ED 359 239 TM 020 008

AUTHOR Sykes, Robert C.; Ito, Kyoko
TITLE Item Parameter Drift in IRT-Based Licensure Examinations.
PUB DATE Apr 93
NOTE 52p.; 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/PC03 Plus Postage.
DESCRIPTORS Analysis of Covariance; *Cutting Scores; *Health Occupations; Item Banks; Item Response Theory; *Licensing Examinations (Professions); Models; *Test Items
IDENTIFIERS *Item Parameter Drift

ABSTRACT

The purpose of this study was to investigate whether systematic, non-zero differences between pairs of item bank b-va ues have occurred in the recent history of two licensure examinations. Licensing examinations were studied for two related health care professions (Program 1 and Program 2). A series of analysis of covariance models was fit to the data in order to investigate the magnitude of changes in item bank b-values and the relationship of any changes to variables indexing factors that have been documented to affect the stability of item parameters. For Program 1, one 300-item form was used, and for Program 2, two 203-item forms were used to obtain a larger number of items. Analyses were first performed to evaluate the relationship, if any, between the dependent variable and the covariates. For both examinations, b-value differences across pairs of item administrations were not influenced by the changing position of the item in different forms. There were systematic changes as a function of time elapsed from a baseline, as discussed in the paper. Item parameter or scale drift was noted for one examination, while bank or pool drift was found for both. Factors that moderated bank drift, and differences between actual form cutscores and cutscores adjusted for documented bank and scale drift ranged from one to five points for the two examinations. Nine tables and five figures present details of the analyses. (SLD)

ITEM PARAMETER DRIFT IN IRT-BASED LICENSURE EXAMINATIONS

Robert C. Sykes
Kyoko Ito

CTB Macmillan/McGraw-Hill

This paper was presented in April, 1993 at the Annual Meeting of the
National Council on Measurement in Education
Atlanta, GA

Introduction

The successful equating of test forms is essential for the validity of test scores produced by testing programs that administer multiple forms. For criterion-referenced tests (CRTs), equating is required to adjust forms for differences in difficulty and subsequently to ensure that performance is evaluated relative to a standard which is fixed over time. A criterion level of performance, often defined on a pool or bank scale, must be translated to a raw cutscore for each administered form in a manner that is not influenced by changes in the ability of candidates taking each form.

For one-parameter Item Response Theory (IRT) or Rasch-based examinations such as the licensure examinations investigated in this paper, equating of test forms may be accomplished by calibrating an anchor set of scored items on a large representative sample and setting the mean of the calibrated item difficulties or b-values equal to the mean bank b-value for the anchor items. Because all of the anchor items in a form have been previously administered at least once before, although not necessarily in the same previous form, bank b-values exist for all of these items. The resulting, post-equated, current bank b-values may then be used to generate a raw-to-IRT(theta)-score correspondence table which determines the number of items that must be answered correctly in order to attain a performance standard on the equated form. This passing standard is, in theory, independent of the set of items constituting the administered form and the group of candidates taking that form.

The realization of a performance standard that can be validly transferred across forms consequently depends upon the difference between a set of bank b-values and a set of calibrated b-values reflecting a difference in levels of a single trait that has substantially determined both the performance of candidates upon whom the items were calibrated (as represented in the set of calibrated item difficulties) and the bank item difficulties. Sets of item difficulties must be represented on a common or single bank scale. The existence of a common bank scale entails that the difficulty of an item, once in the bank, is independent of when it entered the bank, as well as independent of the group of candidates who last took the item and the particular context of items in the form that it was administered. The mean bank b-value used in the equating of a form would therefore not be expected to differ if the items in the form had been administered in a different set of previous examinations.

If the mean bank b-value varies due to the period of time that has elapsed since the items in a form were previously administered, a single bank scale, anchored at a particular point in time, does not exist. Differences within pairs of bank b-values that have been obtained from multiple administrations of pool items will not average zero.

There are two broad types of effects that have been identified as potentially impacting item b-values and hence b-value differences over time. These effects are:

1) changes in examinee ability or component abilities likely induced by changes in curriculum or curricular emphasis, and

2) effects due to the presentation of an item relative to other items in the form.

Form equating constants (i.e., differences in the mean calibrated b-values and mean bank b-values) could be spuriously inflated or attenuated in the presence of one or both of these types of effects.

Scale or item parameter drift, a manifestation of the former effect type, would cause bank b-value differences within a particular content category to increase (or decrease), relative to b-values for items in other content categories, with elapsed time. The mean bank b-value used to equate a form could then differ depending upon the time elapsed since the previous administrations of the items selected for a form from this content category. If previous administrations of these items were staggered over a period of time, as opposed to all at one point in time, the mean bank b-value would capture some average scale drift effect over the period spanned by the previous administrations. Bock, Muraki, and Pfeiffenberger (1988) have documented the presence of drift in item location parameters for items in the College Board Physics Achievement Test over a 10-year period.

Another instance of an effect due to changes in candidate ability was noted by Sykes and Fitzpatrick (1992) for one of the licensure examinations studied in this paper. Their research documented that differences across pairs of Rasch bank b-values for items within a form administered in 1987 were associated with time elapsed from an early period in the testing program. These average non-zero b-value differences over all content categories were believed to be associated with trends in overall candidate performance within the period prior to the administration of the evaluated form and were likely caused by the failure of the equating to completely "detrend" calibrated b-values. The average non-zero b-value differences are a manifestation of a drifting of the complete bank scale (i.e., bank drift).

The potential for b-values to be impacted by the second type of confounding effects, specifically item position within a form, has been documented by studies such as those by Whitely and Dawis (1976) and Yen (1980). Both studies found that context effects due to item position increased or decreased item parameter estimates.

The purpose of the present study was to investigate whether systematic, non-zero differences between pairs of item bank b-values have occurred in the recent history of two licensure examinations. The documentation of non-zero differences would then require establishing the cause and magnitude of these differences. This would allow an evaluation of what effect, if any, these non-zero differences had on past exam cutscores and the proportion of candidates determined to have passed when these cutscores were utilized.

Method

Licensing examinations were studied for two related health care professions. In this paper, the two examinations are referred to as "Program 1 Examination" and "Program 2 Examination."

Forms produced for each of the two programs conformed to the test plan specifications developed from two separate job analyses. Items in both examinations covered two content domains (referred to as "Content Domain 1" and "Content Domain 2"). For the Program 1 Examination, which contains 300 scored items in four booklets, Content Domain 1 consisted of five content areas, and Content Domain 2 of four categories. For the Program 2 Examination normally containing 204 scored items in two booklets, each of the two content domains contained four content areas. Both examinations contain tryout items and are administered more than once a year.

In this study, a series of Analysis of Covariance (ANCOVA) models was fit to the data in order to investigate the magnitude of changes in item bank b-values and the relationship of any changes to variables indexing factors that have been documented to affect the stability of item parameters. The ANCOVA methodology was particularly advantageous for this application in that it permitted the incorporation of different kinds of variables or factors, both quantitative and qualitative. This allowed the study of the two types of effects, mentioned above, that could affect b-value differences: item parameter or bank drift, and context effects. Because these two types of effects constitute all presently known causes of systematic item parameter change and both factors were represented by one or more indices in this study, the fitting of ANCOVA models provided evidence that any significant terms remaining in the model substantively explained the nature of non-zero b-value differences.

The items used were scored items that were administered in a 1991 form or forms and found still usable after the administration(s). For the Program 1 Examination, one 300-item 1991 form was used. For the Program 2 Examination, two 203-item 1991 forms were employed to obtain a larger number of items. In all those evaluated forms, every scored item had previously been administered, either as a tryout or scored item. As with all forms administered in both programs, all scored items served as anchors. All scored items in each program are screened for fit to the Rasch model prior to their assignment to a form.

Each of the studied items had appeared not only in the 1991 form(s) but also in at least one other form administered between 1983 and 1991 for the Program 1 Examination, and between 1982 and 1990 for the Program 2 Examination. For each of the items, sets of bank statistics were obtained for every past scored administration of the item, including the 1991 form(s). Because each item under investigation had been administered at least once prior to its usage in the 1991 form(s), each item had at least two sets of bank item statistics associated with it. The item statistics that constituted each set of bank statistics were:

1) b-value
2) book position (i.e., an item's position within a test booklet)
3) test position (i.e., an item's position within a test form)
4) date of form administration, and
5) content classification codes for the 1991 administration(s).

For each item, sets of item statistics were arranged in reverse chronological order, starting with the statistics for the 1991 form(s) through consecutively earlier administrations. Within each item, sets of bank statistics were then paired up and if the item had an odd number of sets, either the latest or earliest set was deleted. For example, if a usable item in the 1991 form of the Program 1 Examination had previously been administered in the 1990 Form 1 and 1989 Form 1, either the 1991 form set was paired with the 1990 Form 1 set and the 1989 Form 1 set deleted, or the 1990 Form 1 set paired with the 1989 Form 1 set and the 1991 form set deleted. The deletion of item sets was necessary to ensure that data from one administration was not present in more than one pair.

Once all administrations of an item had been paired up, differences were computed for b-values, book position, and test position by subtracting the indices obtained in the earlier administration within each pair from the corresponding indices obtained in the later administration. Thus, if an item had been administered twice, in the 1991 form and 1990 Form 1, a b-value difference (BVDif) was obtained by subtracting the 1990 Form 1 bank b-value from the 1991 form b-value. Similarly, a book position difference (BP1-BP2) and test position difference (TP1-TP2) were created by subtracting the item's 1990 Form 1 book and test position from the item's book and test position in the 1991 form, respectively. These variables, additional variables created from them (described below), and the item's content classification codes and administration date were used to construct one or more indices for each of the three types of variables in the ANCOVA models: dependent, independent, and covariates.

Dependent Variable

B-value differences constituted the dependent, quantitative, variable in the analyses. In all, the Program 1 Examination had 382 pairs of item administrations and consequently 382 independent values of BVDif. There were 487 independent values of BVDif in the analysis of the Program 2 Examination.

Independent Variables

The following three independent qualitative variables were used to classify the analyzed items. The latter two variables were content classifications of the type which may potentially demonstrate item parameter drift over time.

<u>Type of Item Pair (TypePr)</u> Many of the items analyzed contained a tryout administration. Because tryout items are limited, relative to scored items, in the possible positions within a test booklet or test that they may occupy, it is possible that differences between b-values between a tryout and scored item administration pair might differ, on average, from differences between b-values obtained from two scored item administrations. For this reason, each pair of b-values differences was assigned to one of two categories of the TypePr variable: Category 1 if the two b-values were from a tryout and real administration, and category 2 if both b-values were from real administrations. For the Program 1 Examination, there were initially 145 Category 1 b-value differences and 237 Category 2 differences. For the Program 2 Examination, there were 203 Category 1 b-value differences and 281 Category 2 differences after several items were deleted. The variable was included even though Sykes and Fitzpatrick (1992) found no difference in mean b-value differences across the categories of TypePr in their previous research on an earlier administrations of a form from the Program 1 Examination.

Content Domain 1. Each of the analyzed item pairs was classified into one of the Content Domain 1 categories, depending on the item's content coding for the 1991 form(s). For the Program 1 Examination, there were initially 64, 102, 77, 84 and 55 items in the five respective categories. For the Program 2 Examination, there were 164, 86, 148, and 86 items in the four categories, respectively.

Content Domain 2. Each of the item pairs was also classified into one of the four Content Domain 2 areas, again according to the item's coding for the 1991 form(s). There were initially 94, 179, 39, and 70 items, respectively, in the four Content Domain 2 categories of the Program 1 Examination. For the Program 2 Examination, there were 110, 244, 39, and 91 items in the four categories after item deletions.

Covariates

There were two types of covariates. The first type of these quantitative variables consisted of indices of elapsed time that would be expected to demonstrate an association with b-value differences if the average difficulty of at least some items in the pool was changing over time. The second type of covariate consisted of indices of item position. These covariates would be associated with b-value differences if there were relationships between the position of an item in a test, whether tryout or scored, and b-value differences.

Elapsed Time Covariates

TimFr181 or TimFr182. This variable measured the number of months elapsed from the time of administration of the earlier administration in each item pair, and a reference administration. The reference administration was 1981 Form 1 (181) for the Program 1 Examination and 1982 Form 1 (182) for the Program 2 Examination.

A significant positive linear association between the TimFr181 variable and BVDif was found in Sykes and Fitzpatrick's (1992) earlier work on the Program 1 Examination. Since that research utilized items from a 1987 form (as opposed to the 1991 form studied here) and 48 additional months had elapsed between the 1987 form and the 1991 form, the possibility of nonlinear associations between BVDif and "Time Elapsed From 181 or 182" was evaluated by testing quadratic (TimFr181^2 or TimFr182^2), cubic (TimFr181^3 or TimFr182^3) and quartic effects (TimFr181^4 or TimFr182^4).

TimBtAd. This variable measured the number of months elapsed between the earlier and later administrations in each administration pair. As with TimFr181 and TimFr182, the possibility of higher order, nonlinear effects of "Time Between Administrations" was evaluated by additionally testing quadratic through quartic terms (TimBtAd^2, TimBtAd^3, and TimBtAd^4).

Item Position Covariates

BP1-BP2. For each item pair, BP1-BP2 allowed an assessment of whether changes in an item's position in a test booklet was linearly associated with changes in b-values. Quadratic (BP1-BP2^2)[1] and cubic (BP1-BP2^3) "Difference in Book Position" effects were also assessed.

TP1-TP2. Linear (TP1-TP2), Quadratic (TP1-TP2^2), and cubic (TP1-TP2^3) associations between differences in test position and BVDif were evaluated.

Analyses and Results

The fitting of ANCOVA models was a two step process. First, analyses were performed to evaluate the relationship, if any, between the dependent variable (BVDif) and the covariates. These analyses, called the analysis of regression, sought to identify covariates that were significantly related to the dependent variable. In addition to the important function of defining causes of changes in b-value differences, the identification of a set of significant covariates in the analysis of regression

[1] The superscript here and in other item position indices applies to the difference and not the second value in the difference.

permitted error to be reduced in the following phase of testing the associations between BVDif and the independent variables. The reduction of error allows more powerful tests of the independent variables in this second phase, called the analysis of covariance.

Appraisal of the Relations Between Dependent Variable and Covariates

Program 1 Examination: The 382 pairs of b-value differences were plotted against the lowest order covariates (TimFr181, TimBtAd, TP1-TP2, and BP1-BP2) and scanned for outliers. The plots could not definitively rule out nonlinear associations between these four covariates and BVDif. Five observations were deleted as outliers on the elapsed time variables or TP1-TP2.

Means, standard deviations, and intercorrelations between the dependent variable and the four lowest order covariates are given in Table 1-1 for the Program 1 Examination. The mean difference in test position for the 377 administration pairs shows that, on average, an item's position shifted -11.97 positions over consecutive administrations; that is, the item's position shifted toward the beginning of the test in later administrations. A similar effect is noted for booklet position (mean = -17.78). An insignificant difference of -.01 between pairs of b-values is noted across all item content categories.

With respect to the correlations between the dependent variable and covariates, Table 1-1 indicates small but significant correlations (p < .05) between TP1-TP2 and TimFr181, TimBtAd, and BP1-BP2 (r = .23, r = -.21, and r = .24, respectively). The significant correlation between TP1-TP2 and BP1-BP2 is an artifact. When tryout items become scored items, both BP1-BP2 and TP1-TP2 often move in the same direction. The significant correlation between TimBtAd and TimFr181 (r = -.40) is also an artifact. As TimFr181 increases, TimBtAd must decrease.

Program 2 Examination: A total of 487 pairs of b-value differences were plotted against the lowest order covariates (TimFr182, TimBtAd, TP1-TP2, and BP1-BP2) and scanned for outliers. The plots could not definitively rule out nonlinear associations between these four covariates and BVDif. They also suggested the removal of two outliers. One additional pair was deleted because of a coding error, leaving 484 pairs of observations.

Means, standard deviations, and intercorrelations between the dependent variable and the four lowest order covariates are shown in Table 1-2. The mean difference (-28.01) in test position (TP1-TP2) for the 484 administration pairs shows that as with the Program 1 Examination, an item's position in the Program 2 Examination shifted toward the beginning of the test in later administrations. Similarly, the difference in booklet position (BP1-BP2) has a negative mean, -25.05, indicating an item's movement toward the beginning of a book. Once again, b-value differences, on average, did not differ from 0.

With respect to the correlations between the dependent variable and covariates, Table 1-2 indicates a similar significant, artifactual correlation (p < .05) between TP1-TP2 and BP1-BP2 (r = .28) as found in the Program 1 Examination. The significant correlation between BP1-BP2 and TimBtAd (r = -.32) may be attributed to the fact that large values of BP1-BP2 are associated with small values of TimBtAd. This phenomenon is due to the shifts in booklet position which come about when tryouts become reals for the first time and the fact that the time between a tryout and the first scored administration is often as short as 12 months. Finally, as mentioned for the Program 1 Examination, the significant correlation between the "Time Elapsed From 182 (TimFr182)" and "Time Between Administrations (TimBtAd)" indices (r = -.53) is an artifact.

Analysis of Regression

The analysis of regression assesses the association between the covariates and the dependent variable after the influence of all design effects (or independent variables) on the dependent variable has been removed. The association between each covariate and the dependent variable is then evaluated, controlling for the effect of any covariate which preceded the covariate of interest in the model. Because of the order-dependent nature of the tests of the covariates, an ordering of covariates on a priori basis is desirable. To the degree that the ordering of the covariates can be prespecified, the number of reorderings of the covariates, necessary for tests of the effect of each covariate independent of other covariates, can be minimized. Both the nature of the covariates and results from previous research provided an a priori basis for an ordering of the covariates.

Of the covariates related to elapsed time, the "Time Between Administrations" variables were expected to be less significantly associated with b-value differences because they tended to measure shorter periods of time than TimFr181 (or TimFr182). In fact, only TimFr181, out of the four lowest-order covariates studied by Sykes and Fitzpatrick (1992), was significantly related to b-value differences.

Of the covariates related to item position effects, TP1-TP2 was believed to be a less sensitive measure of fatigue or motivational effects than BP1-BP2 because candidates were given breaks between booklet administrations. This would reduce the likelihood that item position in the total test would be related to BVDif due to these effects.

Consequently, a plausible a priori ordering of the lowest-order covariates would have TimFr181 (or TimFr182) entered into the model first, followed by BP1-BP2, TimBtAd, and TP1-TP2. The higher order terms would then enter the model, lower through higher order terms for each of the four lowest-order covariates. This ordering is presented as the first ordering in Table 2-1 and Table 2-2.

Program 1 Examination: The covariates are evaluated from the bottom of the ordering up. Because of the large number of terms evaluated and the several required reorderings of terms, a significance level of $p \leq .01$ was set as a criterion for inclusion in the model. A $p \leq .01$ criterion would reduce the chance that a covariate was judged to be significantly associated to b-value differences by chance alone.

Starting from the bottom of the first ordering, neither quartic ($TimBtAd^4$) nor cubic ($TimBtAd^3$) "Time Between Administrations" was significantly associated with BVDif ($p = .41$ and $p = .81$, respectively). Differences in b-values were significantly ($p = .01$) associated with squared "Time Between Administrations" ($TimBtAd^2$), however.

In order to evaluate any additional contribution of the book position, test position, and "Time Elapsed From 181" variables to explaining variance in BVDif, these 12 covariates were reordered after the quadratic and accompanying linear TimBtAd terms. This second ordering, also presented in Table 2-1, indicated that the higher order book position terms ($BP1-BP2^3$, $BP1-BP2^2$) and the quartic TimFr181 term ($TimFr181^4$) were not significant. Cubic differences in "Time Elapsed From 181" ($TimFr181^3$) was significant at the criterion $p \leq .01$ level, however ($p = .00$). The remaining, linear, difference in book position index (BP1-BP2) could be eliminated because the variable did not explain significant variation after TimFr181 was in the model (reading down from the top).

Because of the higher-order (i.e. above linear) elapsed time effects that significantly explained BVDif variation, the possibility existed that interactions between TimFr181 and TimBtAd were present. Consequently, five new covariates were created in order to evaluate this possibility. These additional terms were:

(1) linear "Time from 181" by linear "Time between Administrations" (TmxTb)

(2) quadratic "Time from 181" by linear "Time between Administrations" (Tm^2xTb)

(3) linear "Time from 181" by quadratic "Time between Administrations" ($TmxTb^2$)

(4) quadratic "Time from 181" by quadratic "Time between Administrations" (Tm^2xTb^2), and

(5) cubic "Time from 181" by quadratic "Time between Administrations" (Tm^3xTb^2).

When individually evaluated in the reverse order, after the significant covariates were in the model, the third ordering in Table 2-1 indicated that none of the five additional covariates were significant at the criterion level (all p's $\geq$ .13).

An assumption of the ANCOVA model is that the regression of the dependent variable on each of the covariates used in the model has the same slope within each of the cells created by the design effects or independent variables. The large number of covariates prevented an initial evaluation of parallelism of slopes. Parallelism was evaluated for the combined final five covariates, however, and could not be rejected ($F_{130,207} = .64$, $p > .90$). There was no sign that the slopes of the regression of BVDif on each covariate were not parallel across cells created by the design effects.

Program 2 Examination: As shown in the first ordering of Table 2-2 (going up from the bottom), quartic ($TP1-TP2^4$), cubic ($TP1-TP2^3$), and quadratic ($TP1-TP2^2$) differences in test position were not significantly associated with BVDif (all p's $\geq .48$). As with the Program 1 Examination, TP1-TP2 did not explain significant variation in BVDif after the three other lowest order covariates had been entered in the model ($p = .20$). Hence differences in test position could not explain b-value differences.

Quartic through quadratic "Time Between Administrations" ($TimBtAd^4$ through $TimBtAd^2$) were similarly nonsignificant (all p's $\geq .81$). An insignificant ($p = .90$) linear TimBtAd term, when read from top down, indicated no effect of time between administrations on b-value differences.

The three higher-order "Difference in Book Position" variables were not significant ($BP1-BP2^4$ through $BP1-BP2^2$, all p's $\geq .18$) and BP1-BP2 was also not significant reading down the list ($p = .55$). Quartic ($TimFr182^4$) and cubic ($TimFr182^3$) time elapsed from 182 variables were not significant ($p = .45$ and $p = .62$) although the quadratic term ($TimFr182^2$) was, necessitating the inclusion of this term and the TimFr182 term in the model.

Once again, the large size of the initial model prevented an evaluation of parallelism with all the original covariates. Parallelism was evaluated when the set of covariates had been reduced to the four linear covariates and $\mathrm{TimFr182}^2$. Parallelism of slopes could not be rejected ($F_{130,317} = 1.02$, $p > .44$). There was no sign that the slopes of the regression of BVDif on each of the five evaluated covariates were not parallel across 32 (= 4x4x2) cells created by the independent variables.

Analysis of Covariance

Program 1 Examination: The five covariates, which collectively explained significant variation in BVDif ($F_{5,333} = 6.19$, $p = .00$ in Table 3-1), were entered initially in the model and each of the independent variables evaluated independently of the covariates and other independent variables. An a priori basis existed for the ordering of these variables, presented in Table 3-1. Because prior research and the literature had documented instances of item parameter drift, the content categories were entered prior to TypePr.

Reading up from the bottom of Table 3-1, the three-way interaction term was insignificant ($p = .17$). Reading down from the top, no other effect involving Content Domain 1 was significant ($p = .37$, $p = .22$, and $p = .10$ for the main effect and two two-way interactions, respectively). Consequently, Content Domain 1 was removed from the model and the remaining effects reassessed. Neither the "Content Domain 2 x TypePr" interaction nor the TypePr main effect was significant ($p = .06$ and $p = .19$, respectively). The Content Domain 2 main effect was significant at $p = .01$.

The final fitted ANCOVA model consisted of the five covariates and an effect due to the second content domain. The model explained a small, though significant, amount of variation in BVDif ($R^2 = .091$, $F_{8,368} = 4.59$, $p = .00$). Additional checks on the adequacy of the fitted model were obtained by plotting the model residuals against the predicted (model) differences and checking the normality of the residuals. Figure 1 contains the plot of the residuals and provides no sign that the size of the residuals are associated with the magnitude of the predicted differences. The Shapiro-Wilk statistic, W (SAS, 1985), indicated that the residuals were distributed normally ($W = .979$, $p = .08$).

The coefficients of the final fitted model are presented in Table 4-1. The estimates for the four Content Domain 2 categories, based on simple contrasts, indicate that only the A category coefficient significantly differs from zero ($p < .05$) when compared against its standard error.

Program 2 Examination: The two covariates, which together explained significant variation in BVDif ($F_{2,452} = 3.39$, $p = .03$ in Table 3-2), were entered initially in the model and each of the independent variables evaluated independently of the covariates and other independent variables. As explained earlier in the Program 1 Examination section, these variables were entered into the mcdel on an a priori basis. Reading up from the bottom of Täble 3-2, no main effect or interaction was significant (all p's $\geq$ .23).

The final fitted ANCOVA model consisted of linear and quadratic "Time Elapsed From 182" terms and an intercept. The model explained small though significant variation in BVDif (R^2 = .014). Additional checks on the adequacy of the fitted model were obtained by plotting the model residuals against the predicted (model) differences and checking the normality of the residuals. Figure 2 cuntains the plot of the residuals and provides no sign that the size of the residuals are associated with the magnitude of the predicted differences. The Shapiro-Wilk statistic, W (SAS, 1985), indicated that the residuals were distributed normally (W = .982, p = .20).

The coefficients of the final fitted model are presented in Table 4-2.

Discussion

. Because the effects found for the Program 1 Examination are more complicated in nature and subsume those noted for the Program 2 Examination, the following discussion focuses on the Program 1 Examination.

Program 1 Examination: Although the insignificance of the book and test position covariates indicated that there was no systematic influence of item position on the stability of b-values, an effect due to item parameter drift and a more general bank drift effect were observed. Item parameter drift was represented by b-values changing significantly for only one of the four Content Domain 2 categories. A main Content Domain 2 effect was also noted by Sykes and Fitzpatrick (1992). They found that drift in the A category approached significance at p = .11, while significant (p < .05) item parameter drift was observed for the B category of Domain 2. A bank drift effect was also found in the previous research but that effect was of a simpler nature than the changes documented here.

Predicted B-Value Differences Based on the Fitted Model for Program 1 Examination

In order to better understand the nature of these effects, coefficients of the fitted model were used to predict b-value differences. B-value differences for each of the Content Domain 2 categories were predicted for two different times between administrations, 12 and 41 months, across the span of time from 181. The two times between administrations were chosen because large numbers of data points were available at these times and because between 80% to 100% of the items in a form have time between administrations falling within this range. No items can be readministered in fewer than 12 months and all items are reviewed for currency after four years or 48 months.

The predicted differences are plotted in Figure 3 with the plotting characters A and O connected with either a solid or dotted line. A's denote Content Domain 2 Category A, and O's denote Categories B through D in the same domain. A solid line denotes 12 months between administrations, and a dotted line 41 months between administration.

B-value differences are predicted to be, on average, approximately -.25 to -.10 for those items that had been administered in 1984 Form 1 (TimFr181 = 36 months) and readministered in 1985 Form 1 (TimPtAd = 12 months). For those items that were administered in 1984 Form 1 and readministered 41 months later, in 1987 Form 2, the mean b-value difference was predicted to be approximately between 0.00 and 0.15.

B-value differences for items in the drifting A category of Content Domain 2 (labeled A in Figure 3) are consistently smaller (more negative) than b-value differences in the other content categories (labeled O). This pattern is true for both the items that were readministered 12 months after a prior administration (solid line in the figure) and those readministered 41 months after a prior administration (broken line).

Smaller (more negative) b-value differences are expected for items in a content category subjected to drift. To illustrate this point, suppose a 50-item form is specified to have approximately 30% of its items (i.e., 15 items) from the drifting content category, A. This percentage is similar to the quota set for Category A of the Program 1 Examination. Assume that items in Category A are getting easier for the candidates because of increased curricular emphasis of the content and that consequently, item difficulty on the Rasch logit scale for these items is decreasing .1 logit a year. For the sake of simplicity, assume that the items for a form are selected from a very large item pool, so that no items are selected more than once for a form over a period of, say, five years. Furthermore, suppose no trends in overall ability are occurring over this five-year period.[2]

In the base year (0), prior to the onset of the scale drift, the mean of the calibrated items, $\bar{b}_c$, will equal the mean of the bank b-values, $\bar{b}_b$, producing an equating constant of 0 when the former is subtracted from the latter. In each of the next five years, $\bar{b}_c$ will decrease by .03 logits (15 items times .1 logit divided by 50). By the end of the fifth year, $\bar{b}_c$ will have declined by .15 logits, so that the calibrated b-values before equating will be, on average, .15 logits more difficult than they were in the base year. Items in the drifting content category in the form administered in the fifth year will actually have b-values that are each .35 less than what they were in the base year (.5 logit decrease over five years plus the .15 equating constant). Thus, if the difference in bank b-values was computed for a non-drifting item in the form administered in the fifth year, relative to its bank b-value in the base year, it would be .15 logit higher than the base-year value, while the corresponding b-value for a drifting item would be .35 lower than the base-year b-value.

[2] These latter two assumptions are not realistic for the Program 1 Examination.

A relaxing of the assumption that an item has not been readministered more than once within the five-year period would either modestly increase or decrease these b-value differences for both drifting and non-drifting items under typical Program 1 item reuse procedures. The direction of the effect would depend upon how many of each type of item had been selected for a readministration between the base and fifth year and how early or late within the period of drift that they were readministered.

Nonlinearity of Predicted B-Value Differences for Program 1 Examination

In addition to the constant differential in b-value differences attributed to the drift of the A content category of the second content domain, the predicted b-value differences reveal the effect of the significant cubic trend over time elapsed from 181. This trend may most likely be attributed to changing population levels of candidate performance, as opposed to any change in test development procedures or specifications. The average difficulty of forms administered for the Program 1 (and Program 2) Examinations are constrained to fall within a certain relatively narrow difficulty range, and content category range quotas have been fixed for forms administered in those examinations.

The evidence for changing population[3] levels of candidate ability is of two types. The first type of evidence consists of passing rates and mean candidate ability expressed on the Rasch (theta) scale. Both indices broadly suggest an increase in candidate performance in 1984 and 1985, a decline in performance between 1987 and 1988, followed by an increase in performance that continued through 1991. This performance trend coincides with the pattern of predicted b-value differences indicated in Figure 3 for both the two 12-month TimBtAd groups and the two 41-month TimBtAd groups.

Both passing rates and mean thetas are linked to the theta scale that is believed to be drifting, however. The link for passing rates is indirect, through the raw cutscore set for each form from the theta passing standard.

The second type of evidence for changing ability levels of the population does not utilize the theta scale, although it too is subject to a degree of drift with changing candidate ability. This evidence is the difference between average form p-values computed from a large representative sample of candidates taking the form (post) and the average p-value for the administration of the items previous to the evaluated form administration (pre). These "post - pre" differences in average p-values indicate increases (+) in candidate performance relative to a past

[3] The population referred to is the large group of first-time, U.S.-educated candidates that constitutes the reference population upon which item parameters are calibrated and examination performance statistics compiled.

generalized performance over a number of last previous administrations, or decreases (-) in that performance. In Figure 4, the available average p-value differences have been drawn over the trend lines of Figure 3. The pattern of p-value differences substantiates the trend of declining, followed by stabilized, then increasing candidate performance visible in the right half of Figure 4.

The trend of differences in average p-values verifies that the cubic pattern of predicted b-value differences over time may be attributed to overall bank drift due to changing population levels of ability. The specific manner in which increases or decreases in candidate overall ability impact differences in item bank b-values may be seen by simulating simplified forms administered over time, in a way similar to that used to evaluate the effects of scale drift and testing a program model, as was done to illustrate the effects of scale drift (p. 25). If, for example, the effects of decreasing candidate ability was modeled, pool item difficulties might be assumed to be increasing by a constant amount per unit of time (items are becoming more difficult because candidates are becoming less able). By selecting several sample tests and computing what post-equated bank b-values would be for forms that differed in composition, it becomes evident that the obtained pattern of cubic b-value differences is not a simple function of changes in candidate performance over time.

In addition to changing candidate performance, two factors substantially determine the observed trend in b-value differences. The first factor is the extent to which anchor items (in the case of both the Program 1 and 2 Examinations, all scored items) were administered in more than one common previous administration. It can be readily seen that an overall bank drift, as opposed to an item parameter drift, can not occur under the limiting case of all anchor items being previously administered in the same previous form. Under these circumstances, the equating constant that is obtained by subtracting the mean calibrated b-value from the mean bank b-value contains the entire effect of bank drift on the calibrated b-values over the period spanned between the first and second administrations.

Because of the potential exposure problems that a repeat administration of a whole form of scored items poses, either through the compromise of a form or memory effects, repeat administrations are not deemed viable by many licensure testing programs, including the two assessed here. Licensure programs often choose to minimize these risks and increase the diversity of sub-test plan content by constructing new forms from items previously administered in a number of previous forms, as well as the set of successfully tried-out items. Hence, forms may contain items from a number of previous administrations. This results in a set of previous administrations that are dispersed over time. The dispersion of last previous administration dates over time is the second factor that impacts b-value differences in bank b-values.

By varying the distribution of elapsed time between the previous administrations and the modeled current administration, it can be seen that declining (or increasing) candidate ability, mediated through a constant drift of bank items becoming increasingly difficult (or easy) over time, will induce a decline (or increase) in bank b-value differences. The decrease in b-value differences that occur with declining ability is substantially brought about through the presence of newer items in a form. Bank b-values for the newer items decrease more than those for older items, that have not been previously administered for a length of time, because of the effect of increasingly negative form equating constants caused by the increase in the mean b-value of the calibrated items. Thus, after equating, the difference between the diminished current bank b-value for a newer item and its previous bank b-value will be smaller than that for an older item.

The equated bank b-value for an item that has not been readministered since the onset of the decline in candidate ability will actually increase due to the accumulation of drift over a number of years that is not eliminated by the equating constant. However, since the number of these items in any form decreases as the declining trend in candidate ability continues,[4]

[4] Another factor that would often limit the number of these items in a form is currency restrictions.

the increase in bank b-value differences which they contribute is increasingly overwhelmed by the decreasing b-value differences being produced by the newer items.

The presence of newer items is also manifest in the quadratic effect of TimBtAd on b-value differences, visible in Figure 3. As the length of time between administrations increases for any period of time elapsed from the earlier bank b-value and 181, b-value differences between bank b-values increase. Conversely, b-value differences decline as TimBtAd diminishes. Those newer items that have been administered on only a few previous occasions contribute more negative b-value differences because of the greater reducing effect of form equating constants on their equated b-values.

The effect of newer items is substantial for the Program 1 (and Program 2) Examinations because approximately 50% of the items in each form consist of successful tryout items. Moreover, the importance of assessing current content results in the selection of relatively few items that have not been administered within three years of their prior selection of a form. The effect of a distribution of elapsed time between administrations that is less skewed than the distribution for the Program 1 Examination would probably be to moderate the magnitude of b-value differences. Additional investigation of such distributions is required. Results from these additional studies may suggest that item reuse policies (that stipulate the length of time that must elapse between administrations) may impact the degree that trends in candidate ability are incorporated in item pools.

Predicted Effect of B-Value Differences on Program 1 Forms

The approach of mean b-value differences to their expected value of 0 that is evident in Figure 3 is a desirable outcome in terms of the validity of test scores. Mean differences that are near zero will result in small differences between the cutscores used for more recent Program 1 Examination forms and those cutscores that would have been produced if there was no scale or bank drift. The fitted model for the Program 1 Examination was used to obtain cutscores adjusted for scale and bank drift in the following manner.

Adjusted b-values were derived by using the ANCOVA model to predict b-value differences that were observed (on average) for the items administered in each of four recent forms, using each item's previous administration date. B-value difference predictions were then compared to a benchmark b-value difference. The benchmark difference was arbitrarily chosen as that for an item administered in 1990 Form 2 and then again in 1991 Form 2. The difference between the benchmark and predicted b-value difference was then added to the bank b-value for each item in each of the four forms. Table 5-1 contains mean bank b-values for these four Program 1 Examination forms as well as the cutscores and passing rates for populations of first-time U.S. educated.

As can be seen in Table 5-1, mean adjusted b-values differ from actual mean b-values by a maximum of .09 logits (1990 Form 2). Adjusted cutscores differ by as much as five points (1990 Form 2) from form cutscores. The largest difference in the percentage of first-time U.S. educated that would have passed in the absence of both types of drift versus the percentage that actually passed was 2.7%, again for the 1990 Form 2.

Program 2 Examination: Unlike the Program 1 Examination, there was no sign of item parameter drift over any category in either of the two content domains. However, the Program 2 Examination also demonstrated an effect due to an overall bank drift. The bank drift for the Program 2 Examination was of a less complicated nature than the effect found for the Program 1 Examination. The absence of an effect due to the amount of time between administrations (TimBtAd) may perhaps be due to differences in the distribution of elapsed time between administrations for items in a Program 2 form relative to a Program 1 form. The distribution of elapsed time between administrations for items in a form is not necessarily similar to the distribution of TimBtAd across pairs of bank b-values, because of the potential for changes in item reuse policies for a program during the evaluated time. These latter elapsed times (TimBtAd) are presented in Tables 1-1 and 1-2 to be similar across the programs.

B-value differences were again predicted using the coefficients of the fitted model. The predicted differences are plotted in Figure 5.

The predicted values in Figure 5 depict b-value differences that average less than -.12 beginning at 0 months elapsed from 1982 Form 1 and steadily increase to a peak of slightly over .02 at approximately 65 months after 1982 Form 2. After 72 months, predicted b-value differences begin to fall.

The quadratic trend depicted for b-value differences in Figure 5 deviated from the expected value of 0 in a manner that reflected changes in ability levels of the Program 2 Examination candidate population during this period. These predicted b-value differences aligned with a pattern of post-pre average p-values that has not been shown.

Predicted Effect of B-Value Differences on Program 2 Forms

The items in four Program 2 Examination forms were adjusted for the bank drift in order to evaluate whether the recent trend of declining b-value differences had resulted in forms that had a different cutscore than what they would have had in the absence of the bank drift. Using the same procedure applied to the four evaluated Program 1 Examination forms, adjusted b-values were derived by using the fitted Program 2 ANCOVA model to predict a b-value difference for each item in each of four forms. This b-value difference had been observed, on average, for items administered in that form and previously administered in the same prior form. Each b-value difference was then compared to a benchmark b-value difference. The benchmark difference was for an item administered in 1990 Form 2 and then again in 1991 Form 2. The difference between the benchmark and predicted b-value difference was then added to the bank b-value for each item in each of the four forms.

As can be seen in Table 5-2, mean adjusted b-values differ from actual mean b-values by no more than .05 logits (1990 Form 1) with the average difference declining through 1991 Form 2, with its difference of .02. The adjusted cutscores for each of the four forms was one point above the actual cutscore, indicating that the items (and the administered forms constituted by them) have been made easier. This adjustment would have resulted in slightly lower passing rates if each form had been adjusted for bank and/or scale drift.

Conclusions

(1) For both examinations investigated (i.e., Program 1 and Program 2), b-value differences across pairs of item administrations were not influenced by the changing position of the item in different forms.

(2) Although b-value differences averaged approximately zero across all pairs of paired administrations for both programs, they changed systematically as a function of time elapsed from a baseline year and time between administrations for the Program 1 Examination, and as a function of time elapsed from a baseline year for the

Program 2 Examination.

(3) Item parameter or scale drift over one Content Domain 2 category was noted for the Program 1 Examination. No parameter drift was observed in the Program 2 Examination over the categories of either of the two content domains.

(4) Bank or pool drift was noted for both examinations.

(5) The magnitude of bank drift may be moderated, or accentuated, by item reuse policies that determine distributions of elapsed time between administrations.

(6) For one of the two programs (i.e., Program 1), predicted mean b-value differences approached zero near the end of the evaluated period of time. Differences that average zero are expected in the absence of bank drift. For the other program (i.e., Program 2), predicted b-value differences increased from negative to positive then declined to below zero.

(7) Differences between actual form cutscores and cutscores adjusted for documented bank and scale drift ranged from one to five points for the two examinations investigated.

References

Bock, R. D., Muraki, E., & Pfeiffenberger, W. (1988) Item pool maintenance in the presence of item parameter drift. Journal of Educational Measurement, 25, 275-285.

CTB Macmillan/McGraw-Hill. (1992) RN Scale Drift Study. Monterey, CA: CTB Macmillan/McGraw-Hill.

Eignor, D. R., & Cook, L. L. (1983). An investigation of the feasibility of using item response theory in the pre-equating of aptitude tests. Paper presented at the annual meeting of the American Educational Research Association, Montreal.

Goldstein, H. (1983) Measuring changes in educational attainment over time: Problems and possibilities. Journal of Educational Measurement, 20, 369-377.

SAS Institute Inc. (1985). SAS user's guide: Statistics, 1985 edition. Cary, NC: SAS Institute.

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

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case of testlets. Journal of Educational Measurement, 24, 185-201.

Whitely, S. E., & Dawis, R. V. (1976). The influence of test context on item difficulty. Educational and Psychological Measurement, 36, 329-337.

Winer, B. J. (1971). Statistical Principles in Experimental Design. New York: McGraw-Hill

Wise, L. L., Chia, W. J., & Park, R. K. (1989). Effects of item position on IRT parameter estimates and item statistics. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Yen, W. M. (1980). The extent, causes, and importance of context effects on item parameters for two latent trait models. Journal of Educational Measurement, 17, 297-311.

Table 1-1

Means, Standard Deviations, Standard Errors, and Intercorrelations of Dependent Variable and Covariates: Program 1 Examination (N = 377)

	BVDif	TimFr181	TimBtAd	BP1-BP2	TP1-TP2
Mean	-.01	89.90	20.19	-17.78	-11.97
Standard Dev.	.28	26.14	7.76	27.68	147.43
Standard Error	.01	1.35	0.40	1.43	7.59
BVDif	1.00				
TimFr181	.01	1.00			
TimBtAd	.10	-.40*	1.00		
BP1-BP2	-.08	.04	-.06	1.00	
TP1-TP2	-.08	.23*	-.21*	.24*	1.00

*p <.05

Table 1-2

Means, Standard Deviations, Standard Errors, and Intercorrelations of Dependent Variable and Covariates: Program 2 Examination (N = 484)

	BVDif	TimFr182	TimBtAd	BP1-BP2	TP1-TP2
Mean	-.00	75.63	22.02	-25.05	-28.01
Standard Dev.	.26	26.45	10.98	27.37	92.14
Standard Error	.01	1.20	0.50	1.24	4.19
BVDif	1.00				
TimFr182	.05	1.00			
TimBtAd	-.03	-.53*	1.00		
BP1-BP2	.02	.02	-.32*	1.00	
TP1-TP2	-.06	.05	-.07	.28*	1.00

*p <.05

Table 2-1

Results of Regression of Covariates on BVDif:
Program 1 Examination

First Ordering

Source	df	Type I SS	F Value	Pr>F
Design effects	38	4.272		
Regression	14			
TimFr181	1	.000	0.00	.99
BP1-BP2	1	.146	2.07	.15
TimBtAd	1	.161	2.27	.13
TP1-TP2	1	.047	0.66	.42
TP1-TP2^2	1	.017	0.24	.63
TP1-TP2^3	1	.146	2.06	.15
TimFr181^2	1	.442	6.25	.01**
TimFr181^3	1	.556	7.86	.01**
TimFr181^4	1	.465	6.58	.01**
BP1-BP2^2	1	.000	0.00	1.00
BP1-BP2^3	1	.005	0.07	.80
TimBtAd^2	1	.550	7.78	.01**
TimBtAd^3	1	.004	0.06	.81
TimBtAd^4	1	.048	0.69	.41

Second Ordering

Source	df	Type I SS	F Value	Pr>F
Design effects	38			
Regression	12			
TimFr181	1	.000	0.00	.99
BP1-BP2	1	.146	2.07	.15
TimBtAd	1	.161	2.28	.13
TimBtAd^2	1	.552	7.83	.01**
TP1-TP2	1	.046	0.65	.42
TP1-TP2^2	1	.017	0.24	.62
TP1-TP2^3	1	.129	1.83	.18
TimFr181^2	1	.195	2.77	.10
TimFr181^3	1	1.218	17.29	.00**
TimFr181^4	1	.070	0.99	.32
BP1-BP2^2	1	.000	0.00	.97
BP1-BP2^3	1	.000	0.00	.95

Third Ordering

Source	df	Type I SS	F Value	Pr>F
Design effects	38			
Regression	10			
TimFr181	1	.000	0.00	.99
TimBtAd	1	.213	3.04	.08
TimBtAd^2	1	.544	7.77	.01**
TimFr181^2	1	.100	1.43	.23
TimFr181^3	1	1.311	18.73	.00**
TmxTb	1	.158	2.26	.13
Tm^{2}xTb	1	.098	1.39	.24
TmxTb^2	1	.064	0.91	.34
Tm^{2}xTb^2	1	.028	0.41	.52
Tm^{3}xTb^2	1	.027	0.39	.53
Error	328	22.960		

** $p \le .01$

Table 2-2

Results of Regression of Covariates on BVDif:
Program 2 Examination

First Ordering					Second Ordering				
		⟵Type I	F				⟵Type I	F	
Source	df	SS	Value	Pr>F	Source	df	SS	Value	Pr>F
Design effects	31	1.589			Design effects	31			
Regression	16				Regression	5			
TimFr182	1	.088	1.37	.24	TimFr182	1	.088	1.39	.24
BP1-BP2	1	.023	0.36	.55	TimFr182^2	1	.365	5.76	.02
TimBtAd	1	.001	0.02	.90	BP1-BP2	1	.064	1.01	.32
TP1-TP2	1	.106	1.66	.20	TimBtAd	1	.001	0.01	.91
TimFr182^2	1	.404	6.29	.01**	TP1-TP2	1	.104	1.65	.20
TimFr182^3	1	.016	0.25	.62	Error	447	28.262		
TimFr182^4	1	.036	0.57	.45					
BP1-BP2^2	1	.002	0.03	.87					
BP1-BP2^3	1	.034	0.54	.46					
BP1-BP2^4	1	.114	1.78	.18					
TimBtAd^2	1	.003	0.05	.83					
TimBtAd^3	1	.004	0.06	.81					
TimBtAd^4	1	.003	0.05	.83					
TP1-TP2^2	1	.032	0.50	.48					
TP1-TP2^3	1	.002	0.04	.85					
TP1-TP2^4	1	.013	0.20	.66					

** $p \leq .01$

Table 3-1

Results of Analysis of Covariance:
Program 1 Examination

Source	df	Type I SS	F-Value	Pr > F
Regression	5	1.914	5.46	.00**
Design Domain 1	4	.303	1.08	.37
Domain 2	3	.819	3.89	.01**
Domain 1 * Domain 2	12	1.091	1.30	.22
TypePr	1	.091	1.30	.26
Domain 1 * TypePr	4	.545	1.94	.10
Domain 2 * TypePr	3	.595	2.83	.04
Domain 1 * Domain 2 * TypePr	11	1.083	1.40	.17
Total (corrected)	376	29.775		

** p ≤ .01

Table 3-2

Results of Analysis of Covariance:
Program 2 Examination

Source	df	Type I SS	F-Value	Pr > F
Regression	2	.429	3.39	.03*
Design Domain 1	3	.003	0.02	1.00
Domain 2	3	.120	0.63	.60
Domain 1 * Domain 2	9	.591	1.04	.41
TypePr	1	.093	1.47	.23
Domain 1 * TypePr	3	.218	1.15	.33
Domain 2 * TypePr	3	.015	0.08	.97
Domain 1 * Domain 2 * TypePr	9	.572	1.01	.43
Total (corrected)	483	30.474		

* p ≤ .05

Table 4-1

Coefficients of the Final Fitted Model
Program 1 Examination:

Parameter	Estimate	Standard Error of Estimate
Intercept	-1.350788	.296213
TimBtAd	0.044827	.011067
TimFr181	0.042821	.012139
TimBtAd^2	-0.000702	.000192
TimFr181^2	-0.000696	.000201
TimFr181^3	0.000003	.000001
Content Domain 2		
A	-0.101869	.043278
B	0.005238	.039100
C	0.009531	.055484
D	0.0	--

Table 4-2

Coefficients of the Final Fitted Model
Program 2 Examination:

Parameter	Estimate	Standard Error of Estimate
Intercept	-0.127195	.051713
TimFr182	0.004818	.001867
TimFr182^2	-0.000037	.000015

Table 5-1

Form and Adjusted B-Value Form Statistics:
1990 Form 1 - 1991 Form 2:
Program 1 Examination

	Form			
	1990 Form 1	1990 Form 2	1991 Form 1	1991 Form 2
B-Values				
Mean Form	-1.17	-0.97	-1.01	-1.13
Mean Adjusted	-1.11	-0.88	-1.00	-1.14
(Largest Abs. Difference: B-Value-Adjusted)	-.27	-.25	-.18	-.11
Cut Score				
Form	189	178	182	190
Adjusted	186	173	181	191
Pass Percentage				
Form	86.4	91.9	91.0	91.2
Adjusted	89.0	94.6	91.7	90.5

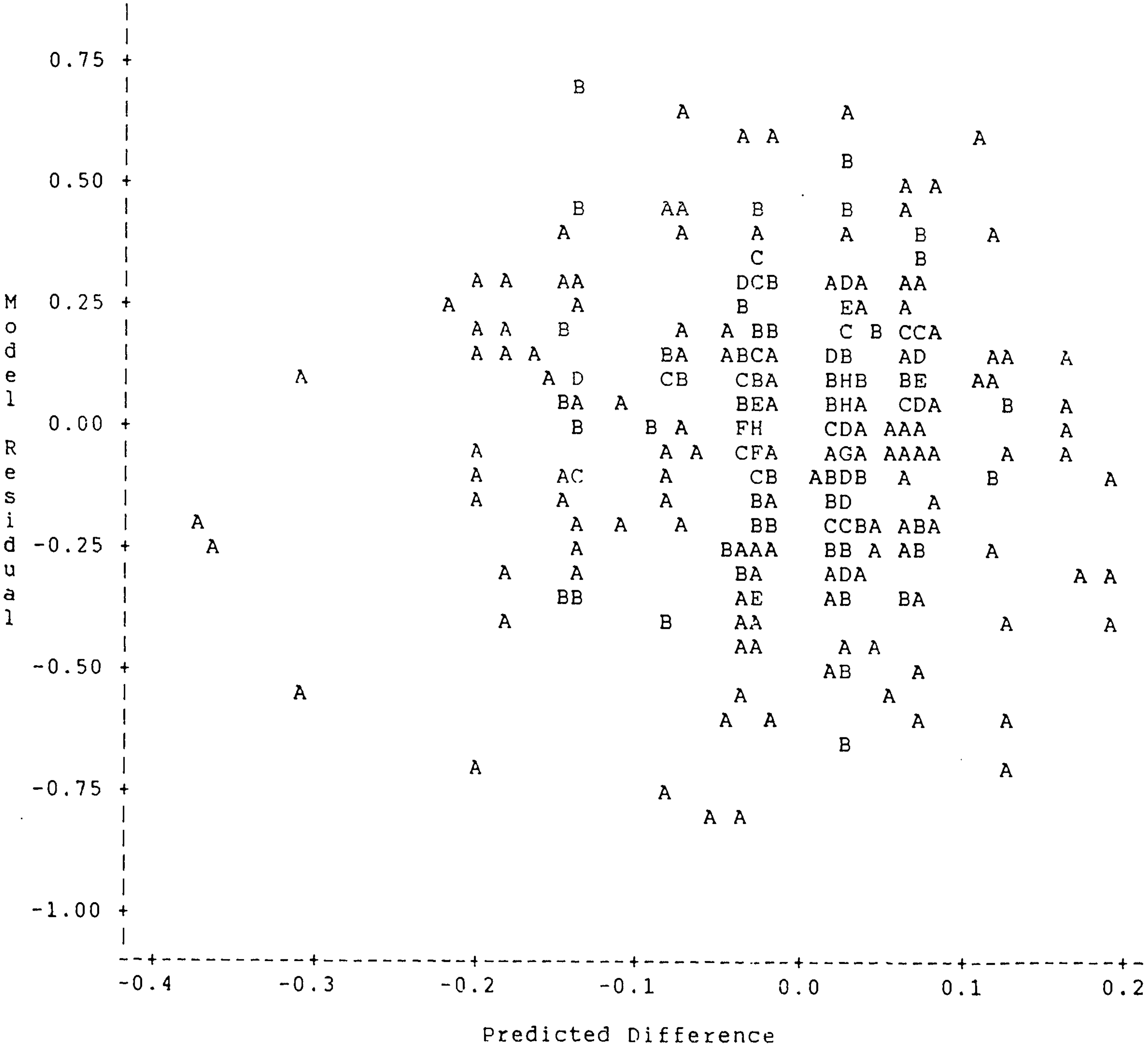
Table 5-2

Form and Adjusted B-Value Form Statistics:
1990 Form 1 - 1991 Form 2:
Program 2 Examination

	Form			
	1990 Form 1	1990 Form 2	1991 Form 1	1991 Form 2
B-Values				
Mean Form	-1.36	-1.26	-1.21	-1.06
Mean Adjusted	-1.41	-1.30	-1.24	-1.08
(Largest Abs. Difference: B-Value-Adjusted)	.05	.05	.05	.05
Cut Score				
Form	130	130	127	121
Adjusted	131	131	128	122
Pass Percentage				
Form	89.8	87.0	85.7	89.0
Adjusted	88.6	85.9	84.7	88.0

Figure 1

Plot of Model Residuals by Predicted Differences: Program 1 Examination



Note: A = 1 obs, B = 2 obs, etc.

Figure 2

Plot of Model Residuals by Predicted Differences: Program 2 Examination

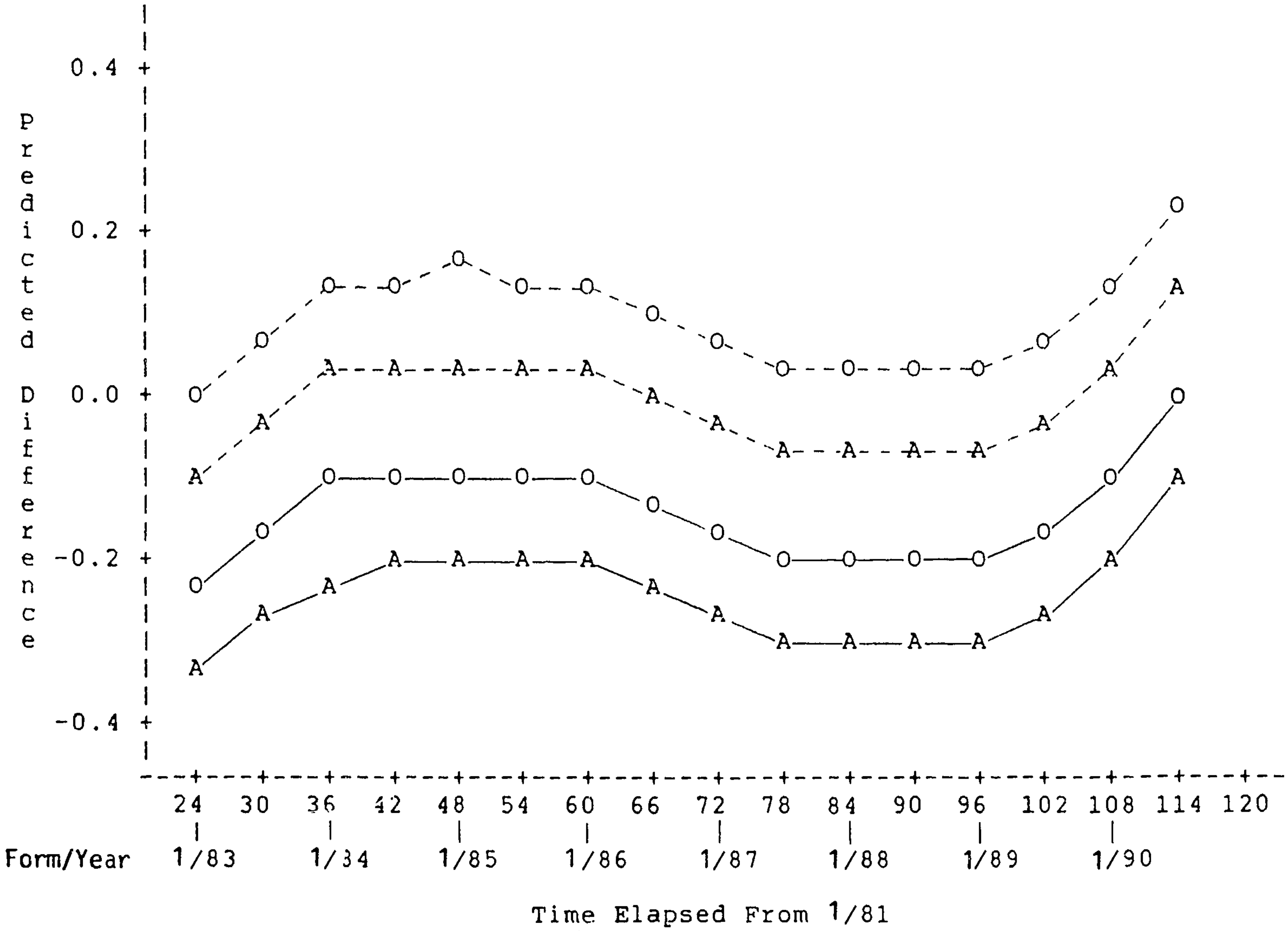
```
  0.8 +
      |
      |
      |                                                 A       A
      |         A                                               A
  0.6 +                                                         A A
      |                                 A                     AAAA
      |                                           A
      |                                                B   A      B
      |                                             A  A      A B
  0.4 +                  A                        B        A
      |         A                                 C    A   B  B A A
      |                                           D B  E   C  A  AA
      |                                           B    BA  C    DBB
      |                  A              A         C B  FA  AA DABBA
  0.2 +                  A                        F    DA  C  B A C
      |                           B     A         A    CA  CA CAB C
      |         A                                 C    EA  CA G C B
      |         A                             B   E    AA  GA CACDG
      |                  B                        D A  J   CB BBEBA
  0.0 +                           B     A         E    BA  CA DABBB
      |                  A                        C    DA  H  B DAD
      |                  A              A         C A  F   I  BBE E
      |         A        A       A            A   D    EA  D  D AAE
      |         A                A                B A  DA  B  BBB
 -0.2 +                          A                F    CA  B  CAB B
      |                                 A         C    H   CA B   A
      |                  B              A         C    C   DA B   C
      |                                       A   B    A   A  A DBA
      |                  A                        A    A   A  BADAA
 -0.4 +                                           A    B        B A
      |                                       A   A        A  B   A
      |                                       A            A    B
      |                  A                        A           A
      |                                                AA     A   A
 -0.6 +                          A                A           A   A
      |                                                      A    A
      |
      |
      |
 -0.8 +
      ---+-------+-------+-------+-------+-------+-------+-------+-------+--
      -0.150  -0.125  -0.100  -0.075  -0.050  -0.025   0.000   0.025   0.050

                              Predicted Difference
```

Model Residual (y-axis)

Note: A = 1 obs, B = 2 obs, etc .

Figure 3

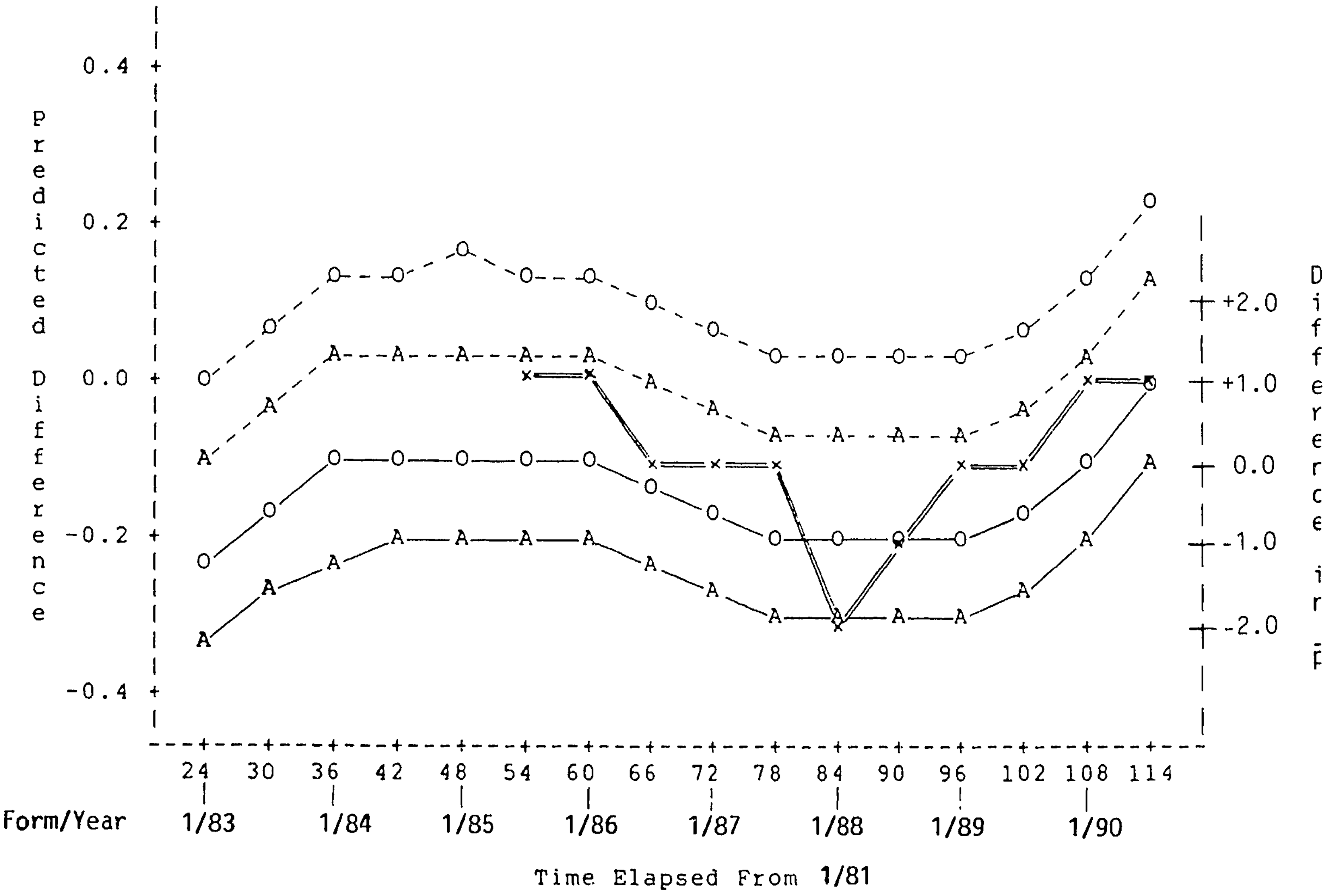
Program 1 Examination:
Plot of Predicted Differences Across Time From 181



Note: 12 months between administrations:
A with solid line : Domain 2 Category A
O with solid line : Domain 2 Categories B, C, & D
41 months between administrations:
A with dotted line : Domain 2 Category A
O with dotted line : Domain 2 Categories B, C, & D

Figure 4

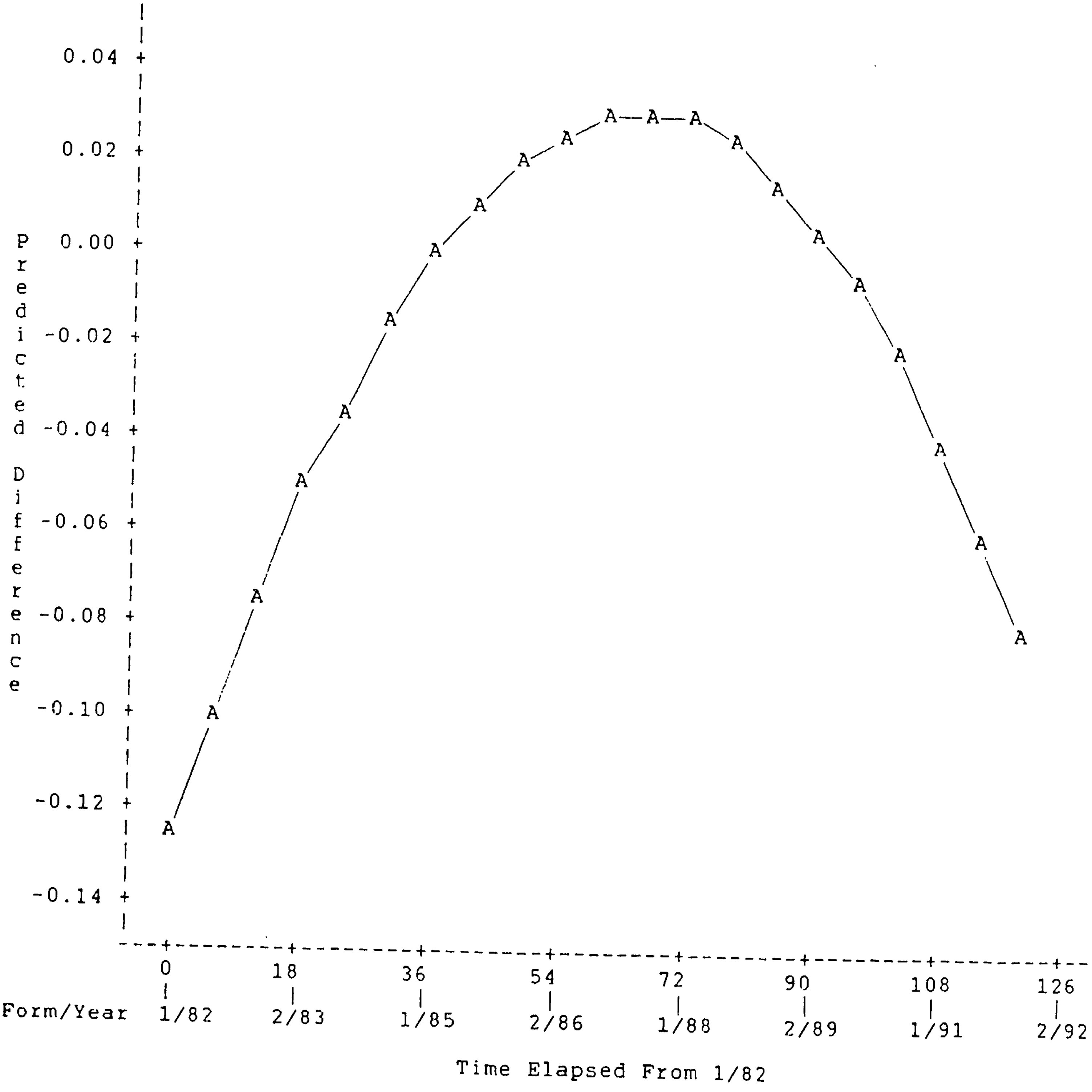
Program 1 Examination:
Plot of Predicted Differences Across Time From 181
and Difference in $\bar{p}$ (post-pre)



Note: 12 months between administrations:
A with solid line : Domain 2 Category A
O with solid line : Domain 2 Categories B, C, & D
41 months between administrations:
A with dotted line : Domain 2 Category A
O with dotted line : Domain 2 Categories B, C, & D
Difference in $\bar{p}$:
x with double solid line

Figure 5

Program 2 Examination:
Plot of Predicted Differences Across Time From 182


Predicted Difference
A
Form/Year
1/82
2/83
1/85
2/86
1/88
2/89
1/91
2/92
Time Elapsed From 1/82