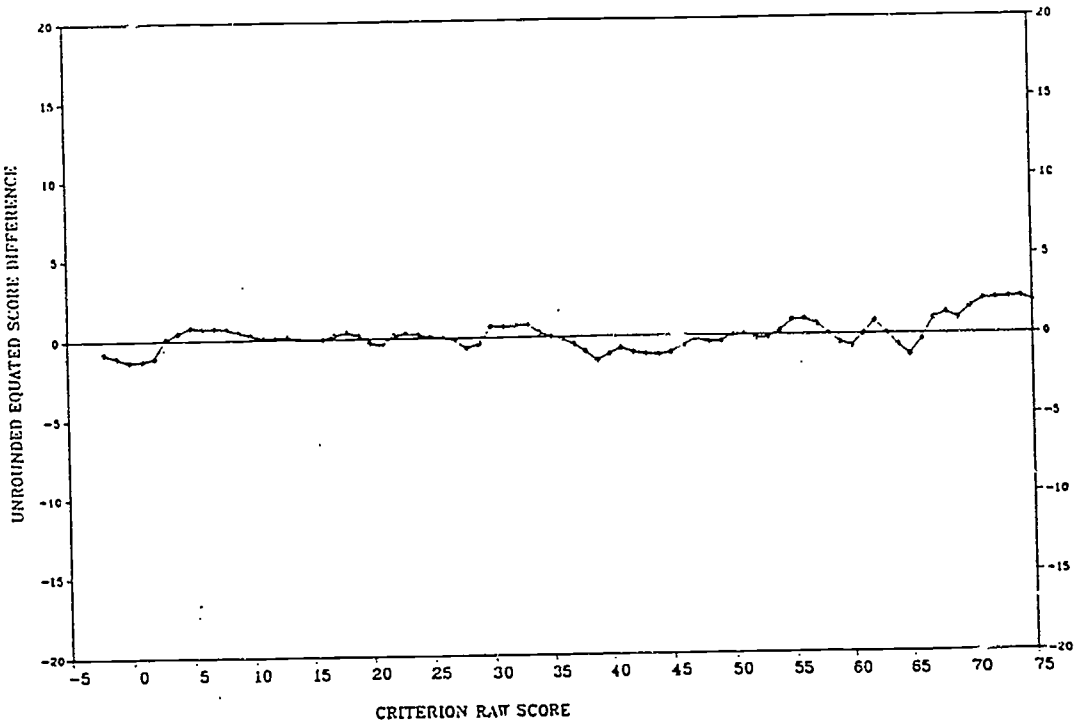


Figure 8: Grouped frequency distributions of SAT-M formula scores.

CAT Taken First



Paper-and-Pencil Taken First

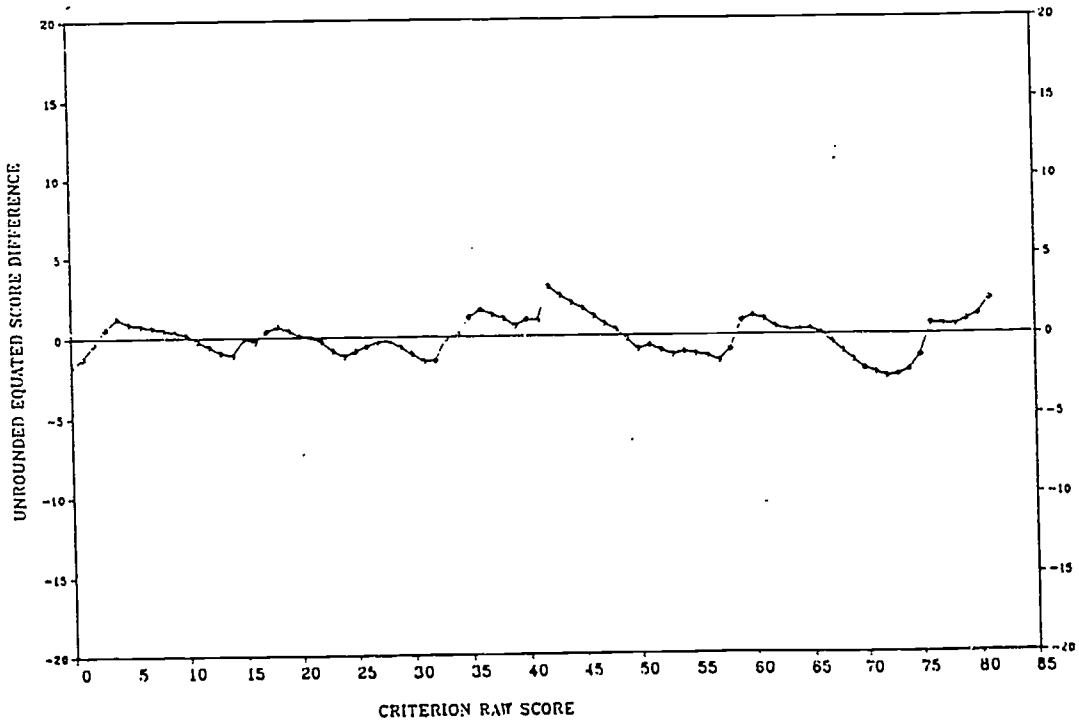
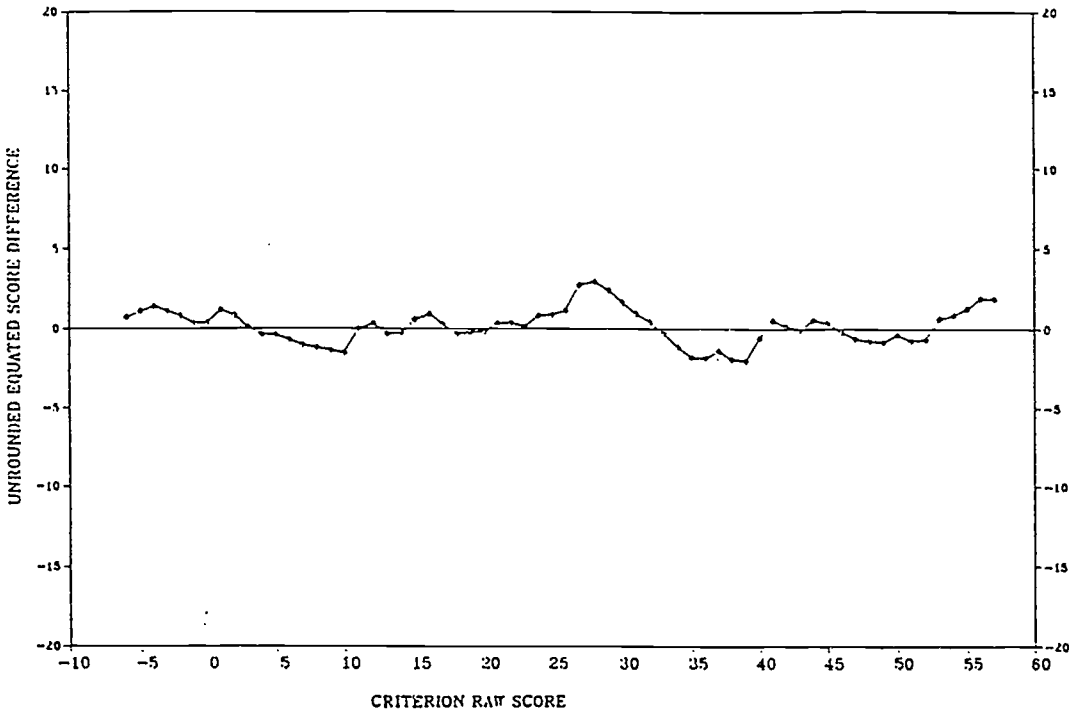


Figure 9: Raw-to-raw SAT-V difference plots, constructed using the linear conversion as the criterion.

CAT Taken First



Paper-and-Pencil Taken First

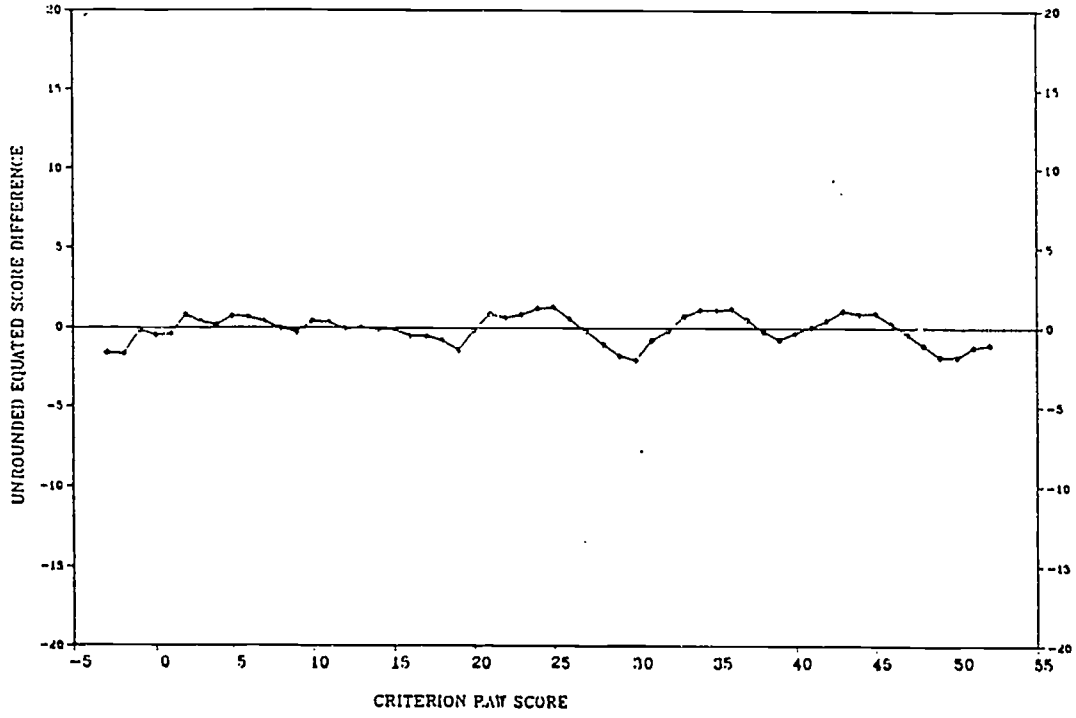


Figure 10: Raw-to-raw SAT-M difference plots, constructed using the linear conversion as the criterion.