Stability of Item Calibrations for the NCLEX-RN™ and NCLEX-PN™ using Computerized Adaptive Testing[1]

Kathleen A. Haynie
Walter D. Way

The Chauncey Group International
March 1996

---

[1] This paper will be presented at a symposium entitled "Supporting Ongoing Computerized Adaptive Testing: Psychometric Research on the NCLEX", at the annual meeting of AERA, New York, 1996.

Abstract

In maintaining item pools for computerized adaptive testing (CAT) for the NCLEX-RN and NCLEX-PN examinations, an important technical issue concerns the recalibration of Rasch item difficulties using data based on operational CAT administrations. Previous research related to the NCLEX has examined recalibrations of both simulated and real CAT data (Haynie & Way, in preparation). Although this research suggests that CAT-based item recalibration may be feasible for the NCLEX, it was limited to single calibration samples and did not examine the stability of CAT-based item recalibrations over time. In considering the long-term effects of CAT item recalibration, it is important to examine the stability of item recalibrations over time and different configurations of item pools. The purpose of this study was to undertake such an investigation.

Comparisons of recalibrated item difficulties based on different pairs of pretest and CAT pools indicated very high correlations ($r = .97$ to .99). Correlations between CAT or pretest estimates and paper-and-pencil estimates were somewhat lower, ranging from .92 to .96. The associations, however, were higher when paper-and-pencil estimates were restricted to items calibrated after 1990. These results suggest that item recalibration in the context of CAT may be supported, although further investigation of scaling issues is needed.

Stability of Item Calibrations for the NCLEX-RN and NCLEX-PN using Computerized Adaptive Testing

Kathleen A. Haynie
Walter D. Way

Introduction

In April of 1994, the National Council of State Boards of Nursing (NCSBN) implemented computerized adaptive testing (CAT) for licensing both registered and practical nurses. These tests, known as the NCLEX-RN and NCLEX-PN are variable-length computerized adaptive tests in which item selection and item and examinee estimation are based on a one-parameter logistic, or Rasch, item response theory (IRT) model. Currently, there are several thousand NCLEX-RN and NCLEX-PN test items available for use in CAT. These items have been divided into several parallel item pools that are used operationally.

The item pools for the NCLEX-RN and NCLEX-PN consist of multiple-choice items that have been calibrated using the Rasch model. The estimated item difficulties for these items have been based on either: 1) calibrations of paper-and-pencil test data prior to the implementation of CAT; or 2) calibrations of data based on unscored items that were administered nonadaptively as part of the NCLEX delivered via computerized adaptive testing. In maintaining the NCLEX item pools, it is desirable to periodically recalibrate the estimated item difficulties that have been administered via CAT. However, a variety of research has suggested that the calibrations of adaptively-administered item response data raises issues that require careful study (Divgi, 1986; Stocking, 1988; Bergstrom & Lunz, 1991; Mislevy, Wingersky & Kingston, 1990; Holland & Wingersky, 1991; Ito and Sykes, 1994).

A previous research study provided an initial investigation of item calibration procedures for the NCLEX delivered using computerized adaptive testing (Haynie & Way, 1995; Haynie & Way, in preparation). This study evaluated both real and simulated data for items administered nonadaptively and adaptively. The principal results of the study can be summarized as follows:

1. Rasch calibrations of simulated data where items were administered nonadaptively (i.e., tryout items) resulted in estimates of item difficulty that were highly correlated with true item difficulty ($r = 0.99+$) and with previous estimates of item difficulty ($r = 0.99+$).

2. Rasch calibrations of real data where items were administered nonadaptively resulted in estimates of item difficulty that were less highly correlated with previous estimates of item difficulty ($r = 0.94$); however, the relationships between the two sets of item difficulty estimates were linear.

3. Rasch calibrations of simulated CAT data suggested curvilinear relationships between estimates of item difficulty and true item difficulty, although the estimates were highly correlated to the true values ($r = 0.99$). Similar curvilinear relationships also existed between CAT-based estimates of item difficulty and previous estimates of item difficulty ($r = 0.99$).

4. Rasch calibrations of real CAT data also suggested curvilinear relationships between CAT-based estimates of item difficulty and previous estimates of item difficulty. The correlation between

CAT-based estimates of item difficulty and previous estimates of item difficulty was 0.92.

5. CAT simulations designed to further investigate the curvilinear relationship between the CAT-based and original item difficulty estimates suggested that this relationship may have been due to a combination of small sample sizes and poor targeting of items to candidate abilities. That is, items at extreme difficulty levels were administered less often and to candidate abilities that were either higher on average (for easier items) or lower on average (for difficult items) than would be optimal.

Based on this previous study, the calibrations of real CAT data did not recover the previous estimates of item difficulty as well as the calibrations of simulated CAT data; however, the results appeared to be acceptable enough for the NCLEX program to consider on-line item recalibration based on CAT data. The study was limited in that it investigated item calibration results for items based on a single item pool and did not consider the stability of on-line recalibrations over time.

At present, nearly two years worth of CAT data have been collected on the NCLEX. These data include items that have been utilized in a variety of formats. For some subsets of items that were originally calibrated in the paper-and-pencil examination, both adaptive and non-adaptive data are available for analysis. Some of these items were administered as pretest pool anchor items, or items with existing difficulty estimates, in 1994 and as operational CAT items in 1995. Other items were administered as operational CAT items in 1994 and as pretest pool anchor items in 1995. For some new items developed after CAT was implemented, both adaptive and nonadaptive data are available. In addition, some items administered as part of CAT in 1994 and 1995 may have data available from different times of the year. That is, items may have been used between October of 1994 and March of 1995 in one pool and between April and September of 1995 in another pool.

The current operational procedures for NCLEX do not yet include routine recalibration of CAT data. Rather, items that suggest misfit of the existing item difficulty estimates to the CAT data are removed from use as live items and re-pretested to obtain updated item difficulty estimates. Clearly, it would be preferable to directly recalibrate such items, as this would make it possible to continue using the items with the updated estimates. However, in considering the long term effects of CAT item recalibration, it is important to examine the stability of item recalibrations over time and different configurations of item pools. The purpose of this study was to undertake such an investigation.

Sources of Data

In this study, operational NCLEX-RN and NCLEX-PN CAT data collected between April 1994 and September 1995 were utilized. For each measure, three CAT pools were used in the field during this period: one pool from April through September 1994 (these will be referred to as R0494CAT and P0494CAT, respectively), a second pool from October 1994 through March 1995 (R1094CAT and P1094CAT), and a third pool from April through September 1995 (R0495CAT and P0495CAT). The April 1994 and October 1994 pools consisted of nonoverlapping sets of items, while the April 1995 pools consisted of some items from each of the two 1994 CAT pools, plus items that were administered non-adaptively in 1994.

Pretest data collected between April 1994 and September 1995 were also utilized. For NCLEX-RN and NCLEX-PN, there were six pretest pools, each pool was used for approximately three months.

The six pretest pools will be referred to as *0494PRE, *0794PRE, *1094PRE, *0195PRE, *0495PRE, and *0795PRE, where * refers to R or P, respectively. Each pretest pool consisted of a group of previously calibrated anchor items and a larger set of untried items. The pretest and CAT data consisted of only candidates educated in the U.S. and testing for the first time, as these constitute the data samples that have historically been used for the paper-and-pencil calibrations.

In preparation for CAT implementation, large pools of multiple-choice test items were calibrated over a period of years, from before 1991 through 1994, as part of the paper-and-pencil NCLEX. These paper-and-pencil item difficulty estimates were available for all of the items in the April 1994 and October 1994 CAT pools, most of the items in the April 1995 CAT pools, and all of the anchor items in the pretest pools.

Data Preparation

In preparation for the pretest and CAT item calibrations, the item responses for each pool were reformatted into person-by-item matrices, in which correct items were scored as '1', incorrect items were scored as '0', and items that were not administered were scored as '3'. Pretest and CAT items with at least 100 candidate responses were included in these matrices.

For each RN and PN item pool, Table 1 provides the overall sample sizes, numbers of items, and numbers of available items. The first six lines of Table 1 provide this information for the pretest pools, while the last three lines provide this information for the CAT pools. For the RN pretest pools, the numbers of candidates that took April or July pools was between 30,000 and 40,000, and far exceeded the numbers that took the October or January pool. This yearly pattern corresponds to the graduation rates of NCLEX-RN candidates. For the PN pretest pools, the numbers of candidates that took each pool was generally less than the corresponding RN pretest pool volumes. The greatest volume of PN candidates (21,165) took the July 1994 pool and the lowest volume of PN candidates (8,467) took the January 1995 pool. The volumes for the other PN pretest pools fell in between these two extremes.

Insert Table 1 About Here

Table 1 indicates that for every pretest pool, all of the items were selected for the item calibration procedure, since every item was seen by at least 100 candidates. Since the expected candidate volumes were known prior to the creation of each pretest pool, the numbers of items for each pool were chosen to ensure adequate per item sample sizes. The per item sample sizes can be determined by the following. Each candidate receives 15 RN or 25 PN pretest items. Since the items are randomly selected, there is little variation in the per item sample sizes. Therefore, the approximate per item sample size is the number of pretest item events (the total sample size multiplied by 15 for RN or by 25 for PN), divided by the number of pretest items in the pool. For example, the RN April 1994 pretest pool had 491,370 item events (32,758 candidates by 15 items) for a per item sample size of approximately 1,068 (491,370 item events divided by 460 items in the pool), well above the 100 candidate minimum.

For RN CAT, the sample sizes for the April pools were just under 46,000 and the sample size for the October 1994 pool was 22,455. Although the numbers of candidates that took the April 1994 and 1995 pools were considerably larger (71,745 and 72,850, respectively), it was necessary to select a

stratified data sample to accommodate the memory limitations of the item calibration program. For PN, the sample sizes of CAT pools were slightly less than those of the corresponding RN pools and ranged from 17,604 to 36,620. Across all of the RN and PN CAT pools, the percentages of selected items ranged from 70% to 99% of the total number of items in each pool. This is due to the adaptive nature of item selection using CAT. For the RN and PN 1994 pools, from 24% to 30% of the items were not seen by at least 100 candidates. In the case of the April 1995 CAT pools, simulations were used to pre-select items on the basis of item exposure rates. Therefore, only 1% of the RN items and 6% of the PN items were not seen by enough candidates.

Item Calibrations

The computer program LOGIST was used for the item calibrations because it handles large numbers of items and simulees, it provides flexibility in terms of options, and because the item difficulty estimates for the paper-and-pencil NCLEX were previously obtained using a LOGIST-based estimation procedure. Candidate ability estimates based on CAT items were used to fix the scale for estimating both pretest and CAT item difficulties. The use of CAT-based ability estimates has been explored in previous NCLEX research and has provided satisfactory results (Haynie & Way, 1995; Way, 1994).

Once recalibrated difficulty estimates for all the pretest and CAT items were obtained, a further scaling step was executed to adjust for differences between the paper-and-pencil and recalibrated item difficulty estimates. For the pretest items, the means and standard deviations of the recalibrated and paper-and-pencil anchor item estimates were used to transform the recalibrated estimates. Similarly, for the CAT items, the means and standard deviations of the recalibrated and paper-and-pencil item estimates were used to transform the recalibrated estimates. The mean-and-sigma transformations applied the following formulas:

$$B^* = \frac{SD_{B_{OLD}}}{SD_{B_{NEW}}}; \qquad A^* = MEAN_{B_{OLD}} - B^* \times MEAN_{B_{NEW}};$$

$$\hat{B}_{NEW} = A^* + B^* \times B_{NEW}$$

Because scalings of common item parameter estimates with the Rasch model typically do not involve adjustments to the standard deviation of the estimates (as the scale is set by the constant discrimination factor), the use of the mean-and-sigma transformations in this study deserves some comment. Research with the NCLEX using CAT (Haynie & Way, 1995) as well as a study reported by Bergstrom and Lunz (1991) indicated that the standard deviations of recalibrated Rasch item difficulties based on CAT data were markedly greater than the corresponding standard deviations of item difficulties based on calibrating non-adaptive data. The reasons for this are not entirely clear, but are probably related to the targeting of items to candidates that occurs with CAT, which is obviously different than the targeting that occurs in a typical paper-and-pencil exam. In particular, the targeting of items in CAT will cause the slopes of the regressions of the item scores on the criterion scores (estimated abilities) to be higher than they would be in a non-adaptive situation. This may be especially true because the NCLEX paper-and-pencil tests had average p-values of .70 or higher. Because the Rasch model forces a constant item discrimination, the increased slope manifests itself in the variance of the estimated difficulties. The mean-and-sigma transformations serve to reset the scale of the item difficulty estimates to the scale defined by the paper-and-pencil exams.

Tables 2 and 3 provide the A^* and B^* values for the mean-and-sigma transformations. For each set of estimates, the B^* entry in Tables 2 and 3 represents the ratio of the paper-and-pencil to recalibrated estimate standard deviations and the A^* is based on the difference in the means of the estimates. In Table 2, all of the pretest item B^*'s were relatively close to one for both measures, with the exception of the RN October 1994 pool which was based on a much smaller per item sample size than the other RN pretest pools. The pretest A^*'s were close to zero based on all six pools and both measures. The pretest transformation values in Table 2 suggest noted, that the standard deviation adjustments made to the Rasch scale for the pretest items may not have been necessary. In Table 3, all of the CAT item B^*'s are less than 1.0, ranging from .8461 to .9102. This indicates that the variation of the recalibrated CAT estimates is greater than the variation of the paper-and-pencil estimates. The CAT A^*'s were close to zero based on all three pools and both measures.

Insert Tables 2 and 3 About Here

Analysis of the Data

The analyses carried out with the item calibration data consisted primarily of correlations between sets of estimates for items that appeared in multiple item pools. In addition, the relationships between selected sets of estimates were explored graphically using bivariate plots and difference plots. Means and standard deviations of the differences between the sets of estimates were calculated to summarize scaling effects. One limitation of these analyses was that the recalibrated estimates were based on CAT data that were administered and scored using item difficulty estimates obtained from calibrations of non-adaptive data. That is, it was not feasible in this study to recalibrate CAT data to obtain new item difficulty estimates, utilize these new estimates in a subsequent CAT administration, and recalibrate the same items a second time using CAT data.

An Overview of the Calibration Results

The intercorrelations of item difficulty estimates and numbers of common items for each pair of item pools are provided in Table 4 for RN and Table 5 for PN. Pairings with fewer than 15 common items were excluded. Figure numbers are provided for pairings of item pools that are subsequently analyzed in more detail. All correlations included in Tables 4 and 5 were significant ($p < .001$). For each measure, the comparisons can be classified as within administration-type (CAT to CAT or pretest to pretest) or between administration-type (pretest to paper-and-pencil, CAT to pretest, or CAT to paper-and-pencil).

Insert Tables 4 and 5 About Here

Since paper-and-pencil item difficulties were only estimated once for each item, sets of paper-and-pencil item difficulty estimates common to two item pools were perfectly correlated ($r = 1.000$). The exception to this is the RN and PN paper-and-pencil item difficulty estimates common to the 4/94 pretest and 4/95 CAT pools ($r = .999$). For the 4/95 CAT pool, the small number of paper-and-pencil estimates based on pretest anchor items were calibrated using the 7/94 pretest data, not the 4/94 pretest data. It

should be noted that for both measures, the paper-and-pencil estimates used for comparison with the 4/95 CAT pools included a small number of pretest item estimates, not administered via paper-and-pencil. For the purposes of identifying general patterns in the pairs of estimates, these pretest item estimates were included in this group of paper-and-pencil estimates. In a latter section of this paper, the differences between the paper-and-pencil and pretest estimates, in relation to the 4/95 CAT estimates, will be considered.

Within Administration-Type Correlations

The correlations of item difficulty estimates common to the different pairs of CAT pools are provided in the upper left corners of Tables 4 and 5. For RN, the estimates in the 4/95 pool correlated .978 with the 4/94 pool estimates (N=311) and correlated .989 with the 10/94 estimates (N=480). For PN, the estimates in the 4/95 pool correlated .977 with the 4/94 pool estimates (N=330) and correlated .985 with the 10/94 estimates (N=383). The correlations of item difficulty estimates common to the different pairs of pretest pools are provided in middle, center areas of Tables 4 and 5. For every pair of pretest pools, the common items are simply the set of anchor items, re-calibrated using the pretest data. For RN, correlations of estimates based on six pairs of pretest pools were possible. These correlations ranged from .983 to .990. For PN, seven correlations were possible and these correlations ranged from .986 to .991. Since the patterns between the different pairs of pretest pools were very similar, representative pretest pool pairings are subsequently provided in Figures 1 through 4.

Between Administration-Type Correlations

The correlations of item difficulty estimates common to pairings of pretest item pools and paper-and-pencil estimates are provided in the bottom center areas of Tables 4 and 5. For RN, the six pairings of interest are the correlations of paper-and-pencil item difficulty estimates with re-calibrated items using the pretest data; these correlations range from .930 to .962. For PN, the ten pairings of interest involve pretest anchor items and pretest anchor items also included in CAT pools; these item difficulty estimate correlations ranged from .940 to .962. Since the patterns between the different pairs of pretest pools and paper-and-pencil estimates were very similar, representative pairings are subsequently provided in Figures 5 through 8.

The correlations of item difficulty estimates common to pairings of CAT and pretest item pools are provided in the left middle areas of Tables 4 and 5. Two types of comparisons are represented in this pairing: (1) for both measures, common items administered in a pretest pool and subsequently administered in a CAT pool (shown in Figures 13 through 16) and (2) for PN, common items were administered in a 1994 CAT pool and subsequently used in a pretest pool as anchor items (using the paper-and-pencil estimates). For RN, the estimates in the 4/95 CAT pool had a correlation of .974 with the 4/94 pretest pool estimates (N=87) and of .982 with the 7/94 estimates (N=104). For PN, the estimates in the 4/95 CAT pool had a correlation of .969 with the 4/94 pretest pool estimates (N=108) and of .972 with the 7/94 estimates (N=152). The correlations of PN CAT items subsequently administered as pretest (anchor) items ranged from .943 to .971; the patterns between these pairs of estimates are similar to the other CAT to pretest pairings and are not further illustrated in figures.

The correlations of item difficulty estimates common to pairings of CAT items pools and paper-and-pencil estimates are provided in the bottom left corners of Tables 4 and 5. Three types of comparisons are represented in this pairing: (1) all items administered in a CAT pool are compared to the paper-and-pencil estimates (shown subsequently in Figures 17 through 22), (2) subsets of items administered in a CAT pool are compared to the paper-and-pencil estimates (not represented by additional

figures) and (3) for PN, common items administered in a 1994 CAT pool are compared to items subsequently used as 4/95 pretest pool anchor items (these comparisons are similar to other CAT to paper-and-pencil estimate comparisons and are not shown in additional figures). For RN, the correlations between each CAT pool and full set of paper-and-pencil estimates were .929 for 4/94, .926 for 10/94, and .946 for 4/95. For PN, the correlations between each CAT pool and full set of paper-and-pencil estimates were .932 for 4/94, .922 for 10/94, and .942 for 4/95.

Graphical Results of the Calibrations

Graphical and statistical results of the calibrations are provided in Figures 1 through 22 for item pool pairings of interest. The figures on the left are plots of the transformed item difficulty estimates. The figures on the right are plots of the difference between the two sets of item difficulty estimates on the Y-axis, and the paper-and-pencil, pretest, or older set of difficulty estimates (in that order of precedence) on the X-axis. Figures 1 through 4 are comparisons of pretest estimates, Figures 5 through 8 are comparisons of pretest and paper-and-pencil (labelled 'Anc') estimates, Figures 9 through 12 are comparisons of CAT estimates, Figures 13 through 16 are comparisons of CAT and pretest estimates, and Figures 17 through 22 are comparisons of CAT and paper-and-pencil estimates. Calibrations results based on RN data are shown in Figures 1, 2, 5, 6, 9, 10, 13, 14, 17, 18, and 19, and results based on PN data are shown in Figures 3, 4, 7, 8, 11, 12, 15, 16, 20, 21, and 22. The accompanying tables to each of the figures show summary statistics for the difference between the item difficulties and the item difficulty estimates themselves. Correlations between the item difficulties and item difficulty differences are also provided. Significant correlations ($p < .01$) are indicated by a '*'.

Insert Figures 1 through 22 About Here

The plots shown in Figures 23 and 24 show the 4/95 CAT and paper-and-pencil estimate comparisons for different subsets of paper-and-pencil estimates based on calibration date. The following ranges of paper-and-pencil calibration dates are represented in the plots: (1) items pretested within the context of CAT in 1994 (upper left), (2) items calibrated in 1993 or 1994 based on paper-and-pencil data (upper right), (3) items calibrated in 1991 or 1992 (bottom left), and (4) items calibrated before 1991 (bottom right). Correlations between the CAT and paper-and-pencil estimates are provided for each figure. The overall 4/95 CAT and paper-and-pencil estimate comparisons are shown in Figure 19 for RN and Figure 22 for PN.

Insert Figures 23 and 24 About Here

Pretest to Pretest Estimate Comparisons

Item difficulty estimate plots for pairs of pretest pools are presented in Figures 1 through 4. The plots on the left indicate that the estimates have a linear relationship and are highly correlated ($r = .98$ to .99). The difference plots on the right indicate no systematic patterns in the differences. The summary statistics also describe these relationships. The differences between the item difficulty estimates have standard deviations of .13 to .16. The differences are not significantly correlated with any of the sets of estimates. Finally, due to the item difficulty estimate transformation procedures, the means of the item difficulty estimates are identical for the items in Figures 1, 3, and 4. For Figure 2, the average

difference of +.039 indicates that the small set of items (N=20) was found to be slightly easier from January to March 1995 than from April to June 1994. (These 20 items were part of the full set of anchor items that were used from April to June of 1994.)

Pretest to Paper-and-Pencil Estimate Comparisons

Item difficulty estimate plots for pairs of pretest pools with paper-and-pencil estimates are presented in Figures 5 through 8. The plots on the left indicate that the relationships between the estimates are linear and fairly strong (r=.94 to .96). Based on the difference plots, there appear to be no systematic patterns in the differences. The differences of the item difficulty estimates have standard deviations of about .25 to .31, which are considerably greater than the standard deviations of the differences between the pretest to pretest item difficulty estimates. The differences are not significantly correlated with any of the sets of estimates. Finally, the means of the item ability estimates are identical or nearly identical for each pair of items in Figures 5 through 8.

CAT to CAT Estimate Comparisons

Figures 9 through 12 provide item difficulty estimate plots for all possible pairs of CAT pools. The plots on the left indicate that the relationships between the estimates are strong (r=.98 to .99). The relationships also appear to be linear, although the difference plot in Figure 11 suggests a small curvilinear tendency. The relationships are also more diffuse at the lower end of the difficulty scale. For both measures, the relationship between the 10/94 and 4/95 estimates is slightly stronger than the relationship between the 4/94 and 4/95 estimates. The summary statistics also describe this relationship. The differences of the item difficulty estimates have standard deviations of .11 to .15, which are slightly lower than the standard deviations of the differences between the pretest to pretest item difficulty estimates. The differences are positively correlated with the 4/95 difficulty estimates in Figures 9, 10, and 12 and the 10/94 estimates in Figure 10 (r=+.19 to +.33). The differences are negatively correlated with the 4/94 difficulty estimates in Figures 11 (r=-.21). The means of the item ability estimates are nearly identical for the RN CAT estimates; however, the PN CAT estimates are slightly easier (.015) from April through September 1995 than from April through September 1994 (Figure 11) or from October 1994 through March 1995 (Figure 12).

CAT to Pretest Estimate Comparisons

Figures 13 through 16 provide item difficulty estimate plots for all possible pairs of CAT and pretest pools. The plots on the left indicate that the relationships between the estimates are strong (r= .97 to .98). The relationships are linear, and are, perhaps, slightly more spread out at the low end of the scale. Based on the summary statistics, the differences of the item estimates have standard deviations of .18 to .20, indicating moderate spread, overall. The differences are negatively correlated with the RN 4/94 and 7/94 pretest estimates in Figures 13 and 14 (r=-.41 and r=-.38, respectively). The means of the RN 4/95 CAT estimates and RN 7/94 pretest estimates (Figure 14) differ by -.016; while the means of the PN 4/95 CAT estimates differ from the PN 4/94 and 7/94 pretest estimates by -.046 and -.049, respectively.

CAT to Paper-and-Pencil Estimate Comparisons

The item difficulty estimate plots for all pairs of CAT pools with paper-and-pencil estimates are presented in Figures 17 through 22. The plots on the left indicate that the relationships between the estimates are somewhat close (r=.92 to .95). The relationships appear to be relatively linear, although

the right hand plots in Figures 17 and 18 appear to be slightly curvilinear. Based on the difference plots, the spreads of the differences are somewhat closer in the center and more diffuse at the high ends and especially the low ends of the difficulty scales. The differences of the item difficulty estimates have standard deviations of about .25 to .29, which are similar to the standard deviations of the differences between the pretest and paper-and-pencil difficulty estimates. The differences are negatively correlated with each set of paper-and-pencil estimates (r=-.20 to -.16) and positively correlated with each set of CAT estimates (r=+.16 to +.20). Since the CAT difficulty estimates were transformed to the paper-and-pencil estimates using the mean-and-sigma method, the means of the two sets of estimates are identical.

Estimate Comparisons as a Function of Time

Figures 23 and 24 present the relationships of 4/95 CAT estimates to different sets of paper-and-pencil estimates based on the paper-and-pencil calibration date. The upper left plots are based on data from the 4/94 or 7/94 pretest pools. Since this is a CAT to pretest pool comparison, the upper left plot in Figure 23 presents the data from the plots in Figures 13 and 14 combined. Similarly, the upper left plot in Figure 24 is derived from Figures 15 and 16. The relationships between these two sets of estimates are linear and strong (r=.980 and .972). The relationships between CAT estimates and paper-and-pencil estimates calibrated from 1993 and 1994 (upper right plots) are also linear, but indicate a less strong relationship (r=.940 and .947). The plots in the lower left provide the relationships between 4/95 CAT estimates and paper-and-pencil estimates from 1991 and 1992 and the plots in the lower right provide the relationships between the CAT estimates and the paper-and-pencil estimates calibrated from before 1991. These four plots indicate weaker relationships (r=.924 to .936) between the CAT estimates and older paper-and-pencil estimates.

Discussion of the Results

The purpose of this study was to examine NCLEX item recalibrations over time and different configurations of the item pools. Such changes in item difficulty may be the result of scaling effects, context effects, and/or item parameter drift. Scaling effects may be related differences in the type of data that contribute to the item calibrations (i.e., adaptive or non-adaptive). Context effects can be due to differential item ordering, changes in the candidate population, and/or mode of administration (paper-and-pencil or computer-based test delivery). Item parameter drift may occur due to changes in educational curricula, practice, and/or technology, among other things.

The item pools investigated in this study were classified by item type (CAT, pretest, or paper-and-pencil) and by pool date (different times of the year). For every within administration-type comparison (CAT or pretest), the associations were very high (r=.98 and .99) and the relationships were linear. The associations for each CAT to pretest comparison were only slightly weaker (r=.97 to .98).

The results of the study also indicated that the recalibrated CAT item difficulties correlated highly with paper-and-pencil item difficulty estimates (r= .92 to .95); however, these correlations were slightly lower than the within administration-type correlations. To explore the stability of item recalibrations as a function of time since the previous calibrations occurred, CAT item difficulty estimates from the April 1995 CAT pools were compared to different sets of estimates that were categorized by date of previous calibration. The associations between the CAT estimates and pretest estimates calibrated in 1994 (from computer-administered but nonadaptive data) were considerably stronger than the associations between the CAT estimates and the 1993-1994 paper-and-pencil estimates. Since the pretest and paper-and-pencil

calibration dates are similar, this effect is more likely to be due to the mode of administration and/or differential item ordering than to a factor such as item parameter drift. The comparatively weaker associations between the CAT estimates and the older sets of paper-and-pencil estimates may have been indicative of some item parameter drift, but this trend was more noticeable for PN than for RN.

Although the general results of this study were very encouraging, one puzzling finding was the magnitudes of the correlations between the original and/or recalibrated item difficulties and the difference between the recalibrated and original item difficulties. For example, the differences between the recalibrated difficulties based on CAT administration and the original difficulties based on non-adaptive pretest items were negatively correlated with the original difficulties. Also, the relationships between the CAT recalibrated item difficulties and the original paper-and-pencil item difficulties show a consistent pattern in which the recalibrated difficulty estimates are negatively correlated with the differences between the recalibrated and original estimates. In each of these comparisons, the recalibrations were based on CAT administration and the original calibrations were based on non-adaptive administration. This pattern is seen as well in the comparison of recalibrated and original item difficulties that are both based on CAT administration; the differences between the recalibrated and original estimates have moderately positive correlations with the recalibrated estimates.

One reason these trends may be of concern is that, although the mean differences between the original and recalibrated the estimates tend to be near zero, more systematic differences may exist between the original and recalibrated difficulty estimates as the estimates become more extreme. To some extent, this may be related to the scaling transformations that are applied when the recalibrated item difficulty estimates are linked to the scale of the original estimates. Thus, despite the overall encouraging results seen in this study, caution should be taken to make sure that original transformations of recalibrated CAT data do not introduce systematic effects that could accumulate over repeated calibrations.

For the operational NCLEX, recalibration in the context of CAT will allow the replenishment of item difficulty estimates that are outdated due to reasons such as item parameter drift. The results of this study suggest that item recalibration in the context of CAT is supported; however, the implementation of these procedures should proceed with caution. In particular, more research is needed to explore the implications of negative correlations between difficulty estimates and estimate differences; and further study of the mean-and-sigma scaling method in the context of CAT item recalibrations will be of benefit.

References

Bergstrom, B.A., & Lunz, M.E. (1991, April). Equivalence of Rasch item calibration and ability estimates across modes of administration. Paper presented at the International Objective Measurement Workshop, Chicago, IL.

Divgi (1986). Determining the sensitivity of CAT-ASVAB scores to changes in item response curves with the medium of administration. Alexandria, VA: Center for Naval Analyses.

Eignor, D.R., Way, W.D., & Amoss, K.E. (1994, April). Establishing the comparability of the NCLEX using CAT™ with traditional NCLEX examinations. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Haynie, K.A., & Way, W.D. (1995, April). An investigation of item calibration procedures for a computerized licensure examination. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Haynie, K.A., & Way, W.D. (in preparation). An investigation of item calibration procedures for a computerized licensure examination (NCLEX Joint Research Committee Report). Princeton, NJ: Chauncey Group International.

Holland, P.W., & Wingersky, M. (1991, May). A monte carlo comparison of four approaches to on-line calibration of a CAT. Presentation at the ONR Workshop on Model-based Measurement, Princeton, NJ.

Ito, K., & Sykes, R.C. (1994, April). The effect of restricting ability distributions in the estimation of item difficulties: Implications for a CAT implementation. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Lunz, M. E., & Bergstrom, B. A. (1991). Comparability of decisions for computer adaptive and written examinations. Journal of Allied Health, 20, 15-23.

Mislevy, R. J., Wingersky, M. S., & Kingston, M. (1990). Evaluation of a procedure for calibrating "seeded" test items. Research Triangle Park, NC: U.S. Army Research Office.

Stocking, M.L. (1988). Scale drift in on-line calibration (Research Report 88-28-ONR). Princeton, NJ: Educational Testing Service.

Sykes, R.C., & Fitzpatrick, A.R. (1992). The stability of IRT b-values. Journal of Educational Measurement, 29, 201-211.

Way, W.D. (1994, April). Psychometric results of the NCLEX™ beta test. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Way, W.D. & Haynie, K.A. (1994). NCLEX simulations report for April 1994. Princeton, NJ: Educational Test Service.

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

Wingersky, M.L., Patrick, R., & Lord, F.M. (1995). Logist User's Guide (Version 7.1). Princeton, NJ: Educational Testing Service.

Zara, T. (1992, April). A comparison of computer adaptive and paper-and-pencil versions of the National Registered Nurse Licensure Examination. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Zwick, R., Thayer, D.T., & Wingersky, M. (1994). Effect of Rasch calibration and DIF estimation in computer-adaptive tests (Research Report 94-32). Princeton, NJ: Educational Testing Service.

Table 1
Numbers of Candidates, Items, and Available Items
for NCLEX-RN and NCLEX-PN Item Pools

Pool Name	RN			PN		
	Number of Candidates	Total # Items	Selected Items	Number of Candidates	Total # Items	Selected Items
4/94 Pre	32,758	460	460	15,111	486	486
7/94 Pre	38,559	684	684	21,165	718	718
10/94 Pre	4,576	200	200	9,264	301	301
1/95 Pre	18,192	200	200	8,467	170	170
4/95 Pre	32,636	523	523	9,006	470	470
7/95 Pre	40,112	830	830	17,733	790	790
4/94 CAT	45,920	1798	1289	36,620	1484	1129
10/94 CAT	22,455	1783	1287	17,604	1469	1029
4/95 CAT	45,532	1243	1229	27,282	1115	1045

Table 2
Mean-and-Sigma Transformation Values
for Recalibrated Pretest Item Parameter Estimates

	April 1994	July 1994	Oct. 1994	Jan. 1995	April 1995	July 1995
A^* (RN)	.0468	.0381	.0678	.0548	.0029	.0088
B^* (RN)	1.0468	1.0706	1.1234	1.0658	.9898	1.0093
A^* (PN)	-.0050	-.0199	.0517	-.0004	-.0075	.0094
B^* (PN)	.9438	.9531	.9783	.9533	.9649	.9835

Table 3
Mean-and-Sigma Transformation Values
for Recalibrated CAT Item Parameter Estimates

	April 1994	Oct. 1994	April 1995
A^* (RN)	-.0440	-.0542	-.0377
B^* (RN)	.8533	.8683	.9102
A^* (PN)	-.0405	-.0612	-.0425
B^* (PN)	.8672	.8461	.8892

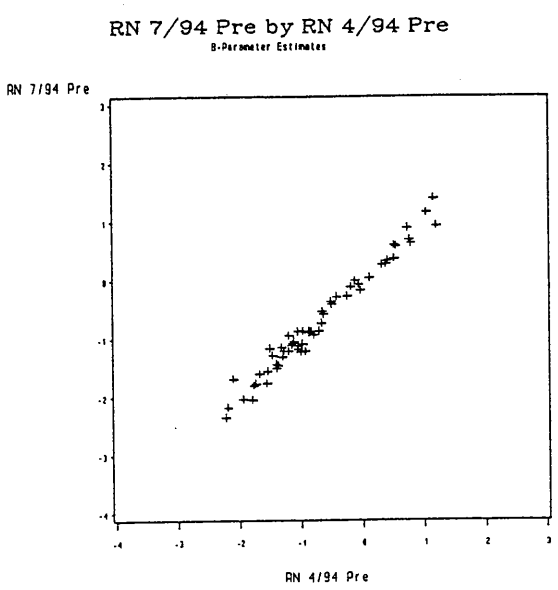
Intercorrelations and Numbers of Common Items for Pairs of NCLEX-RN Pools

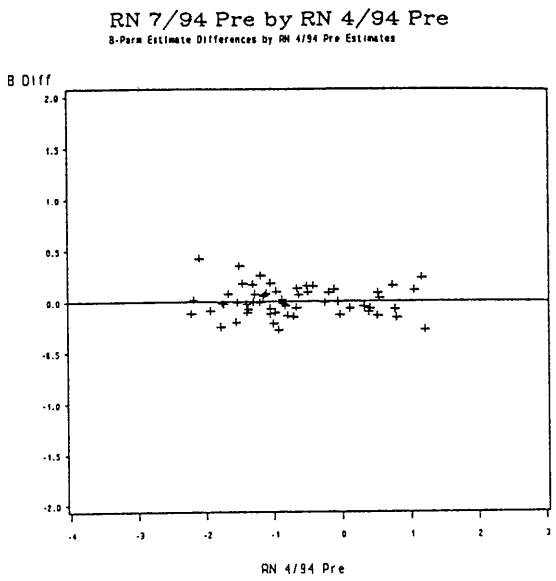
Pool/ Item Type	R0494 CAT	R1094 CAT	R0495 CAT	R0494 PRE	R0794 PRE	R1094 PRE	R0195 PRE	R0495 PRE	R0795 PRE	R0494 P&P-C	R1094 P&P-C	R0495 P&P-C	R0494 P&P-P	Pool/ Item Type
R1094 CAT														.094 CAT
R0495 CAT	Fig. 9 r=.978 N=311	Fig. 10 r=.989 N=480												R0495 CAT
R0494 PRE			Fig. 13 r=.974 N=87											R0494 PRE
R0794 PRE			Fig. 14 r=.982 N=104	Fig. 1 r=.987 N=60										R0794 PRE
R1094 PRE				r=.990 N=20	r=.983 N=20									R1094 PRE
R0195 PRE				Fig. 2 r=.984 N=20	r=.989 N=20									R0195 PRE
R0495 PRE									r=.987 N=70					R0495 PRE
R0795 PRE														R0795 PRE
R0494 P&P-C	Fig. 17 r=.929 N=1289		r=.938 N=365											R0494 P&P-C
R1094 P&P-C		Fig. 18 r=.926 N=1287	r=.936 N=526											R1094 P&P-C
R0495 P&P-C	r=.951 N=311	r=.930 N=480	Fig. 19 r=.946 N=1229	r=.999 N=88	r=1.000 N=104					r=1.000 N=370	r=1.000 N=532			R0495 P&P-C
R0494 P&P-P				Fig. 5 r=.939 N=60	r=.945 N=60	r=.951 N=20	r=.930 N=20							R0494 P&P-P
R0495 P&P-P								r=.962 N=70	Fig. 6 r=.960 N=70					R0495 P&P-P

Intercorrelations and Numbers of Common Items for Pairs of NCLEX-PN Pools

P1094 CAT														P1094 CAT
P0495 CAT	Fig. 11 r=.977 N=330	Fig. 12 r=.985 N=383												P0495 CAT
P0494 PRE			Fig. 15 r=.969 N=108											P0494 PRE
P0794 PRE			Fig. 16 r=.972 N=152	r=.991 N=70										P0794 PRE
P1094 PRE				r=.991 N=70	Fig. 3 r=.990 N=70									P1094 PRE
P0195 PRE				r=.991 N=20	r=.986 N=20	r=.989 N=20								P0195 PRE
P0495 PRE	r=.959 N=16	r=.971 N=33												P0495 PRE
P0795 PRE	r=.943 N=16	r=.962 N=33						Fig. 4 r=.988 N=70						P0795 PRE
P0494 P&P-C	Fig. 20 r=.932 N=1129		r=.938 N=358					r=.957 N=18	r=.944 N=18					P0494 P&P-C
P1094 P&P-C		Fig. 21 r=.922 N=1029	r=.929 N=436					r=.944 N=36	r=.940 N=36					P1094 P&P-C
P0495 P&P-C	r=.950 N=340	r=.928 N=384	Fig. 22 r=.942 N=1045	r=.999 N=110	r=1.000 N=156					r=1.000 N=395	r=1.000 N=463			P0495 P&P-C
P0494 P&P-P				r=.962 N=70	r=.961 N=70	Fig. 7 r=.956 N=70	r=.949 N=20							P0494 P&P-P
P0495 P&P-P	r=.976 N=16	r=.952 N=33						Fig. 8 r=.958 N=70	r=.956 N=70	r=1.000 N=18	r=1.000 N=36			P0495 P&P-P
Pool/ Item Type	P0494 CAT	P1094 CAT	P0495 CAT	P0494 PRE	P0794 PRE	P1094 PRE	P0195 PRE	P0495 PRE	P0795 PRE	P0494 P&P-C	P1094 P&P-C	P0495 P&P-C	P0494 P&P-P	Pool/ Item Type

Figure 1

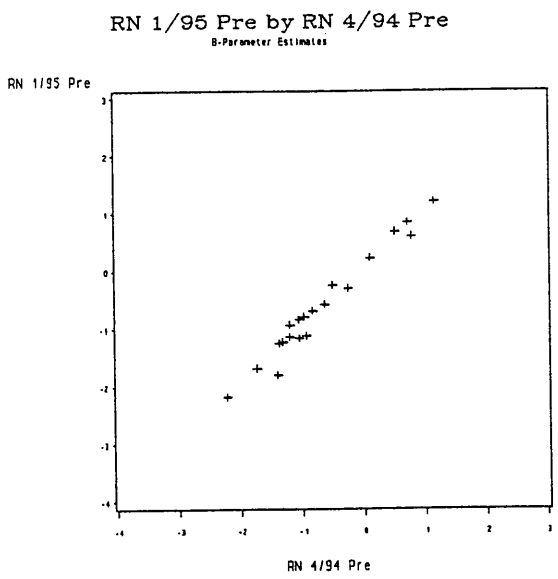


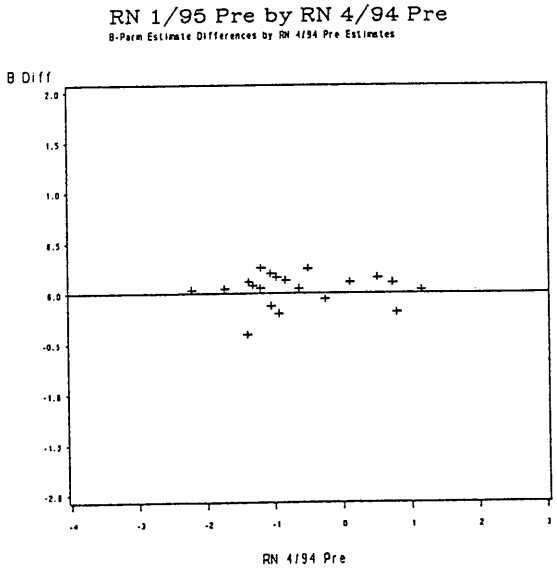


Variable	N	Minimum	Maximum	Mean	Std Dev
DIF	60	-0.2799840	0.4209030	0.0000000	0.1414273
R0494PR	60	-2.2201490	1.2072630	-0.7117764	0.8784298
R0794PR	60	-2.3401400	1.3931150	-0.7117762	0.8784298

Correlations	DIF	R0494PR
R0494PR	-0.08050	
R0794PR	0.08050	0.98704*

Figure 2

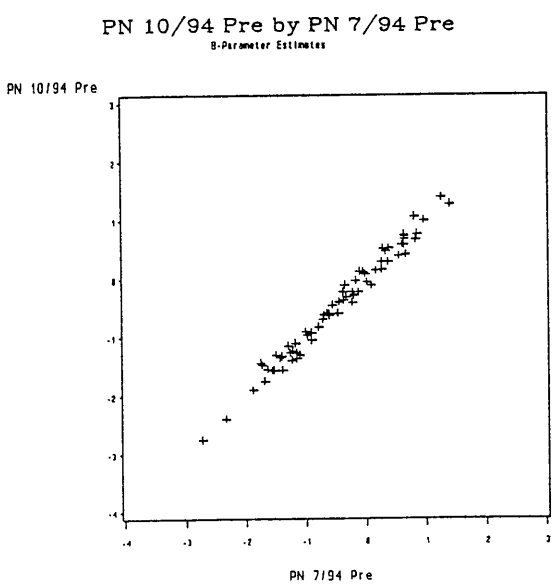


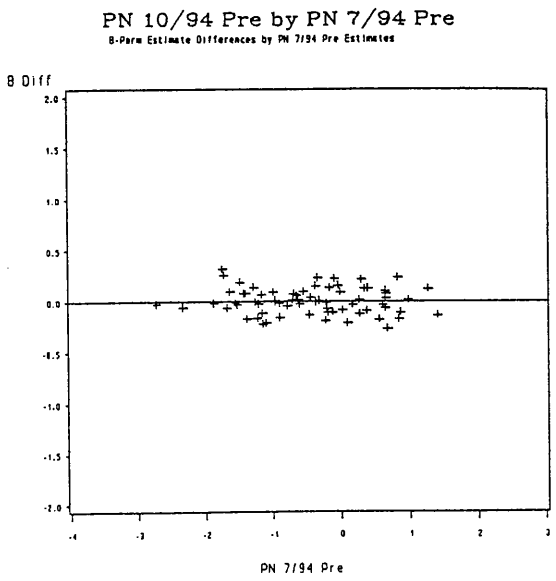


Variable	N	Minimum	Maximum	Mean	Std Dev
DIF	20	-0.4083210	0.2539310	0.0385278	0.1608587
R0494PR	20	-2.2201490	1.1614660	-0.6645324	0.8854963
R0195PR	20	-2.1848950	1.1936700	-0.6260046	0.9041419

Correlations	DIF	R0494PR
R0494PR	0.02630	
R0195PR	0.20367	0.98406*

Figure 3

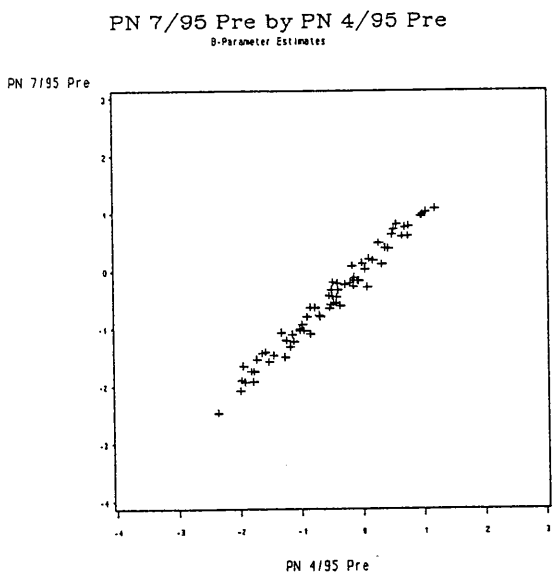


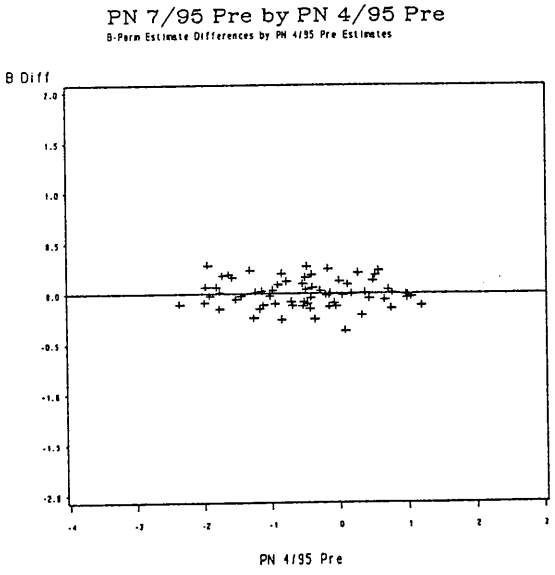


Variable	N	Minimum	Maximum	Mean	Std Dev
DIF	70	-0.2670850	0.3145620	0.0000000	0.1278141
P0794PR	70	-2.7243790	1.4078360	-0.4568555	0.8970536
P1094PR	70	-2.7523510	1.3883510	-0.4568555	0.8970536

Correlations	DIF	P0794PR
P0794PR	-0.07124	
P1094PR	0.07124	0.98985*

Figure 4

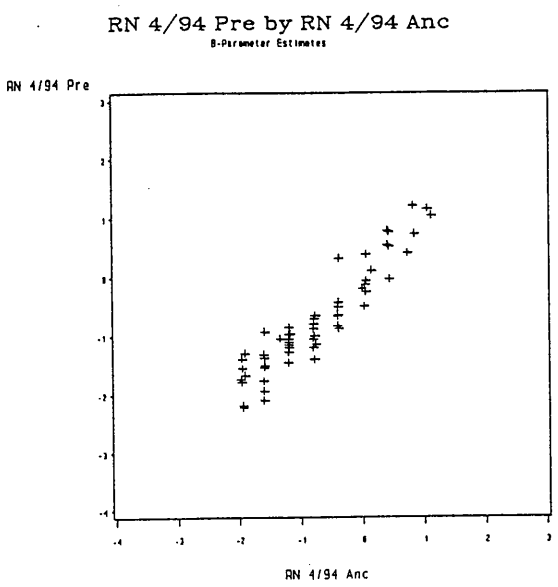


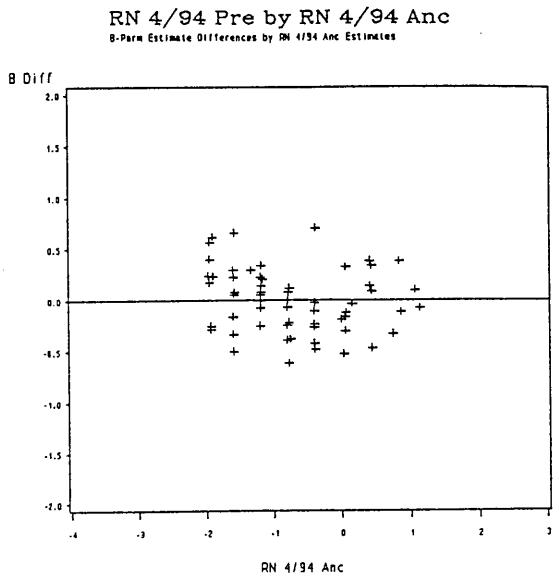


Variable	N	Minimum	Maximum	Mean	Std Dev
DIF	70	-0.3723090	0.2857440	0.0000000	0.1381251
P0495PR	70	-2.3611550	1.1988100	-0.4901982	0.8798797
P0795PR	70	-2.4666310	1.0764100	-0.4901984	0.8798800

Correlations	DIF	P0495PR
P0495PR	-0.07849	
P0795PR	0.07849	0.98768*

Figure 5

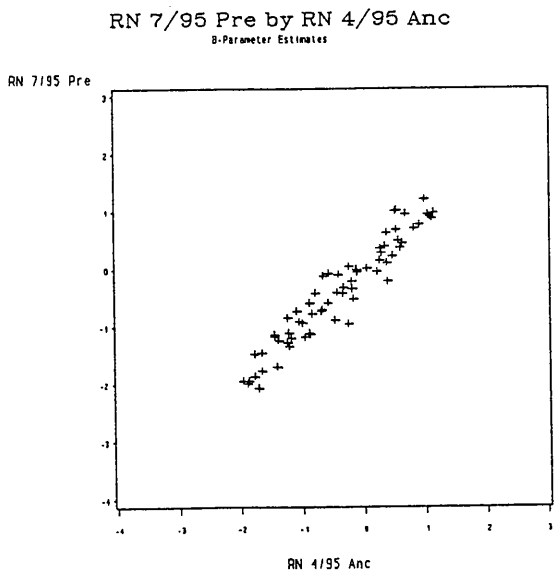


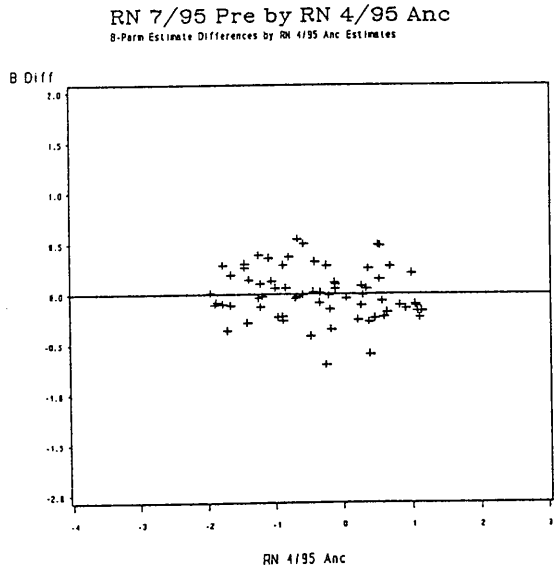


Variable	N	Minimum	Maximum	Mean	Std Dev
DIF	60	-0.6218360	0.7015980	-0.0043880	0.3090019
R0494PA	60	-1.9696000	1.1296000	-0.7073883	0.8834418
R0494PR	60	-2.2201490	1.2072630	-0.7117764	0.8784298

Correlations	DIF	R0494PA
R0494PA	-0.19106	
R0494PR	0.15962	0.93850*

Figure 6

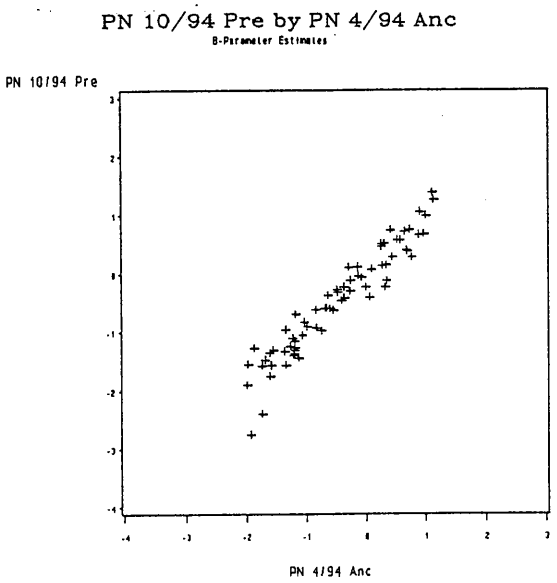


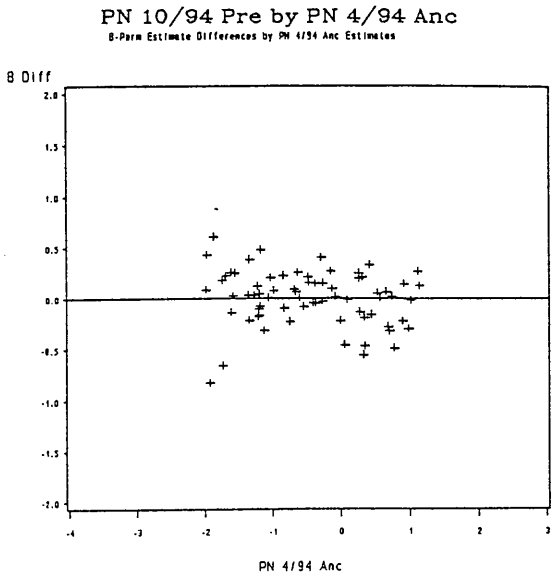


Variable	N	Minimum	Maximum	Mean	Std Dev
DIF	70	-0.7087960	0.5435320	0.0000000	0.2509455
R0495PA	70	-1.9579000	1.1454000	-0.3743929	0.8827597
R0795PR	70	-2.0734630	1.2031200	-0.3743950	0.8827692

Correlations	DIF	R0495PA
R0495PA	-0.14210	
R0795PR	0.14217	0.95959*

Figure 7

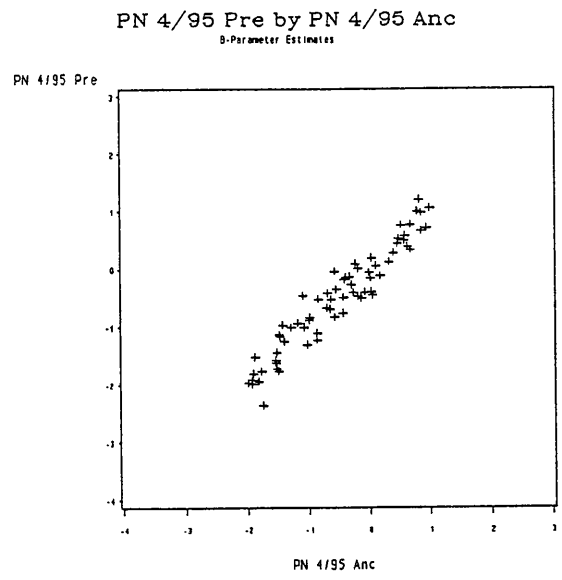


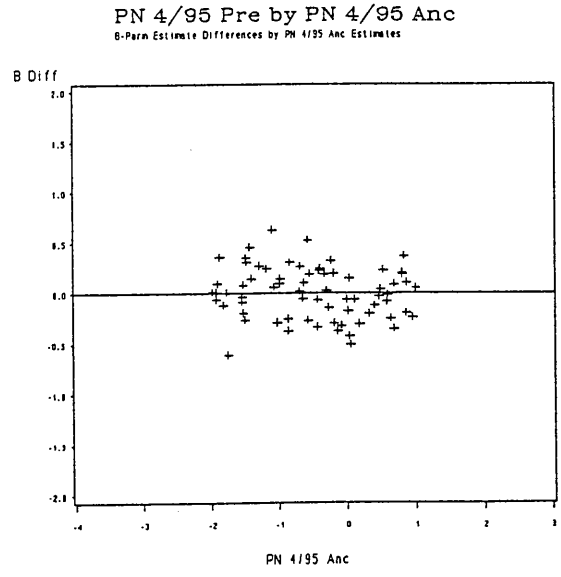


Variable	N	Minimum	Maximum	Mean	Std Dev
DIF	70	-0.8264510	0.6029810	0.0000000	0.2646550
P0494PA	70	-1.9834000	1.1420000	-0.4568557	0.8970540
P1094PR	70	-2.7523510	1.3883510	-0.4568555	0.8970536

Correlations	DIF	P0494PA
P0494PA	-0.14751	
P1094PR	0.14751	0.95648*

Figure 8

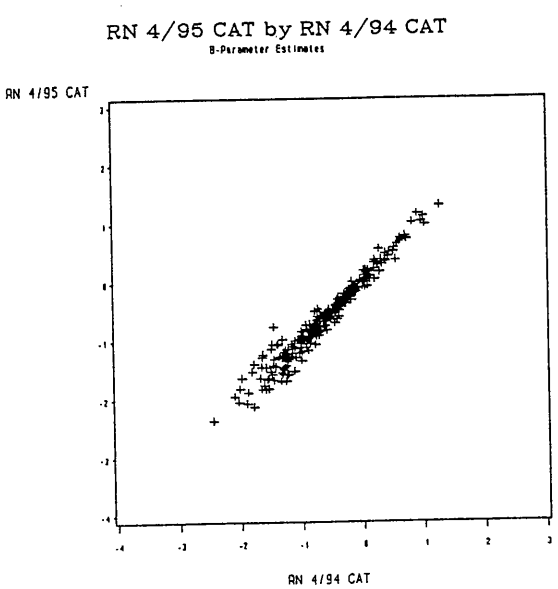


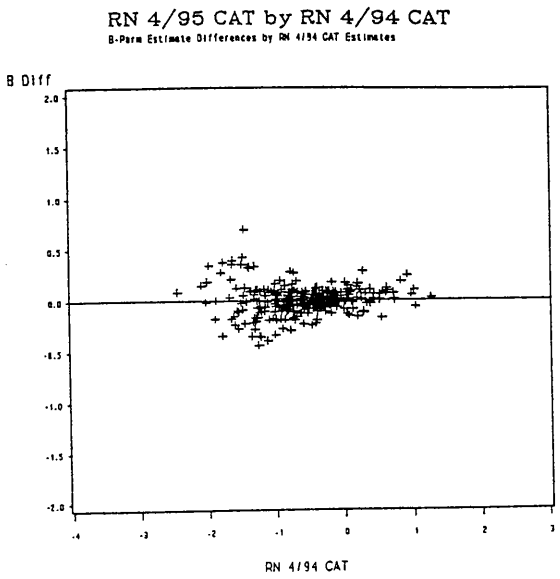


Variable	N	Minimum	Maximum	Mean	Std Dev
DIF	70	-0.6103550	0.6276190	0.0004446	0.2535736
P0495PA	70	-1.9834000	1.0026000	-0.4906429	0.8793424
P0495PR	70	-2.3611550	1.1988100	-0.4901982	0.8798797

Correlations	DIF	P0495PA
P0495PA	-0.14206	
P0495PR	0.14621	0.95845*

Figure 9

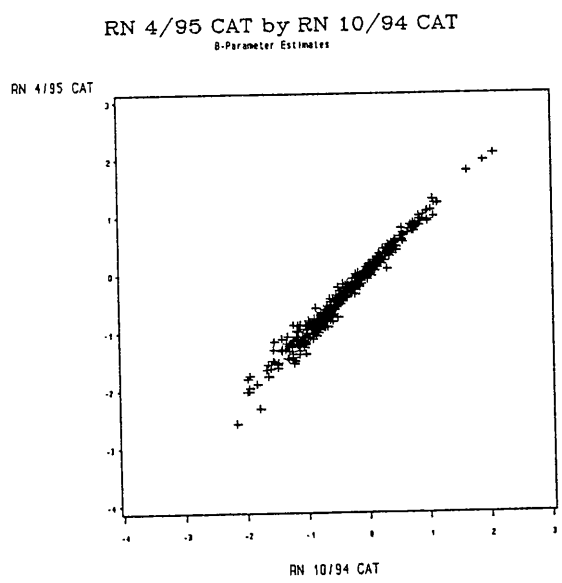


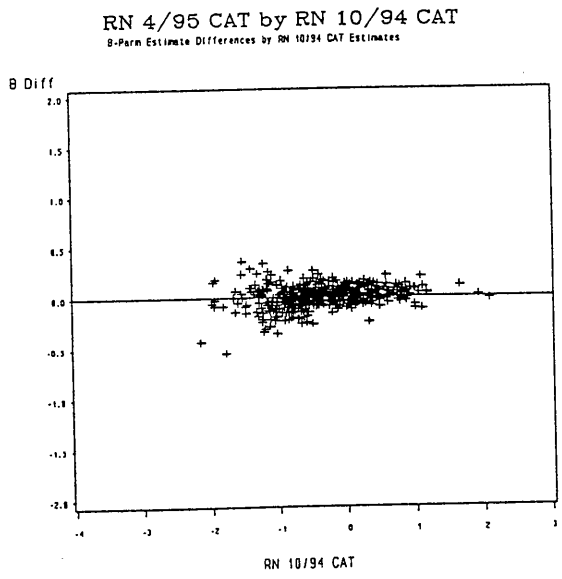


Variable	N	Minimum	Maximum	Mean	Std Dev
DIF	311	-0.4405120	0.6930880	0.0003271	0.1374771
R0494CT	311	-2.4459530	1.2791600	-0.5794600	0.6355661
R0495CT	311	-2.3649870	1.3009850	-0.5791329	0.6555196

Correlations	DIF	R0494CT
R0494CT	0.03927	
R0495CT	0.24779*	0.97780*

Figure 10

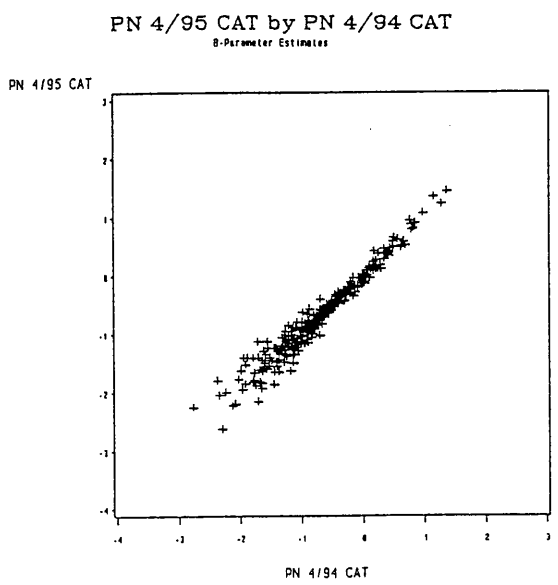


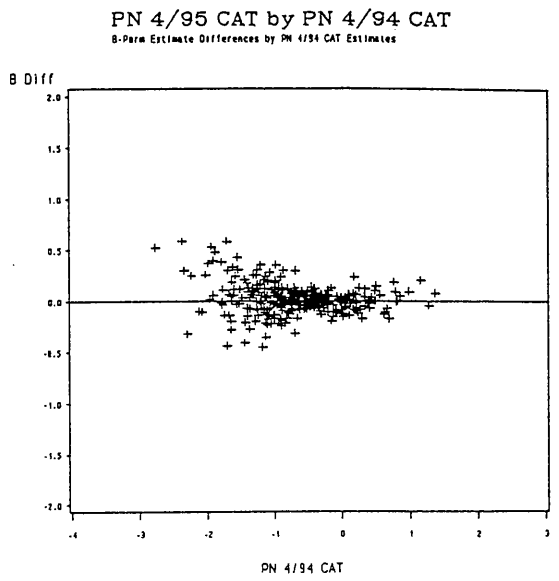


Variable	N	Minimum	Maximum	Mean	Std Dev
DIF	480	-0.5494870	0.3606590	0.0027332	0.1058192
R1094CT	480	-2.1511410	2.0963540	-0.3921798	0.6652381
R0495CT	480	-2.5841060	2.0747460	-0.3894466	0.6931006

Correlations	DIF	R1094CT
R1094CT	0.18928*	
R0495CT	0.33435*	0.98870*

Figure 11

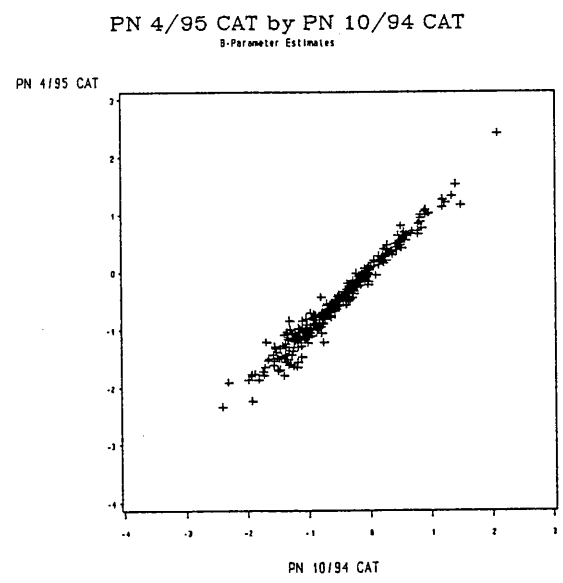


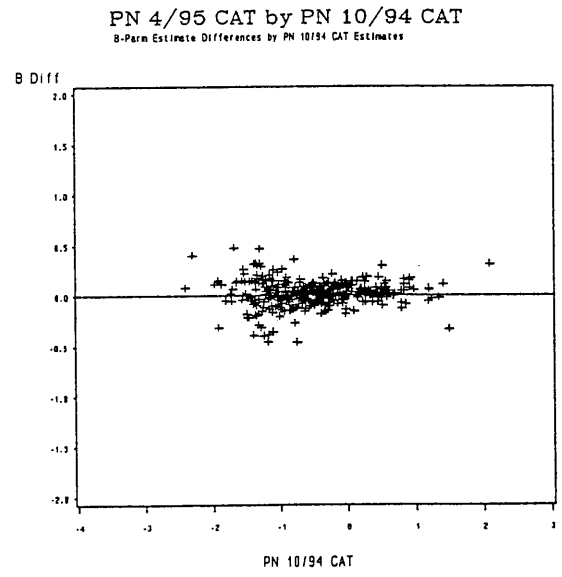


Variable	N	Minimum	Maximum	Mean	Std Dev
DIF	330	-0.4534670	0.5825700	0.0150801	0.1468412
P0494CT	330	-2.7679950	1.3735060	-0.6916491	0.6959584
P0495CT	330	-2.6192550	1.4490150	-0.6765690	0.6801869

Correlations	DIF	P0494CT
P0494CT	-0.21168*	
P0495CT	-0.00071	0.97749*

Figure 12

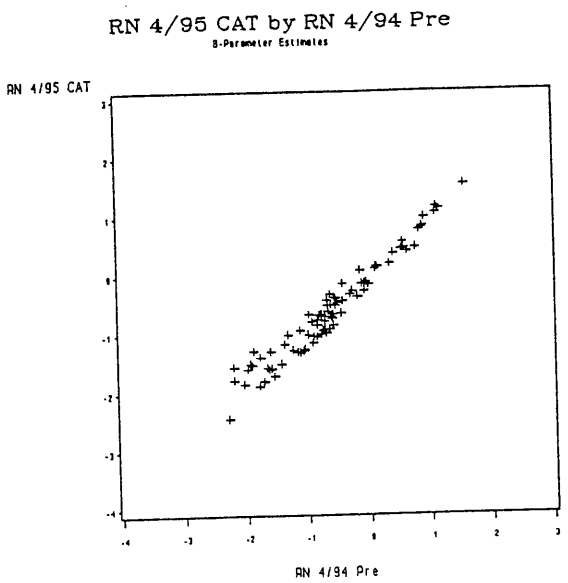


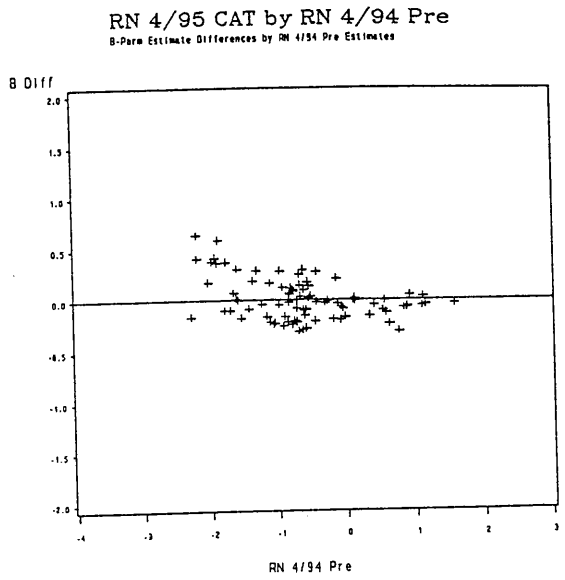


Variable	N	Minimum	Maximum	Mean	Std Dev
DIF	383	-0.4626680	0.4746660	0.0157747	0.1256743
P1094CT	383	-2.4131890	2.0930570	-0.5052673	0.7062528
P0495CT	383	-2.3377140	2.4035750	-0.4894926	0.7196553

Correlations	DIF	P1094CT
P1094CT	0.01868	
P0495CT	0.19297*	0.98464*

Figure 13

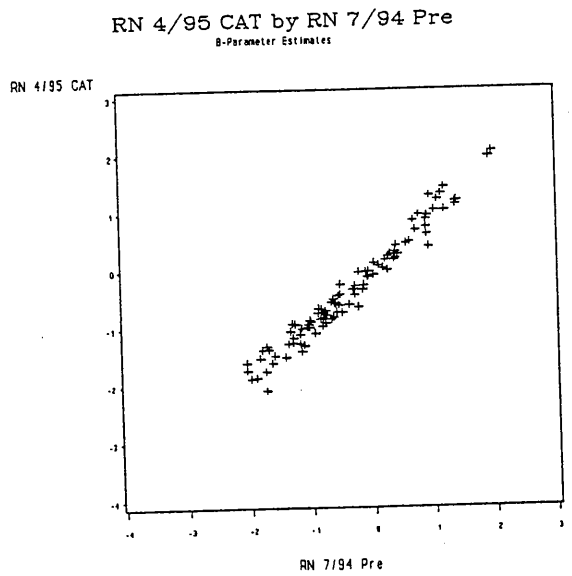


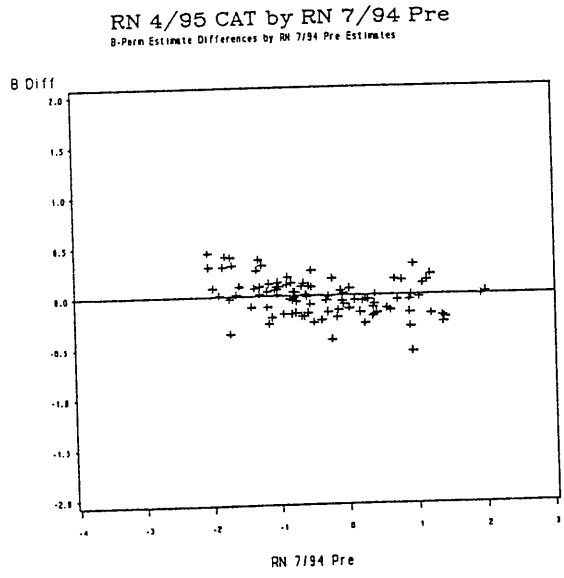


Variable	N	Minimum	Maximum	Mean	Std Dev
DIF	87	-0.3075760	0.6396510	0.0062789	0.2013190
R0494PR	87	-2.2803170	1.5796210	-0.6067845	0.8710893
R0495CT	87	-2.4398040	1.5413290	-0.6005057	0.8106242

Correlations	DIF	R0494PR
R0494PR	-0.40548*	
R0495CT	-0.18737	0.97389*

Figure 14

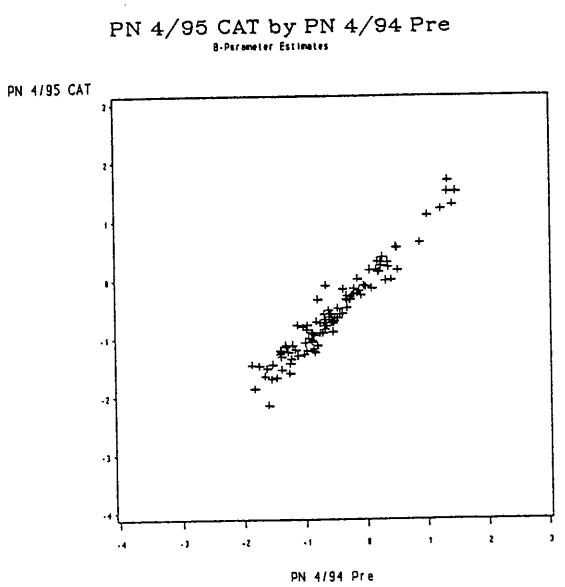


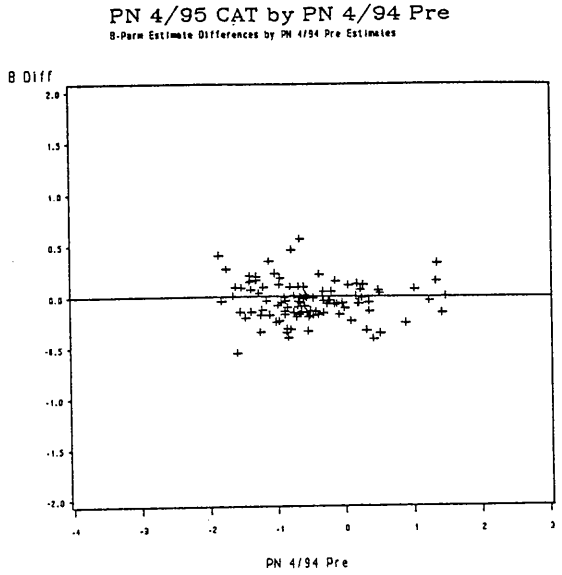


Variable	N	Minimum	Maximum	Mean	Std Dev
DIF	104	-0.5656120	0.4349240	-0.0156995	0.1824930
R0794PR	104	-2.0466740	2.0084370	-0.3366242	0.9547570
R0495CT	104	-2.0956980	2.0316580	-0.3523236	0.9014319

Correlations	DIF	R0794PR
R0794PR	-0.37961*	
R0495CT	-0.19962	0.98230*

Figure 15

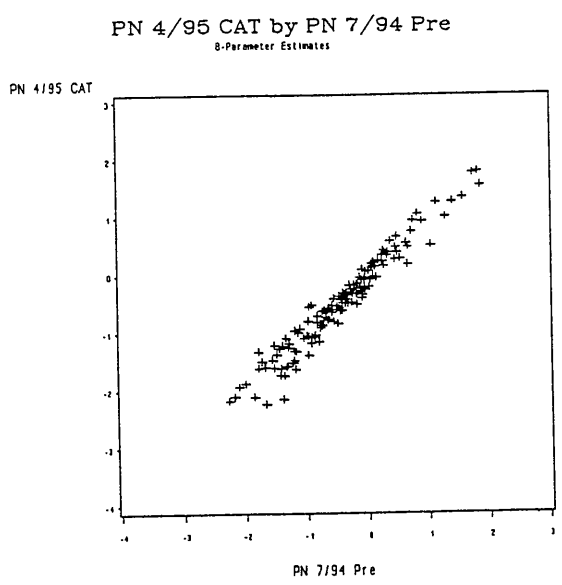


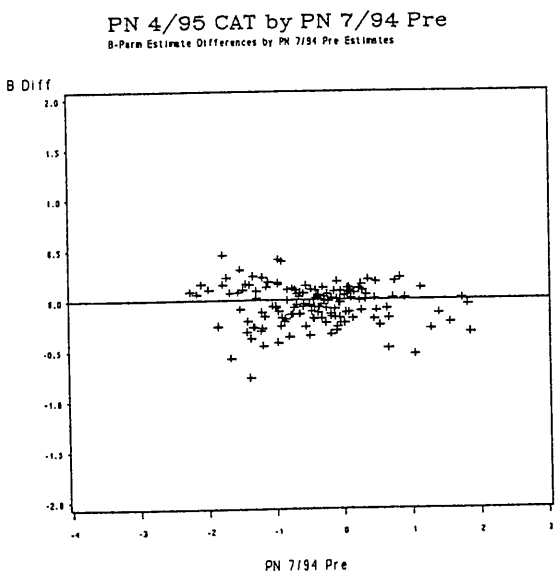


Variable	N	Minimum	Maximum	Mean	Std Dev
DIF	108	-0.5528920	0.5579120	-0.0463635	0.1856034
P0494PR	108	-1.8579640	1.4891000	-0.5119767	0.7356270
P0495CT	108	-2.1484930	1.6843340	-0.5583402	0.7477434

Correlations	DIF	P0494PR
P0494PR	-0.06033	
P0495CT	0.18886	0.96882*

Figure 16

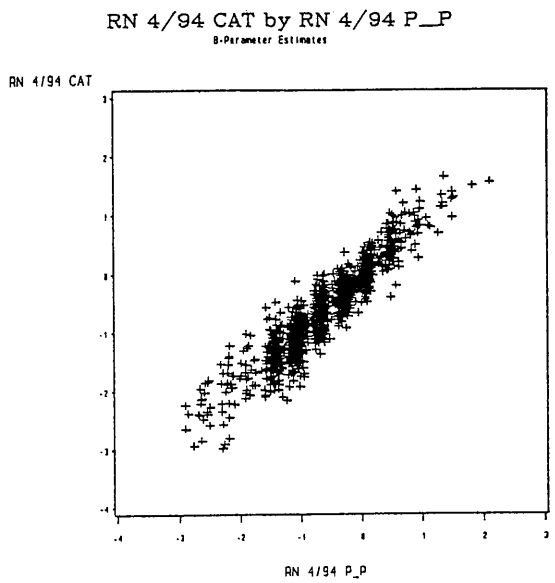


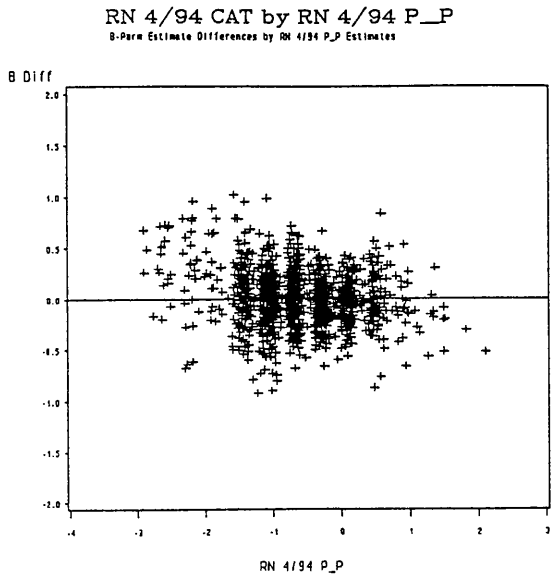


Variable	N	Minimum	Maximum	Mean	Std Dev
DIF	152	-0.7685270	0.4498130	-0.0487554	0.1915367
P0794PR	152	-2.2617190	1.8731020	-0.4161479	0.8103148
P0495CT	152	-2.2403340	1.7840620	-0.4649033	0.8154224

Correlations	DIF	P0794PR
P0794PR	-0.09144	
P0495CT	0.14403	0.97226*

Figure 17

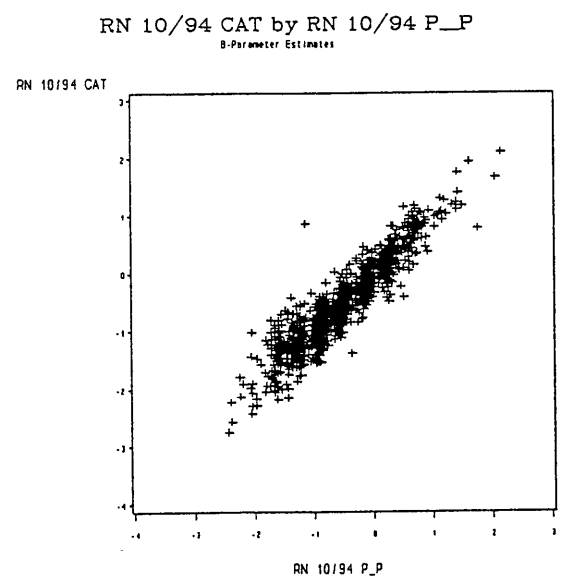


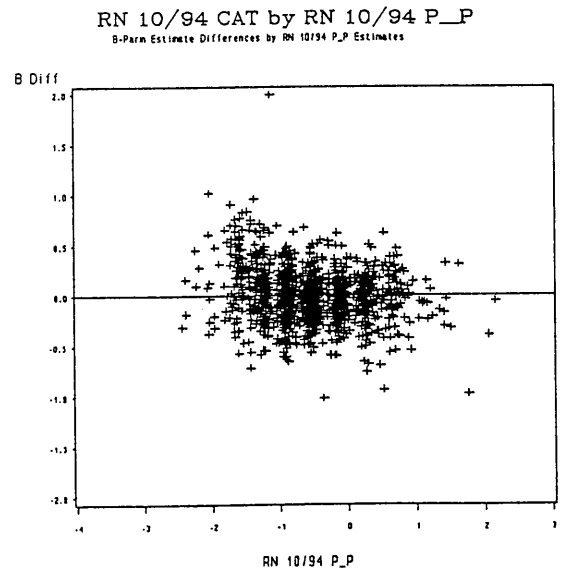


Variable	N	Minimum	Maximum	Mean	Std Dev
DIF	1289	-0.9265100	1.0150160	0.0000000	0.2766823
R0494PC	1289	-2.9061000	2.0997000	-0.6366558	0.7356966
R0494CT	1289	-2.9760330	1.6644180	-0.6366563	0.7356966

Correlations	DIF	R0494PC
R0494PC	-0.18804*	
R0494CT	0.18804*	0.92928*

Figure 18

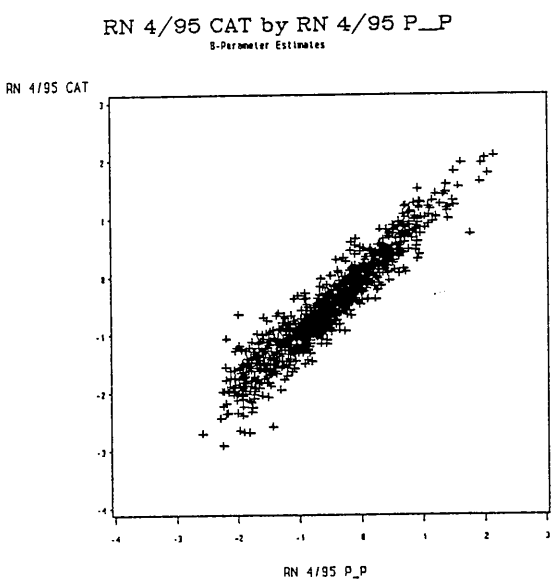


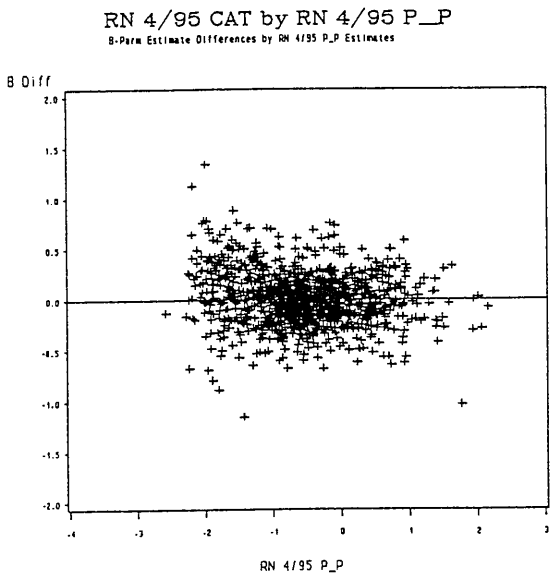


Variable	N	Minimum	Maximum	Mean	Std Dev
DIF	1287	-1.0206900	1.9935830	0.0000000	0.2704289
R1094PC	1287	-2.4365000	2.1558000	-0.5352195	0.7040024
R1094CT	1287	-2.7493340	2.0963540	-0.5352195	0.7040025

Correlations	DIF	R1094PC
R1094PC	-0.19207*	
R1094CT	0.19207*	0.92622*

Figure 19

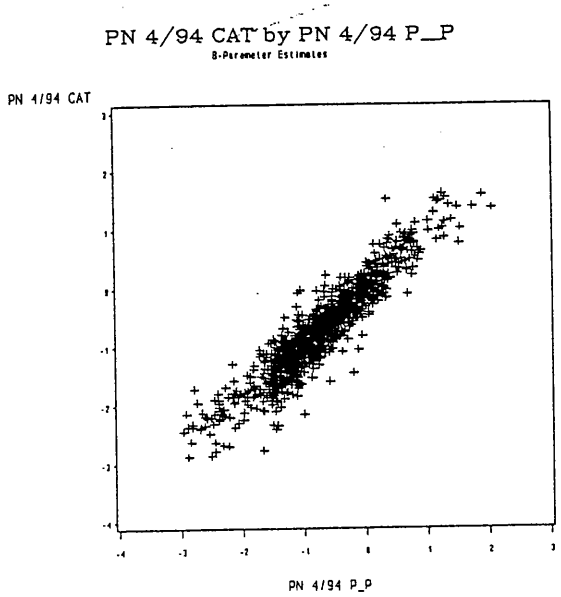


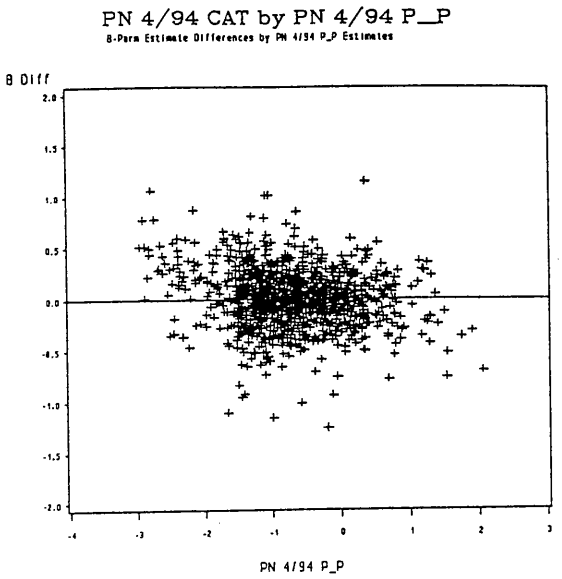


Variable	N	Minimum	Maximum	Mean	Std Dev
DIF	1229	-1.1509060	1.3343070	0.0000000	0.2586107
R0495PC	1229	-2.5778000	2.1558000	-0.5525522	0.7852855
R0495CT	1229	-2.9117010	2.0747460	-0.5525522	0.7852855

Correlations	DIF	R0495PC
R0495PC	-0.16466*	
R0495CT	0.16466*	0.94577*

Figure 20

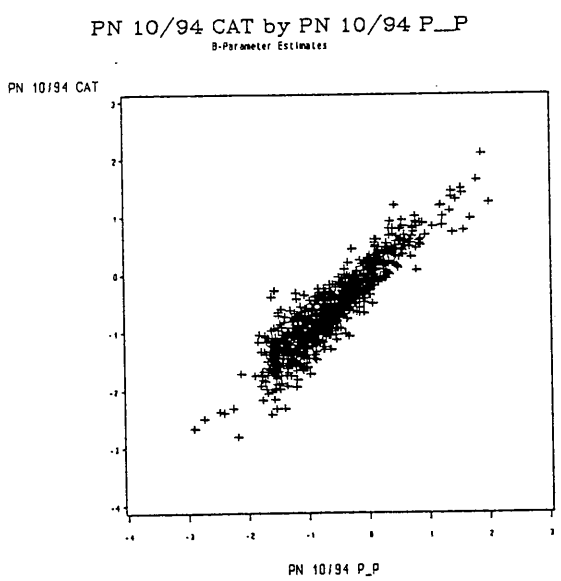


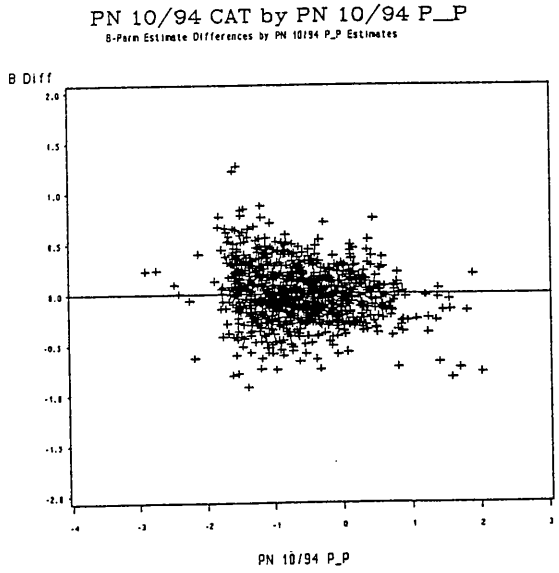


Variable	N	Minimum	Maximum	Mean	Std Dev
DIF	1129	-1.2621000	1.1427590	0.0000000	0.2881337
P0494PC	1129	-2.9545000	2.0622000	-0.7213960	0.7835357
P0494CT	1129	-2.8672290	1.6077470	-0.7213961	0.7835357

Correlations	DIF	P0494PC
P0494PC	-0.18387*	
P0494CT	0.18387*	0.93239*

Figure 21

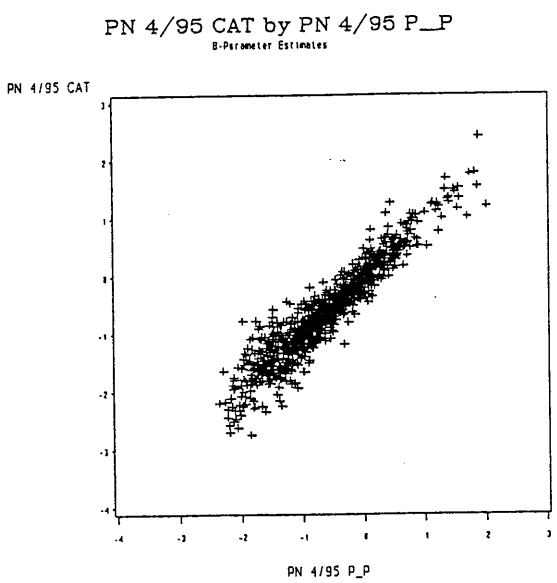


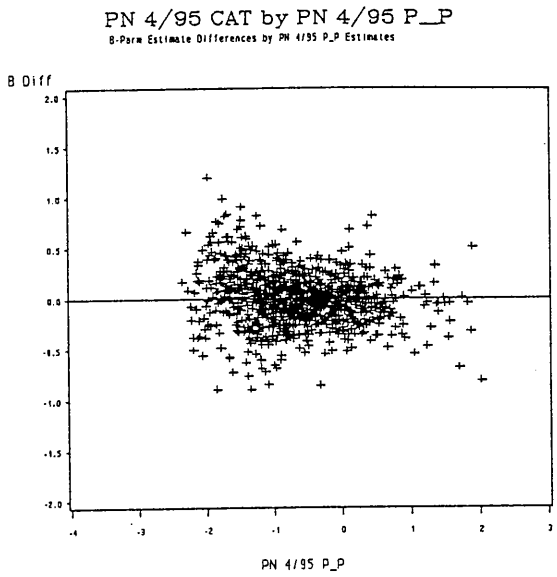


Variable	N	Minimum	Maximum	Mean	Std Dev
DIF	1029	-0.9254400	1.2730080	0.0000000	0.2760665
P1094PC	1029	-2.8988000	2.0209000	-0.6223000	0.7007419
P1094CT	1029	-2.8101970	2.0930570	-0.6223000	0.7007418

Correlations	DIF	P1094PC
P1094PC	-0.19698*	
P1094CT	0.19698*	0.92240*

Figure 22

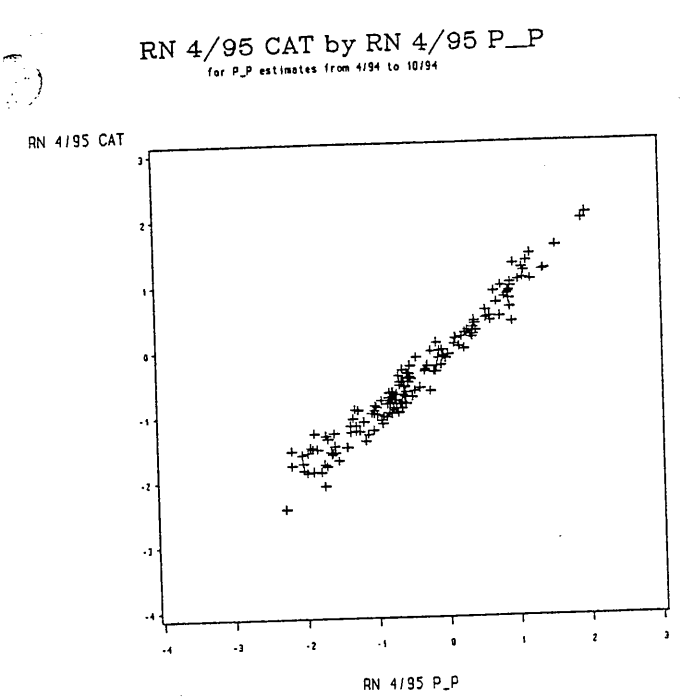




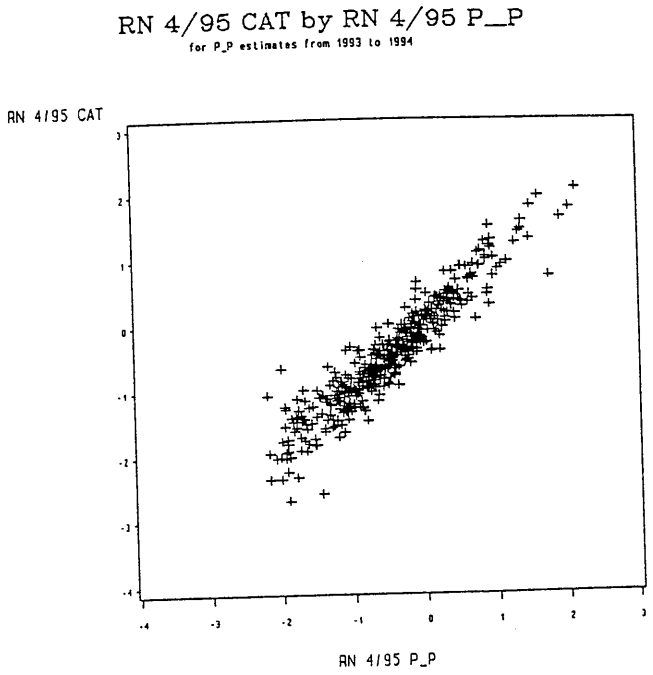
Variable	N	Minimum	Maximum	Mean	Std Dev
DIF	1045	-0.8971570	1.1981650	0.0000000	0.2626538
P0495PC	1045	-2.3577000	2.0209000	-0.6453130	0.7723611
P0495CT	1045	-2.7453310	2.4035750	-0.6453131	0.7723611

Correlations	DIF	P0495PC
P0495PC	-0.17003*	
P0495CT	0.17003*	0.94218*

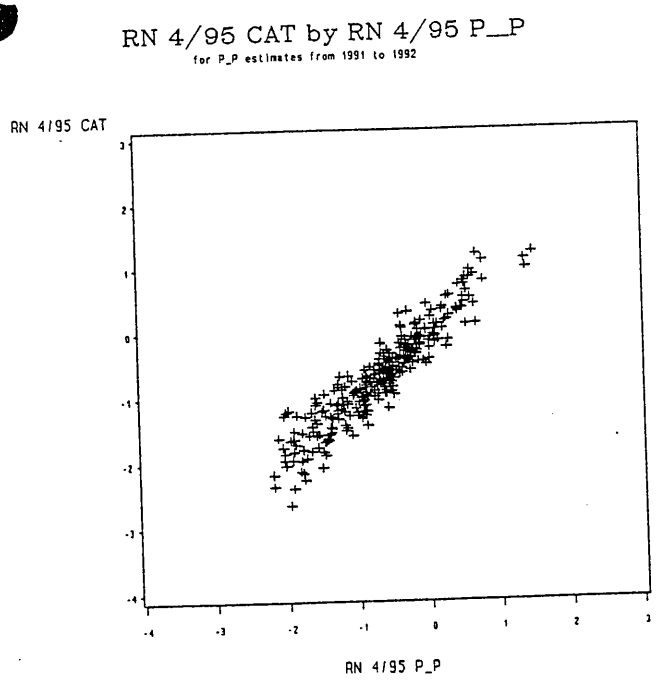
Figure 23



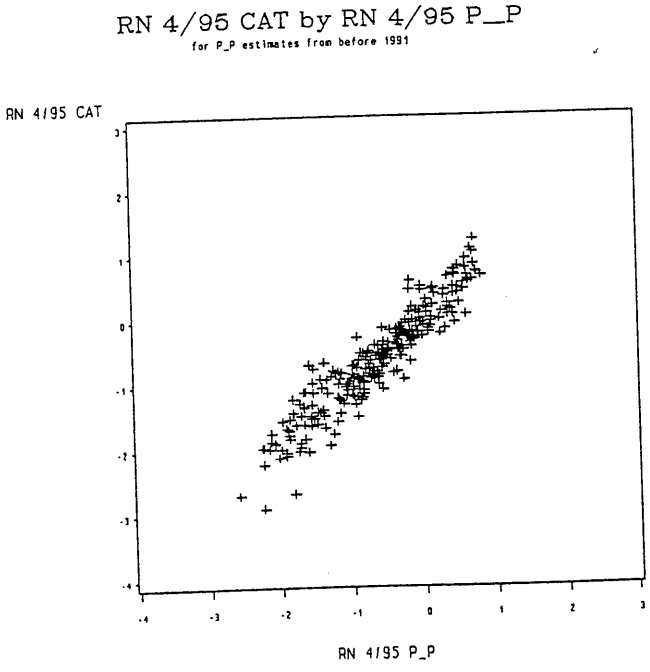
r= .980



r= .940

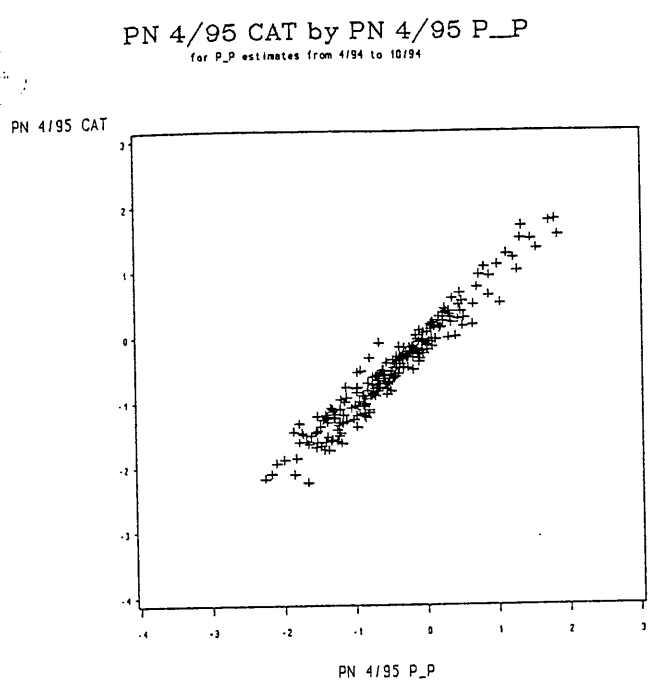


r= .935

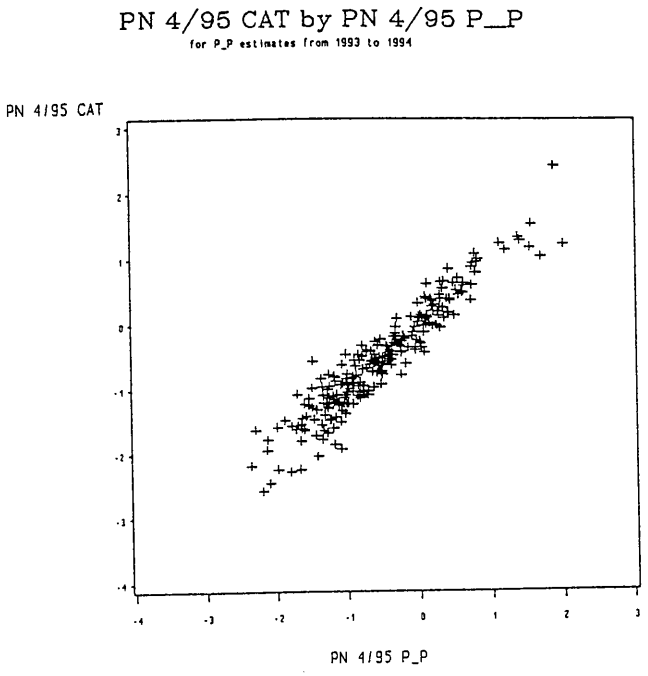


r= .936

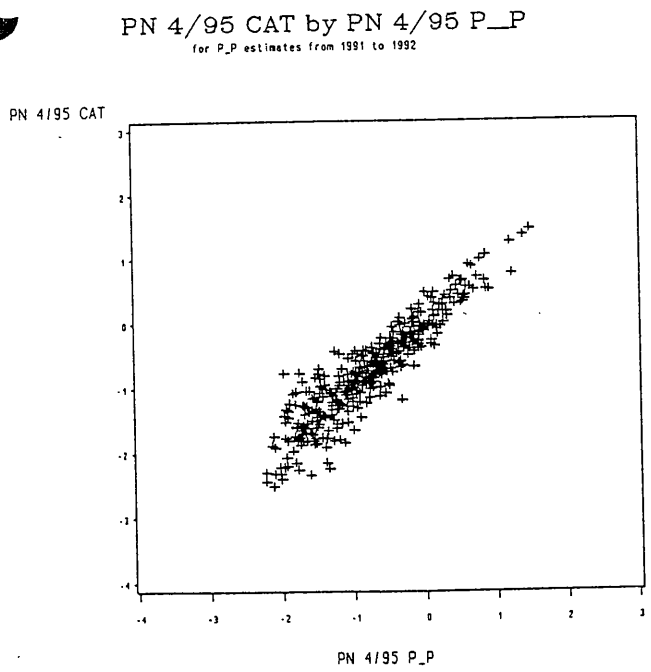
Figure 24



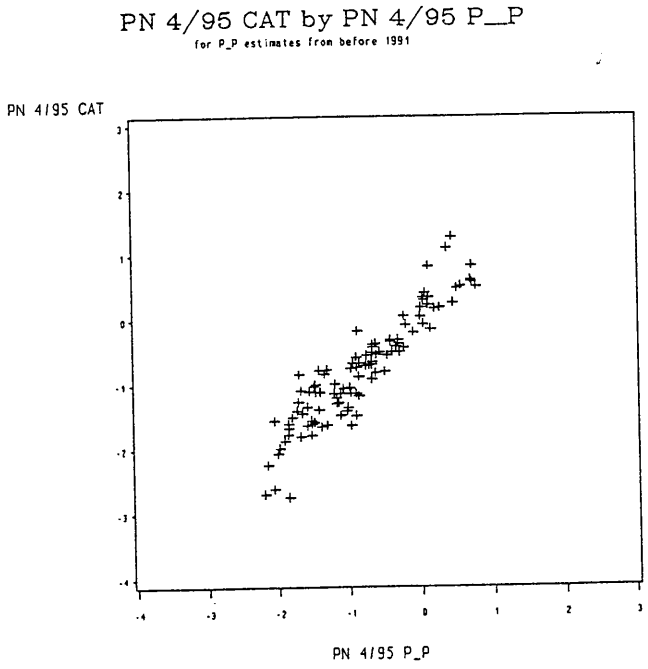
r= .972



r= .947



r= .924



r= .924