SCHOOL OF OPERATIONS RESEARCH
AND INDUSTRIAL ENGINEERING
COLLEGE OF ENGINEERING
CORNELL UNIVERSITY
ITHACA, NY 14853

TECHNICAL REPORT NO. 875

November 1989

State SAT Averages: A Model-Based Exploration

by

Don Edwards and Cynthia B. Cummings*

*Don Edwards is Associate Professor, Department of Statistics, and Cynthia B. Cummings is Statistical Computing Consultant, College of Business Administration, University of South Carolina, Columbia, SC 29208.
Part of this work was completed while the first author was Visiting Scientist, School of Operations Research and Industrial Engineering, Cornell University, Ithaca, NY 14853. Previous versions of this paper were entitled "A long-term study of state SAT scores," with Cynthia B. Cummings listed as Cynthia Beckworth.

Abstract

Considerable attention in the educational statistics literature has been given to the naive use of state average SAT scores as measures of the relative efficacy of state educational systems. The main problem of course lies in the fact that the test-takers are a self-selected sample, the more talented students in the state being more likely to take the test. We explore the bias of the observed state average scores using a simple selection model based on truncated normal distributions. Empirical evidence for this model is found in regional patterns and each year in fourteen years of SAT state averages, 1972-1985. A first-order bias correction is proposed, and general patterns for the fourteen year period examined. A more general model is discussed.

KEY WORDS: Self-selection bias, truncated normal distribution, SAT scores, dynamic interactive graphics, robust regression.

ACKNOWLEDGEMENT: Revisions of this paper benefitted from criticism by Howard Wainer of the Educational Testing Service.

1. INTRODUCTION

The Scholastic Aptitude Test (SAT) , administered by the College Board, is basic as a predictor of an individual's collegiate success. It is widely used for that purpose by colleges and universities all over the United States. Unfortunately, despite warnings from statisticians at the Educational Testing Service (ETS) and elsewhere, another use for SAT test scores has evolved: they have become widely, naively used as measures of relative efficacy of state educational systems. Specifically, the simple averages of state test-takers' scores are routinely used to compare state educational systems, and to monitor educational system changes over time for a state and for the nation as a whole.

The fallacy in these secondary uses is of course that the test-takers from a given state are self-selected; they are not guaranteed or even likely to be representative of the state's student population as a whole. They tend to be the "cream of the crop" among the state's students, and hence their average tends to be higher than the "true" state mean SAT. Moreover, the effect of this bias will vary widely between states, since the percentage of high school seniors actually taking the test varies from under 3% in some states to more than 70% in others.

All this is well known and acknowledged by most reporters of state SAT averages (including ETS), in an accompanying disclaimer. Nevertheless, the scores are reported and decisions are made using them. Why does the use of these potentially misleading state SAT averages continue? We believe there are at least three reasons: (1)

there is an enormous need/demand for objective measures of efficacy of state and national educational systems; (2) educational scientists/statisticians have been unable to agree upon objective and practical alternatives satisfying this need; and (3) though the existence of the self-selection bias is well known, the severity of its effect in making state-to-state comparisons is probably not.

As a case in point for (1), we need look no further than our home state, South Carolina. During the fourteen year period 1972-85 studied here, South Carolina's raw state SAT average ranked 49th, 50th, or 51st among the fifty states and the District of Columbia. There is a certain stinging irony to placing 51st in a country with only fifty states. Yearly, on release day, the state scores (with accompanying disclaimer) typically made front-page news, and were the lead story on local T.V. and radio news. Fingers pointed. Heads rolled. In 1984, with motivation traceable largely to these state SAT averages, then-governor Richard Riley proposed a one cent increase in the state's sales tax (from three to four cents/dollar) to generate $217 million per year to fund his Educational Improvement Act (EIA). The bill passed after much debate, through a legislature infamous for killing tax increases for social reforms. It is no understatement to say that multi-million dollar decisions have been made based largely on these scores, and that, due primarily to them, education has been the number one issue in South Carolina state politics during the 80's.

Section 2 of this article is devoted to a summary of previous work and controversy on bias correction of state SAT scores. In section 3 a simple "cream of the crop" sampling model for the test-

takers is proposed, and the bias of the raw state SAT averages under this model is quantified. In section 4, implicit assumptions underlying a regression bias correction for state SAT averages are examined. In the fifth section, it is verified for 1972-85 that the pattern in the widely-studied 1982 data is not a fluke, and other patterns in the data are informally examined. In the sixth section, a more flexible selection model is discussed.

2. R-WARS, ANTI-R-WARS, AND ANTI-ANTI-R-WARS

In 1984, a report "State Education Statistics" was released by then-U.S. Secretary of Education Bell. It contained the 1982 SAT averages (that is, the averages of individuals taking the test in the 1981-82 academic year; denoted here by Y_i, i=1,..,51), the estimated proportion of high school seniors taking the test ("participation rate", p_i), and a number of other demographics for each of the fifty states and DC. Several articles in the educational literature followed, examining the empirical relationships among the variables (these works were apparently conducted independently); these were in turn followed by a backlash of articles by statisticians objecting to sloppy use of formal statistics. The dominant predictor in each article was participation rate, p. Figure 1 shows a scatter plot of Y vs. p for the 1982 data as a backdrop for discussion.

PLACE FIGURE 1 ABOUT HERE.

Powell and Steelman (1984) used multiple stepwise regression analyses to model Y, arriving at a model in terms of p, $\sqrt{p}$, percent of test-takers female, percent of test-takers minority, and public school expenditures per student. They are the only authors of those discussed here to include the analog of Figure 1 in their discussion. Page and Feifs (1985) independently but by similar means arrive at a different regression model (for the 1983 data) using a linear term for p along with the percentage caucasian, employment rate, and average income. Their plot of predicted-Y versus Y suggests that their model underpredicts states with large Y's.

Wainer et al (1985) provide a variety of exploratory analyses for the state SAT means, including in some of these an indicator function for state "type", SAT or ACT (in 22 states SAT is taken by the majority of college-bound students; in the others, the ACT is predominant. These two state-groups are for the most part identifiable as low-p and high-p clusters in Figure 1). Their ultimate assertion, though, is that immediate conclusions drawn from the data in "State Education Statistics" can be very misleading. They give some reasons, which Wainer (1986; 1989, with discussions) elaborates on. He discusses five pitfalls involved in analyses such as those found in Powell and Steelman (1984) and Page and Feifs (1985). Wainer's five pitfalls are, in his own words:

(1) Promiscuous adjustment without an explicit model.
(2) Adjusting for a posttreatment concomitant variable.
(3) Extrapolating without checking the model.
(4) Inconsistent aggregation bias - measuring the covariates over a different population than the outcome variables.
(5) Just because it's adjusted does not mean it's helpful.

The meanings of most of these pitfalls will be clear to many statisticians without further explanation; others are referred to the articles, which provide several good examples and additional references. Wainer (1986) suggests a correction of his own, based on translating the ACT averages of SAT states to SAT averages (using a separately established, sharp linear regression to predict SAT from ACT). This could effectively place all states in the 30%-70% participation rate category.

Powell and Steelman (1987) responded to Wainer (1986) with some defenses of their own work and some criticisms of his (for example, it also operates in the absence of a selection model, in the hope that states in the 30%-70% participation rate range have equal self-selection bias).

Our own interest in this problem began with the EIA, and has been conducted very gradually since then, with an interim summary by Beckworth (1988). The truncated normal selection model discussed in §3 has recently been independently proposed by Taube and Linden (1989); a similar approach was also alluded to by Wachter (1989). Taube and Linden use it on 1985 data, both as a correction in its own right and in multiple regression analyses similar to those of Powell and Steelman (1984) and Page and Feifs (1985). The work summarized here should complement that of Taube and Linden (1989), since it is more oriented to identification of implicit statistical assumptions underlying the corrections, and includes a look at a long-term data set, and a more general selection model. We have, however confined ourselves to pursuit of the "what" question: "What are the true state mean SAT's?" as opposed to the multiple

regression approaches of the educational specialists, which implicitly combine this with the "why" question ("Why are they what they are?") by including predictors like state expenditures, family income, etc. in their models. The "what" question is important and difficult enough to merit this article's attention. It ought to be resolved beyond reasonable debate before the "why" question is addressed. Also, in restricting to the "what" question, pitfalls 3, 4, and 5 are avoided for the most part, not to mention the interpretational maze of collinearity among the demographic predictors employed by multiple-regression approaches to the "why" question.

Still more recently, Holland and Wainer (1989) have written a combined critique of the truncated-normal selection model as proposed in Taube and Linden (1989) and a draft version of this article. The model they study is technically different from ours, since it assumes there are no testing errors. The patterns they observe in (1) the frequency distributions of SAT scores and (2) state test-taker variances are not inconsistent with the sampling model proposed here under a non-negligible testing error whose variance decreases with increasing ability; they could also be explained in terms of the "fuzzy-boundary" model in §6; or, a combination of these.

3. A "CREAM OF THE CROP" SAMPLING MODEL FOR SAT TEST TAKERS

The truncated-normal selection model proposed here seems too simple to be true, but empirical evidence is surprisingly supportive.

In this section the focus is on a single state, and hence the subscript i is suppressed.

Let A denote the actual SAT-ability of a single randomly selected senior in the given state. If this student takes the test with full motivation, the SAT score will be his/her ability A plus test error ε, whose distribution has mean 0 by definition. The first model assumption is:

> ***(A1) The actual SAT-ability A of a randomly selected senior in the state follows a normal distribution with a mean μ and variance σ^2.***

The primary goal then is simply to estimate μ. Ideally an objective measure of the accuracy of the estimate (e.g. a standard error) should also be provided; this will hopefully be dealt with in a later study. Bias is the most important problem in these test averages; variance comes into play for only about 1 in 4 states, those with low numbers of test-takers (§5).

The test-takers are self-selected, and will tend to be the "cream of the crop" among the state's students. A simplification of this statement is the second model assumption.

> ***(A2) If 100p% of the state's seniors take the test, then this is (approximately) the "best" 100p%, i.e. those with highest SAT-abilities in the population of seniors.***

(A2) is an assumption on the conditional distribution of a single test-taker's ability *given* p; it does not fall prey to the concomitant variable pitfall #2. Its motivation is as follows: the societal reasons

for p being large or small notwithstanding, most students will take the SAT because they aspire to an institution or scholarship requiring it, and believe (no doubt counseled by their parents, teachers, siblings and peers) that their ability is as great or greater than others around them who have achieved these goals successfully.

Assumptions (A1) and (A2) combine formally to the following: the marginal distribution of the ability of a test-taker, A^T, is the conditional distribution of a normal(μ,σ^2) variable given it exceeds its (1-p)th percentile (a (1-p)-truncated normal distribution; note this would not imply that the SAT scores themselves follow a truncated normal distribution; their distribution will be distorted by test errors ε, and by dependencies (we do not assume the test-takers' abilities are independent of each other or of the test errors).

Even the most reasonable of readers will object to the assumption of a "sharp" boundary at the (1-p)th percentile: students close to this boundary will be more indecisive, some above it not taking the test, some below it taking the test. This concept is referred to here as the "fuzzy selection boundary". In §6, a "fuzzy boundary" selection model is studied. Interestingly, it is found that the methods employed here assuming a sharp boundary are not only robust but completely valid under this fuzzy boundary model (using a modified and possibly even more believable version of assumption (A3) of §4).

Let ϕ and Φ denote, respectively, the standard normal p.d.f. and c.d.f., and let z_{1-p} denote the (1-p)th quantile of Φ. Under the formalization of assumptions (A1) and (A2) and using additivity of

the expectation operator the mean Y of the n test-takers has expectation equal to that of the (1-p)-truncated normal variable:

$$E(Y) = E\{(1/n)\sum_{j=1}^{n}(A_j^T+\varepsilon_j)\} = E(A^T) = \mu + \sigma B(p), \qquad (3.1)$$

where the bias function B(p) is given by

$$B(p) = \phi(z_{1-p})/p \qquad (3.2)$$

for $0 < p \leq 1$ (this can be calculated directly without much difficulty, or see for example Johnson and Kotz (1970)). Figure 2 illustrates this "cream of the crop" sampling model and its induced bias on the test-takers' mean.

PLACE FIGURE 2 ABOUT HERE

B(p) can be hand-calculated with a normal table and a scientific calculator, or in any computer language/statistical package that includes the equivalent of the Probit function. A table is handy and is provided (Table 1). These values were computed in a five line Statistical Analysis System (SAS 1985) data step whose third line is:

B = (1/P)*(0.39894228)*EXP(-(PROBIT(1-P)**2)/2).

PLACE TABLE 1 ABOUT HERE

Figure 3 shows a plot of B(p) vs. p; B can be regarded as the function (1/p) flattened at both extremes of p. There is a similarity of shape (for $0 < p < 0.7$) between B(p) and the state SAT vs. p point cloud in Figure 1. The data seems to plead (perhaps even "promiscuously"): "regress me!" The validity of this will be discussed in §4, but for the moment we give in to temptation and show the reader the fitted regression of 1982 state SAT averages Y_i, i=1,..,51 on the transformed participation rates $B(p_i)$, in Figure 4. A "gut" reaction is that the fit of this peculiar, mechanistically-motivated regression function is too precise to be a complete accident (adding that this particular point swarm is quite uncooperative in an empirical curve-fitting exercise; there is visually detectable lack of fit in regressions of Y on p; on p with p^2; on 1/p; on p^γ for any γ; and on p with $\sqrt{p}$).

PLACE FIGURE 3 ABOUT HERE

PLACE FIGURE 4 ABOUT HERE

In consideration of (3.1), a natural approach to bias adjustment might be to use all the test takers scores from the given state to obtain a consistent estimate of σ, insert this into (3.1), and solve to obtain a consistent estimate of μ. Obtaining an estimate of σ, though, involves a number of additional assumptions on the variance-covariance structure of the abilities A_j^T and the test errors ε_j. The first choice would probably be to assume that these A's and ε's are mutually uncorrelated, and that the ε's are of constant variance σ_ε^2;

under these and (A1)-(A2) the variance S^2 of the test takers has expectation

$$E(S^2) = \sigma_\varepsilon^2 + \sigma^2 B_2(p) \tag{3.3}$$

where

$$B_2(p) = 1 - B(p)[B(p) - z_{1-p}]. \tag{3.4}$$

$B_2(p)$ is strictly increasing in p. Solving (3.3) for an estimate of σ^2 first requires some estimate of σ_ε^2. One idea might be to use the known reliability indices of the SAT (which exceed 0.90); the estimate of error variance using these will probably be an underestimate, though, since test error as measured by reliability is instantaneous test error, whereas ε here includes more "low frequency" sources of error, e.g. illnesses, lack of sleep, emotional trauma. Another possibility is to assume (as in §4 below) that state variances σ_i^2, i=1,...,51 are approximately equal, in which case (3.3) would provide another regression for estimation of σ_ε^2 and the common σ^2. As mentioned in Beckworth(1988), these and other schemes have been attempted (using only the test-taker variances S^2) but no sensible answers obtained.

The variance-covariance structure of the A's and ε's is probably not simple. The test takers themselves come in clusters in time and space, which could cause dependencies among abilities. Also, it seems intuitive that test error variance would decrease with increasing ability since, for example, more talented students will do less quessing. At the state level (assuming mutually uncorrelated A_j^T's and ε_j's) this latter would result in

$$E(S^2) = \sigma_\varepsilon^2(p) + \sigma^2 B_2(p), \qquad (3.5)$$

where the function $\sigma_\varepsilon^2(p)$ increases with p. Ignoring the first term in (3.5) would lead to overestimates of σ for low-participation states, and consequently to absurdly "overcorrected" estimates of their state means μ_i using (3.1).

Exploration of the variance/covariance structure of test-taker abilities and test errors is certainly a route for further research, if the entire collection of state scores could be made available for several states/years. If it could be done well the resulting bias corrections would be preferable to the first-order corrections discussed in §4. However, it is seems universally true across applications of Statistics that good models for variances are more elusive and less robust than models for means.

4. ASSUMPTIONS OF A FIRST-ORDER BIAS CORRECTION FOR STATE SAT AVERAGES

The clean fit of the state SAT means Y_i to the transformed participation rates $B(p_i)$ suggests regression; here the implicit assumptions of this are examined. Considering (3.1) and (3.2) for each state and DC gives

$$Y_i = \mu_i + \sigma_i B(p_i) + \varepsilon_i *, \qquad (4.1)$$

i=1,..,51, where $\varepsilon_i *$ represents the averaged test errors for the test-takers of state i, $E(\varepsilon_i *)=0$. Equivalently, define a national mean SAT

μ_N and the state deviations from the national mean $\delta_i = \mu_i - \mu_N$, i=1,..,51. Then (4.1) becomes

$$Y_i = \mu_N + \sigma_i B(p_i) + \delta_i + \varepsilon_i * , \qquad (4.2)$$

i=1,..,51. As these statements stand, they are not very useful for estimation, since there are only 51 pairs $(Y_i, B(p_i))$ for estimation of 102 unknowns μ_i, σ_i. Equation (4.2) states that the variability in the observed averages Y_i derives from variability in the σ_i, $B(p_i)$, δ_i, and $\varepsilon_i *$. If it were the case that the dominant ("first-order") source of this variability was due to the $B(p_i)$, (4.2) becomes a simple linear regression model of the Y_i on the transformed participation rates $B(p_i)$, though the errors about the regression line will not have equal variance, nor be uncorrelated. Relaxing this statement somewhat, the following assumptions would validate the reflex regression of the Y_i on the $B(p_i)$, where the motivation would be to use the residuals (up to a constant) as adjusted state SAT means:

(A3) The state ability standard deviations σ_i are approximately equal, and

(A4) The state deviations from the national mean δ_i have mean 0 (conditonally on p_i).

(A3) is of course a common assumption in analysis of variance settings, where it is often approximately true, and many formal inference procedures are robust to moderate departures. It seems intuitive that it would be approximately true in this setting as well, though there are no claims to robustness here. In (A4), the most

outrageous assumption yet, the constants δ_i have become random variables (hence this might be considered, in a minimal sense, an Empirical Bayes approach). This in itself is not so outrageous, but to believe them to have mean 0 conditionally on the p_i is certainly questionable. Intuition suggests that a state's true SAT mean μ, and hence its δ, would increase with its participation rate p. Looking at the data, though, it is difficult to believe that the relationship could be very strong. High-participation-rate states include several (South Carolina, North Carolina, Georgia) which by other indicators probably have negative δ_i, as well as all the New England states, which probably have positive δ_i. Participation rate is driven primarily by the admissions policies of in-state colleges and universities. Nevertheless, (A4) is probably the weakest of the assumptions underlying the regression corrections discussed here; since we expect some bias to remain in the resulting estimates, they are labeled "first-order" bias corrections.

An exploratory analysis of the 1982 data using the small-scale dynamic interactive graphics program MacSpin$_{TM}$ (Donoho, Donoho, and Gasko(1985)) is encouraging. If (A1)-(A4) hold approximately, there should be some regional correlation among the δ_i. This manifests itself in a plot of Y_i versus $B(p_i)$ (Figures 5 and 6) as fairly well-defined and approximately parallel "sub-lines" corresponding to regionally contiguous groupings of states with similar δ_i's. The grouping shown in Figure 6 is given by:

Northeast:	CT, MA, ME, NH, NJ, NY, RI, VT
Mideast:	DE, IL, IN, KY, MD, MI, OH, PA, TN, VA, WV
Deep South:	AL, AR, GA, LA, MS, NC, SC, TX
Midwest:	IA, KS, MN, MO, ND, NE, OK, SD, WI
West:	AZ, CA, CO, ID, MT, NM, NV, OR, UT, WA, WY
Ungrouped:	AK, FL, HI, DC

These groups were found very quickly using MacSpin$_{TM}$'s capacity for interactive identification, subsetting and re-subsetting of points in a scatter plot. They are is fairly consistent with typical regional classifications, with the exception that the border-Southern states Virginia, Tennessee, and Kentucky seem to belong in the Mideast region here. Several of the regions include a wide range of participation rates, the exceptions being the Midwest (all ACT states) and New England (all SAT states). The patterns in this plot are supportive of the first-order corrections, since states in the same region will have similar adjusted mean SATs.

PLACE FIGURE 5 ABOUT HERE

PLACE FIGURE 6 ABOUT HERE

If assumptions (A1)-(A4) are palatable, (4.2) becomes

$$Y_i = \mu_N + \sigma B(p_i) + \varepsilon_i^{**} , \qquad (4.3)$$

where $\varepsilon_i^{**} = \delta_i + \varepsilon_i^{*}$ has mean 0. Denote the intercept and slope estimates of a regression of Y_i on $B(p_i)$ by $\hat{\mu}_N$, $\hat{\sigma}$ (these should be taken only as a fitted slope and intercept, though, not as estimates of

the "true" μ_N and σ; see §5-6). The ith residual e_i is an unbiased predictor of δ_i. Rather than use these residuals directly, each state's SAT here is adjusted to its predicted value under a 50% participation rate (a typical test-taking rate considering both the SAT and ACT). The ***first-order bias-adjusted state SAT means*** are defined to be

$$\hat{SAT}_i = \hat{\mu}_N + \hat{\sigma}B(.5) + e_i, \qquad (4.4)$$

i=1,..,51. Observations on the pattern in Figure 5 and similar figures for each of the years 1972-85 (§5) lead to the suspicion that there are states with unduly influential δ_i values in some years, and consequently robust regression (with the Bisquare influence function, using the function rreg of the statistical package Splus_{TM} (Becker, Chambers, and Wilks (1988)) was used in regressions here.

Table 2 and Figure 7 summarize the results of the first-order bias corrections for the 1982 data. Note the reduction in SAT range from about 300 points to about 150. Again, regions are much more well-separated and internally homogeneous after the correction.

5. FOURTEEN YEARS: 1972-1985

The intent in examining a long-term data set was to verify repeatability of the pattern of the 1982 data and to do more informal checks of the assumptions (A1)-(A4). Some "eyeballing" observations on national and state trends are irresistible; these should not be taken too seriously until independent data are available to test/refine the bias corrections. Verbal and quantitative SAT

averages and numbers of test takers n_i for states i=1,..,51 were provided by the College Board. The participation rates p_i were estimated by the ratio of number of test takers to an estimate of the number of graduating seniors in the state (N_i); this latter figure was in turn estimated using number of public high school graduates (supplied by the U.S. Department of Education) and an inflation factor to include private high school graduates computed from years for which this data was available. The estimated participation rates so obtained showed good agreement (correlation exceeding 0.99) with those published in Powell and Steelman (1984), Page and Feifs (1985), and Taube and Linden(1989); the largest relative discrepancies, with a few exceptions, could be attributed to roundoff errors at low values of p in these papers.

Figure 8 shows the state SAT averages Y_i plotted against participation rates p_i and transformed participation rates $B(p_i)$ in three-year increments 1972, 1975, 1978, 1981, and 1984, and the fitted robust regression lines for these years. Table 3 summarizes the regressions. Intermediate-year patterns are essentially interpolations of those shown, and the 1985 plots closely resemble those of 1984. Positions of individual states in the Figure 8 point clouds are approximately as shown in Figures 1 and 5, with a few exceptions, three of which (Alabama, Mississippi, and the District of Columbia) are identified. In the '80's these plots are not greatly different from the 1982 data already discussed; in the 70's the pattern is stable for the most part with the exception of the three identified states, which qualify as "outliers" in the regressions of these years: they were severely downweighted in the final iterations

of the robust regressions. Excepting these three, perhaps the most surprising aspect of this data upon cursory inspection is how *little* change seems to occur over time.

PLACE FIGURE 8 ABOUT HERE

PLACE TABLE 3 ABOUT HERE.

Figure 9 is meant primarily to show the trends in national SAT, both unadjusted and using the first-order bias adjustment of (4.4). The national SAT mean is in each year the weighted average of the state means, weighted by state senior class sizes N_i. The unadjusted SAT national means decline throughout the '70s and then rise again in the early '80s. This pattern has received much attention/concern in the news media and among educational specialists. Speculation as to its cause has included the possibility of changing participation rates. Actually, participation rates nationwide did not change much during this period (the national weighted average p, for example, varied only from 0.307 to 0.355). The national mean follows a very similar pattern under the first-order correction, shown at right in Figure 9.

Three of the four states in Figure 9 (Colorado, Indiana, and North Carolina) are representative of most states in that their time plot, both in unadjusted and adjusted means, closely parallels the national. This leads to the belief that whatever factors are influencing the national means to drift are operating consistently at state levels as well. Particularly curious is the large drop from 1974

to 1975, which is noticeable in most of the individual state time plots.

The fourth state in Figure 9, Ohio, is one of the few whose raw SAT time plot does not mirror the national, and also one of the few whose participation rate changed substantially during 1972-85. Ohio's "p" declined fairly steadily from 0.28 in 1972 to 0.16 in 1977, and then stayed more or less constant. Its unadjusted SAT clearly increased relative to the US values during 1972-77, and then stayed more or less constant. Its adjusted SAT is essentially coincident with the national pattern across the entire period. This seems to support the first-order corrections, since parallel-to-the-national is the norm for most states whose p did not change much (not all the states whose participation rates changed were so intuitively supportive of the first-order corrections, though, as discussed below).

PLACE FIGURE 9 ABOUT HERE

Figure 10 is an attempt at a complete description of trends in the first-order adjusted SAT relative to the national level for all states. Plotted here against time are the values (adjusted state mean - national adjusted mean), essentially estimates of the deviations δ_i in (4.2); the horizontal line across each plot now corresponds to the national mean. The subsetting of the states into groups of four or five is not alphabetical, or by region, but simply in such a way as to attempt to obtain as many clean plots as possible in as little space as possible (in a few cases this obviously was not successful). It is easy

to see here the large gains relative to the US for Mississippi, Alabama, and DC as expected from Figure 8.

PLACE FIGURE 10 ABOUT HERE

The majority of the sequences in Figure 10 are stable year to year, but an interesting few are not. The adjusted means for North Dakota, South Dakota, Wyoming, and Utah fluctuate in a random-like fashion by as many as 40 points in consecutive years, and 20 point swings are commonplace. Why? These 4 states, in that order, were consistently lowest in number of test-takers n_i over the 14 year period, with 300-600 test takers per year. There is a "second tier" of states with 1000-2000 test takers per year which corresponds very closely to the "second-order" unstable sequences of Figure 10, i.e. Mississippi, Oklahoma, Alaska, Kentucky, Idaho, Louisiana, Nevada, Arkansas, and West Virginia (passing over DC, which has a single unexplained large discrepancy in 1977, and Nebraska, which has a single large drop in 1975, the year the University of Nebraska dropped the SAT as an admission requirement). 19 of the other states show 3,000 - 10,000 test takers per year, and the remaining 19 from 10,000 to more than 100,000 per year. Figure 10 shows essentially the residuals of the regression fits of (4.3), and their stochastic properties under (A1)-(A4) should reflect those of the $\varepsilon_i^{**} = \delta_i + \varepsilon_i^*$. The stability of the large-n_i sequences (and common sense) leads to the belief that the δ_i do not change much year to year: large changes in a state's true mean SAT (relative to the national average) will typically not come quickly; after all, the "experiment" that

generates any one of these δ_i takes 18 years. There is no physical reason to believe that the δ_i for the Dakotas, Wyoming, and Utah are substantially more unstable than those for, e.g., Montana and Colorado. This leads finally to the belief that the large fluctuations in Figure 10 for the low-n_i states are due to the ε_i*, the averages of the test errors, whose variances will be in at least a gross sense inversely proportional to $\sqrt{n_i}$ (another theory might be that these fluctuations are due to instability in the p_i for these states, which when amplified through the transformation $B(p_i)$ could create an errors-in-variables problem in the regression. Plots of $B(p_i)$ against time do not support this; they are typically smooth in their year-to-year changes. The largest one-year change (excepting Nebraska '74-75), multiplied by the typical $\hat{\sigma}$ of 100, would account for only a 20 point change in adjusted SAT (Montana '73-74), and there are only a handful of such changes in B(p) which could account for even10 point changes in adjusted SAT). High variance in the adjusted SAT scores for the dozen or so low-n states could possibly be repaired by inclusion of translated ACT scores for these states, or by pooling data across years in a time series model of some sort.

As previously mentioned, a few states did experience large (proportionally speaking) trends in participation rate. Some of these seemed to show trends remaining in the first-order adjusted means, and some did not. Alaska's p increased from .22 to .39 after 1977; Florida's p increased from .25 to .45 over the entire period; Illinois' decreased from .25 initially to about .14 in 1983; Michigan's decreased from .23 initially to .10 in 1982; Ohio has already been mentioned; and Washington's p increased from .11 in 1974 to .28 in

1985. The first-order adjusted state means for these states do not seem to be trending in any substantial way in Figure 10. On the other hand, Arizona's p increases after 1975 from .08 to about .15; Minnesota's increased after 1975 from .05 to .13; and Nevada's increased after 1974 from .11 to .21, and these three states do seem to show decreasing trends in Figure 10. These latter indicate the possibility of slight undercorrection in low participation-rate states. The game being played here is a tricky one, though, since some of these trends may be real; several of the other states in Figure 10 seem to have trends relative to the national average in their adjusted SAT.

As mentioned before, In 1975 The University of Nebraska changed its admissions policies. There was a concomitant sudden drop in SAT participation rate from .37 in 1974 to .13 in 1975, followed by a steady decrease from .13 to .06 in 1975-79. The first-order adjusted mean for Nebraska does not ride this period smoothly, but eventually does return to approximately its 1974 level (Figure 10(j)). A possible explanation for this might be a suddenly "very fuzzy" selection boundary caused by uncertainty as to who should or should not take the SAT after the University of Nebraska dropped it as a requirement.

6. A FUZZILY-TRUNCATED-NORMAL SELECTION MODEL

It is counterintuitive to believe that the "cream-of-the-crop" selection boundary is actually as sharp as hypothesized in (A2). A less restrictive selection model may be described as follows. Again, fix a state and let A, ε, μ, σ be as in §3. Let I be the indicator of the

event that a student of ability A takes the test, i.e. I=1 if he/she takes the test, 0 otherwise. Conditionally on A=a, I is taken to be a Bernoulli random variable such that P(I=1) = P(test is taken) = r(a) for a selection function r(a). In a given state r(a) will be increasing in a: the more talented the student, the more likely he/she will be to take the test. A useful parametric form for r(a) in this setting is the Probit function

$$r(a) = \Phi[\,(a-\alpha)/\beta\,], \qquad -\infty < a < \infty, \tag{6.1}$$

where $\beta \geq 0$, α are unknown constants. The case $\beta=0$ is defined as the limiting (discontinuous) r(a) obtained fixing α (or expressing it as a function of β) and taking $\beta \to 0$.

The parameters α, β have the job of quantifying the state-specific factors influencing the decision to take the test, e.g. the admissions requirements of in-state universities. For example, if a statewide proportion p of all students take the test; a natural reaction to this is to require r(a) to satisfy the constraint E[I] = p, i.e.

$$p = E[I] = \int_{a=-\infty}^{\infty} \Phi[(a-\alpha)/\beta]\, d\Phi[(a-\mu)/\sigma] = \Phi[(\mu-\alpha)/(\beta^2+\sigma^2)^{1/2}]. \tag{6.2}$$

This overall-participation-rate constraint then simplifies to a constraint on α,

$$\alpha = \mu + z_{1-p}(\beta^2+\sigma^2)^{1/2}. \tag{6.3}$$

The remaining free parameter β is a "fuzzy boundary" parameter. Taking $\beta \to 0$ under (6.3), the selection function r(a) reduces to the sharp-boundary selection function of §3:

$$r(a) = \begin{cases} 1 & a > \mu + \sigma z_{1-p} \\ 0 & \text{otherwise} \end{cases}$$

and as $\beta \to \infty$ the selection boundary becomes so fuzzy that self-selection is for all intents and purposes independent of ability: the selection model becomes that of a px100% random sample of all seniors. It will be convenient to express β relative to σ as $\rho = [\sigma^2/(\beta^2+\sigma^2)]^{1/2}$; then $(1-\rho)$x100% might be interpreted as "percent fuzziness" in the ability selection model. When $\rho=1$, the boundary has 0% fuzziness, the sharp boundary of §3; when $\rho=0$ there is 100% fuzziness in the selection by ability (the px100% random sample model).

Under (6.1)-(6.3), the conditional density function of an individual's ability A given I=1 (the marginal density function of the ability of a single test taker, called A^T in §3) is

$$f(a) = (1/p)\Phi[(a-\alpha)/\beta](1/\sigma)\phi[(a-\mu)/\sigma], \quad -\infty < a < \infty. \qquad (6.4)$$

Figure 11 shows plots of these density functions for an example p=0.15, μ=800, σ=100, and percent fuzziness $(1-\rho)$x100% =0%, 5%,

10%, and 20%. Apparently the density loses its sharp-peaked, severely skewed appearance with only a modest amount of fuzziness in the selection boundary.

PLACE FIGURE 11 ABOUT HERE

The conditional moment generating function of A given I=1 is

$$m(t|\ I=1) = (1/p)\Phi(\rho\sigma t - z_{1-p})\exp(t\mu+\sigma^2t^2/2) \qquad -\infty < t < \infty. \tag{6.5}$$

Defining $\psi(t)=\ln[m(t)]$, and using the fact that $\psi'(0) = E(A|\ I=1)$ and $\psi''(0) = Var(A|\ I=1)$, the mean and variance of a test-taker's ability can be seen to be

$$E(A|\ I=1) = \mu + \rho\sigma B(p) \tag{6.6}$$

and

$$Var(A|\ I=1) = \sigma^2 B_2(p;\ \rho),$$

for

$$B_2(p;\ \rho) = 1-\rho^2 B(p)[B(p)-z_{1-p}]\ . \tag{6.7}$$

It can be seen taking $\rho \to 0$ (under (6.3)) that expressions (6.4)-(6.7) reduce to the corresponding quantities under a px100% random sample, and as $\rho \to 1$ to the corresponding quantities for the (1-p)-truncated normal random variable; in particular (6.6) becomes (3.1) and (6.7) becomes the second term in (3.4).

What is really interesting is that under this fuzzy boundary selection model, with *any* amount of fuzziness, the mean of the test takers is still linearly related to the same function B(p); the slope of the line simply decreases with ρ (ignoring, as done in §3 without confession, the annoyance that the denominator of the arithmetic mean, called a constant n, is in fact a random variable $\sum_{j=1}^{N} I_j$; Slutsky's theorem will give $E(Y) \cong \mu + \rho\sigma B(p)$ under these large N's).

Considering the national level, the foundational expression (4.2) modified with the additional parameters ρ_i becomes

$$Y_i = \mu_N + \sigma_i \rho_i B(p_i) + \delta_i + \varepsilon_i * ,$$

for states i=1,..,51. If the assumption (A4) is still palatable, and under a modified assumption (A3*):

> ***(A3*) The degree of fuzziness in a state's selection boundary is approximately proportional to its Ability standard deviation ($\sigma_i \rho_i$ constant over i),***

(which seems almost more believable than (A3)), the residuals of a regression of Y_i on the $B(p_i)$ are still unbiased predictors of the δ_i, and the first-order adjusted means of §4 are still valid for comparisons between states. The slope coefficient of the regression now estimates not σ but the common $\sigma_i \rho_i = \sigma *$.

7. CONCLUSION

The most useful and beautiful scientific models are the *simple* ones that provide surprisingly accurate (though still imperfect) predictions to the output of complicated processes. A famous example, admired by Wainer (1989) and probably familiar to most readers, is Wald's World War II aircraft-return model (Mangel and Samaniego (1984)). Briefly, in order to make decisions on where best to armor aircraft, Wald examined the patterns of bullet holes in aircraft that had returned. Under the simple probabilistic assumption that bullets strike target planes according to a uniform distribution, it would follow that the locations of bulletholes on the returning planes must be the locations *least* in need of armoring; the planes struck in the vulnerable spots did not return.

Because they are always technically false, it is usually easy to talk oneself out of believing any given model. Of course the abilities of high school seniors in a given state are not really, exactly, normally distributed; of course if a proportion p of the students take the test, this will not be exactly the top px100% in abilities. At the same time, it is difficult to imagine a Messerschmidt pilot who would randomly scatter his shots over his target plane. Most of them would probably aim. Doubtless this occurred to Wald, but after considering the matter further (the aim of the enemy pilot would be disturbed to some degree by evasive action of the Allied plane, and so on), and more importantly after examining consistency of the model with available data, he and others decided that it was accurate enough to justify decisions as to where planes should be armored. It was not worthwhile to fuss about second-order precision in the model; our

boys were dying out there. In South Carolina, when the proposed 1983 EIA was under hot debate in the legislature, our boys (and girls) were also "dying out there", in school systems underfunded for decades. Even a crude adjustment to state SAT scores would have been preferable to a release-with-disclaimer policy, which only fueled the fires of the legislative ostriches opposing the bill.

The theory discussed here is fairly consistent with the data examined, but it is not by any means rigorously tested yet, and cannot be without a carefully conducted, carefully interpreted, random-selection survey. The first-order corrections defined are not meant as substitutes for estimates from such a survey (even if they were nearly unbiased, the plots of Figure10 show that they have large variance for low-participation rate states). At present, it is merely *interesting* and *promising* that such simple selection models are so consistent with the patterns observed in the SAT state means, and that the "first-order" adjusted SAT's they yield seem to "make a lot of sense". There are other models which could fit this data (particularly sobering is the possibility that the $E(\delta|p)$ taken to be 0 in Assumption (A4) might in fact be an increasing function of p, in which case the plot would still "look good" if this function was similar in shape to B(p), and regression corrections would be more than a little wrong).

As mentioned in Holland and Wainer (1989), the National Assessment of Educational Progress (NAEP) will provide some state-by-state data (for 13 year olds, mathematics scores, for 38 states) by 1990. This survey, if its sample sizes are large enough to provide low-variance estimates, could certainly be used for long-needed

comparisons between states. If so, it could also be used to evaluate and improve adjustment strategies like the ones discussed here. It simply could not be a bad idea to harness the (free) information in more than half a million yearly test scores (considering both ACT and SAT), if this could be done well. If general theories of self-selected testing adjustment could be made reliable (with reliable standard errors), they could be used at other organizational levels as well, e.g. school system levels. They could be used in place of expensive national surveys in alternating years, the savings from this diverted to other education-related concerns. They could conceivably even be used in international comparisons.

REFERENCES

Becker, Richard A., Chambers, John M., and Wilks, Allan R. (1988), *The New S Language*, Pacific Grove, CA: Wadsworth and Brooks/Cole.

Beckworth, Cynthia M. (1988), "Regression corrections for self-selected sampling in state-to-state SAT comparisons", M.S. thesis, Department of Statistics, University of South Carolina, Columbia, SC 29208.

Donoho, Andrew W., Donoho, David L., and Gasko, Miriam (1985), *MACSPIN$_{TM}$ graphical data analysis software*, Austin, TX: D^2 Software, Inc.

Holland, Paul W. and Wainer, Howard (1989), "Sources of uncertainty often ignored in adjusting state mean SAT scores for differential participation rates: the rules of the game," submitted for publication.

Johnson, Norman L. and Kotz, Samuel (1970), *Continuous Univariate Distributions-1*, New York: Wiley.

Mangel, M. and Samaniego, F.J. (1984), "Abraham Wald's work on aircraft survivability," *Journal of the American Statistical Association*, **79**, 259-267.

Page, E.B. and Feifs, H. (1985), " SAT scores and American states: seeking for useful meaning," *Journal of Educational Measurement*, **22**, 305-312.

Powell, B. and Steelman, L.C. (1984), "Variations in state SAT performance: meaningful or misleading?" *Harvard Educational Review*, **54**, 389-412.

Taube, K.T. and Linden, K.W. (1989), "State mean SAT score as a function of participation rate and other educational and demographic variables," *Applied Measurement in Education*, **2**, 143-159.

Wachter, K.W. (1989), Discussion of Dr. Wainer's paper. To appear, *Journal of Educational Statistics* , vol.**14.**

Wainer, H. (1986), "Five pitfalls encountered while trying to compare states on their SAT scores," *Journal of Educational Measurement*, **23**, 69-81.

Wainer, H. (1989), "Eelworms, Bullet Holes, and Geraldine Ferraro: some problems with statistical adjustment and some solutions," to appear, *Journal of Educational Statistics* , vol.**14.**

Wainer, H., Holland, P.W., Swinton, S., and Wang, M. (1985), "On 'State Education Statistics'," *Journal of Educational Statistics*, **10**, 293-325.

TABLE 1

The function B(p)

p	B(p)	p	B(p)	p	B(p)	p	B(p)
0.01	2.6652	0.26	1.24576	0.51	0.78199	0.76	0.40904
0.02	2.4209	0.27	1.22461	0.52	0.76623	0.77	0.39435
0.03	2.2681	0.28	1.20223	0.53	0.75059	0.78	0.37961
0.04	2.1543	0.29	1.18036	0.54	0.73507	0.79	0.36481
0.05	2.0627	0.30	1.15898	0.55	0.71965	0.80	0.34995
0.06	1.9854	0.31	1.13804	0.56	0.70432	0.81	0.33502
0.07	1.9181	0.32	1.11753	0.57	0.68910	0.82	0.32000
0.08	1.8583	0.33	1.09742	0.58	0.67396	0.83	0.30488
0.09	1.8043	0.34	1.07768	0.59	0.65889	0.84	0.28966
0.10	1.7550	0.35	1.05828	0.60	0.64390	0.85	0.27430
0.11	1.7094	0.36	1.03922	0.61	0.62898	0.86	0.25881
0.12	1.6670	0.37	1.02046	0.62	0.61412	0.87	0.24316
0.13	1.6273	0.38	1.00199	0.63	0.59932	0.88	0.22732
0.14	1.5898	0.39	0.98379	0.64	0.58456	0.89	0.21128
0.15	1.5544	0.40	0.96586	0.65	0.56984	0.90	0.19500
0.16	1.5207	0.41	0.94816	0.66	0.55517	0.91	0.17845
0.17	1.4886	0.42	0.93070	0.67	0.54052	0.92	0.16159
0.18	1.4578	0.43	0.91345	0.68	0.52590	0.93	0.14437
0.19	1.4282	0.44	0.89641	0.69	0.51129	0.94	0.12673
0.20	1.3998	0.45	0.87957	0.70	0.49670	0.95	0.10856
0.21	1.3724	0.46	0.86290	0.71	0.48212	0.96	0.08976
0.22	1.3459	0.47	0.84641	0.72	0.46753	0.97	0.07015
0.23	1.3202	0.48	0.83009	0.73	0.45294	0.98	0.04941
0.24	1.2953	0.49	0.81391	0.74	0.43833	0.99	0.02692
0.25	1.2711	0.50	0.79788	0.75	0.42370	1.00	0.00000

TABLE 2
1982 summary

State	Part. rate p	raw SAT	adj. SAT*
AK	0.298	923	884
AL	0.061	964	839
AR	0.040	999	855
AZ	0.113	981	885
CA	0.386	899	879
CO	0.168	983	909
CT	0.691	896	927
DC	0.507	821	822
DE	0.545	897	904
FL	0.374	889	866
GA	0.489	823	821
HI	0.471	857	852
IA	0.028	1088	929
ID	0.071	995	876
IL	0.138	977	892
IN	0.471	860	855
KS	0.053	1045	914
KY	0.062	985	861
LA	0.058	975	847
MA	0.655	888	913
MD	0.499	889	889
ME	0.472	890	885
MI	0.105	973	874
MN	0.075	1028	912
MO	0.107	975	877
MS	0.027	988	827
MT	0.087	1033	924
NC	0.465	827	821
ND	0.028	1068	910
NE	0.059	1045	918
NH	0.556	925	934
NJ	0.646	869	893
NM	0.081	997	885
NV	0.167	917	843
NY	0.617	896	915
OH	0.161	958	882
OK	0.048	996	860
OR	0.415	908	893
PA	0.513	885	887
RI	0.616	877	896
SC	0.481	790	787
SD	0.025	1075	912
TN	0.083	999	888
TX	0.323	868	835
UT	0.038	1022	876
VA	0.509	888	889
VT	0.530	904	909
WA	0.187	982	914
WI	0.100	1011	909
WV	0.072	968	850
WY	0.055	1017	887

*adjusted to 50% participation

TABLE 3
regression coefficients

Year	ordinary least squares intercept	ordinary least squares slope	robust regression intercept	robust regression slope
1972	854.4498	83.7359	857.2259	90.8954
1973	837.8490	90.4910	835.6243	98.9523
1974	826.2051	98.8425	830.3144	101.3089
1975	803.6141	103.7184	812.9794	101.9049
1976	802.0601	103.9493	816.5903	99.2290
1977	787.4048	112.5855	801.6902	107.2442
1978	790.3189	110.4743	797.4391	108.7202
1979	789.7724	108.1541	799.6151	104.9519
1980	791.0731	106.0608	799.8509	102.5413
1981	789.6363	107.7641	795.3780	105.4776
1982	793.4512	107.9001	797.3674	106.1055
1983	792.7431	109.7363	797.3722	107.8103
1984	799.8647	110.6512	805.1600	108.4082
1985	811.9221	111.9337	816.2615	109.8183

FIGURE 1

State SAT averages vs. participation rates, 1982

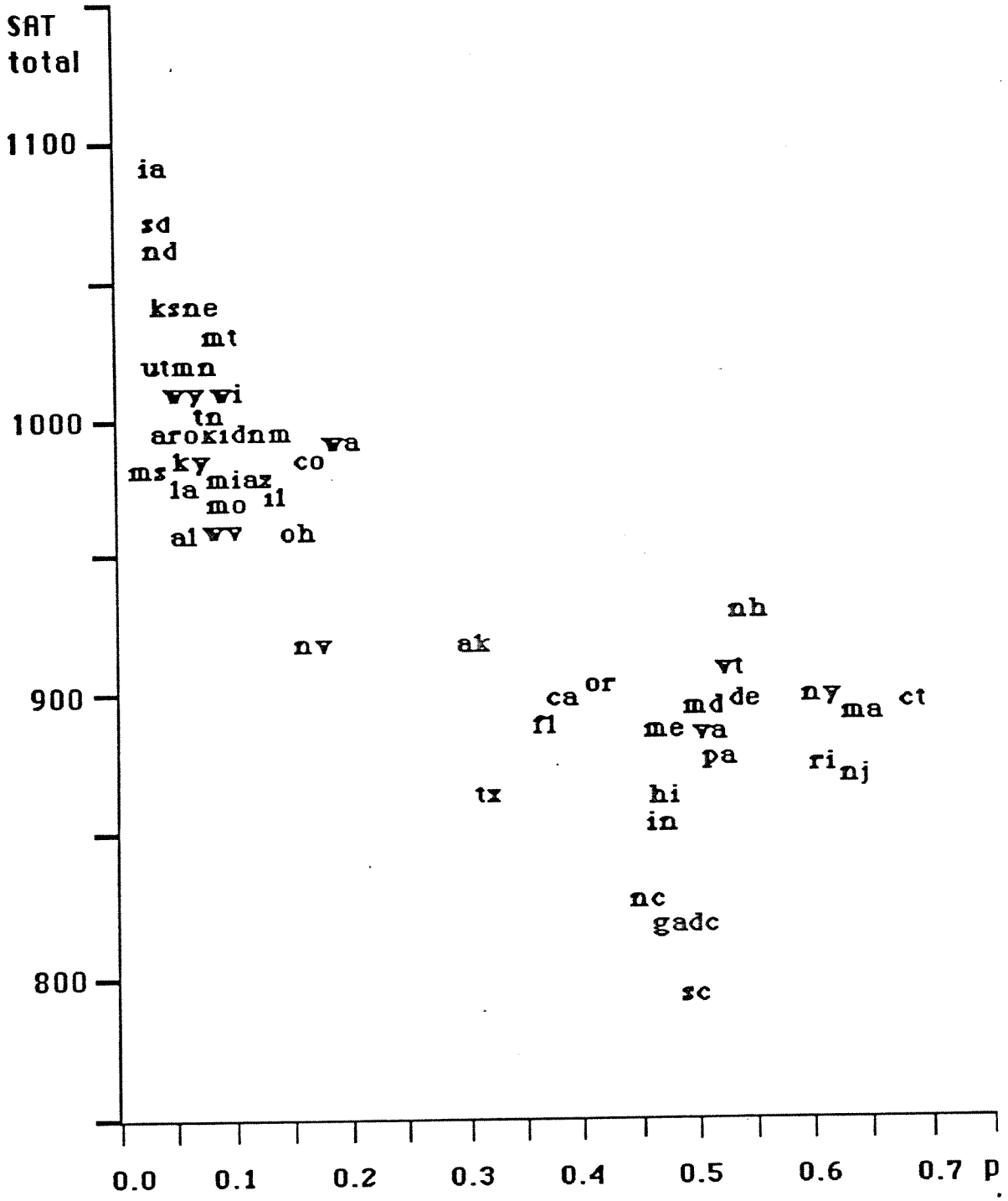
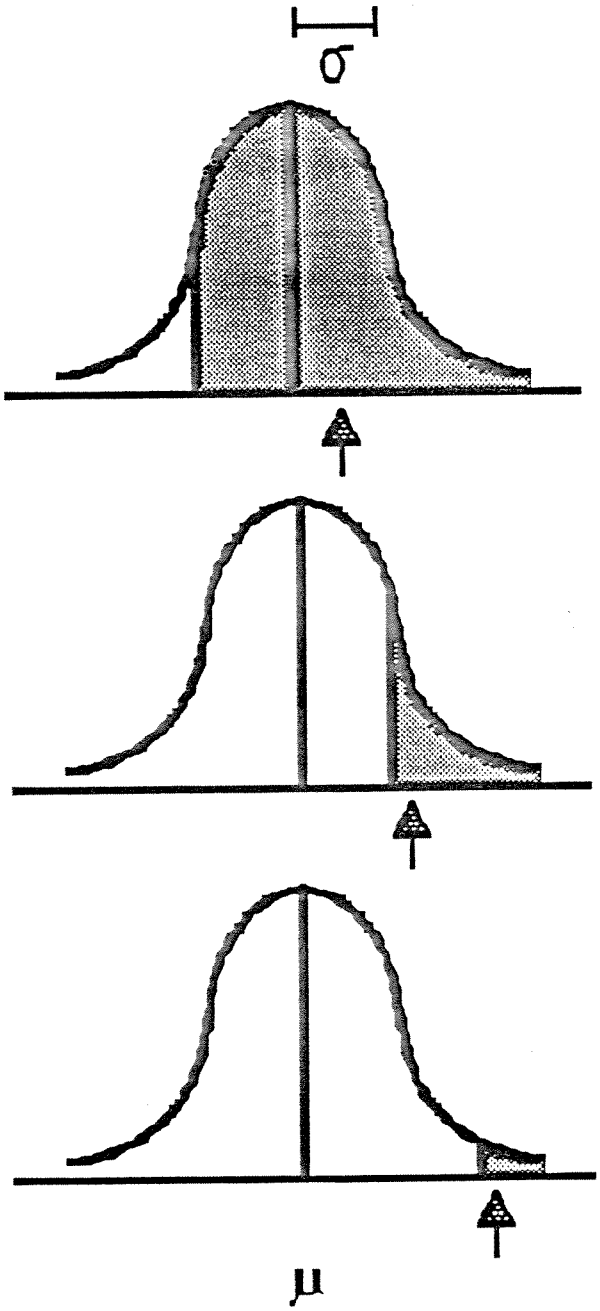


FIGURE 2

Cream of the Crop Sampling Model

Identical Ability Distributions



70% participation

Mean of Test Takers (denoted ↑)

$\approx \mu + (0.496)\sigma$

30% participation

Mean of Test Takers

$\approx \mu + (1.159)\sigma$

5% participation

Mean of Test Takers

$\approx \mu + (2.063)\sigma$

FIGURE 3

The bias function B(p)

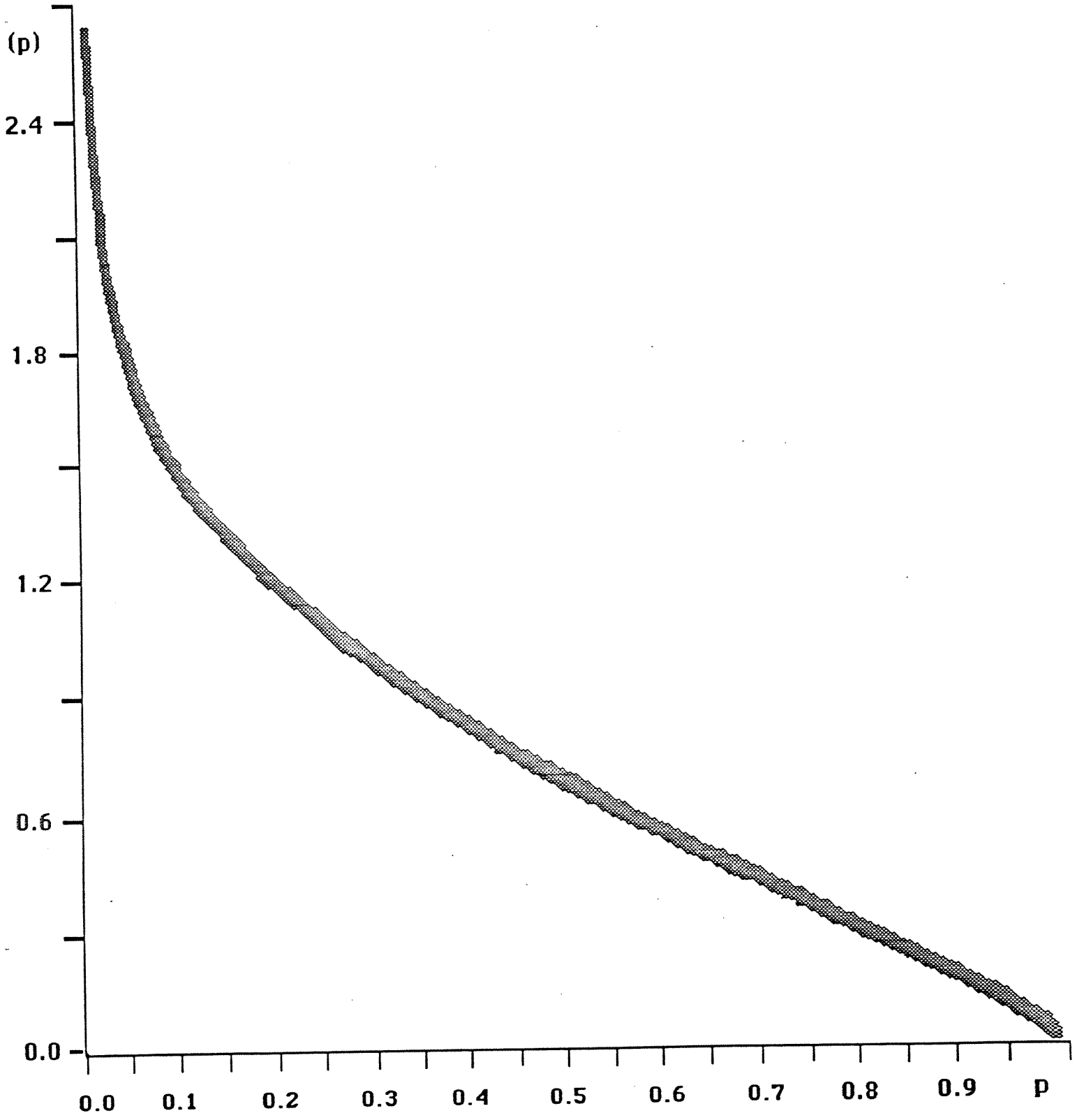


FIGURE 4

Fitted regression, 1982 scores

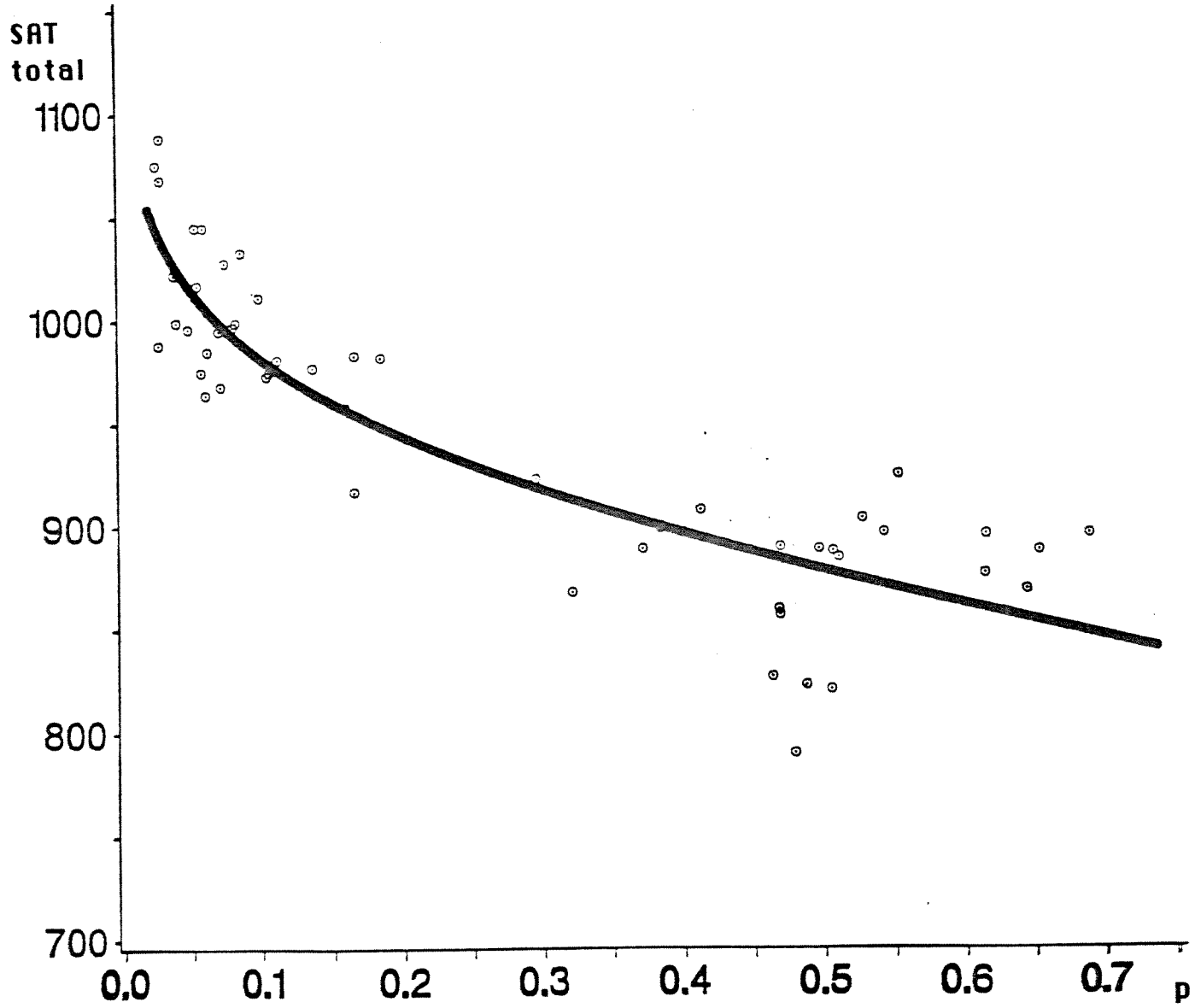


FIGURE 5

State SAT averages vs. transformed participation rates, 1982

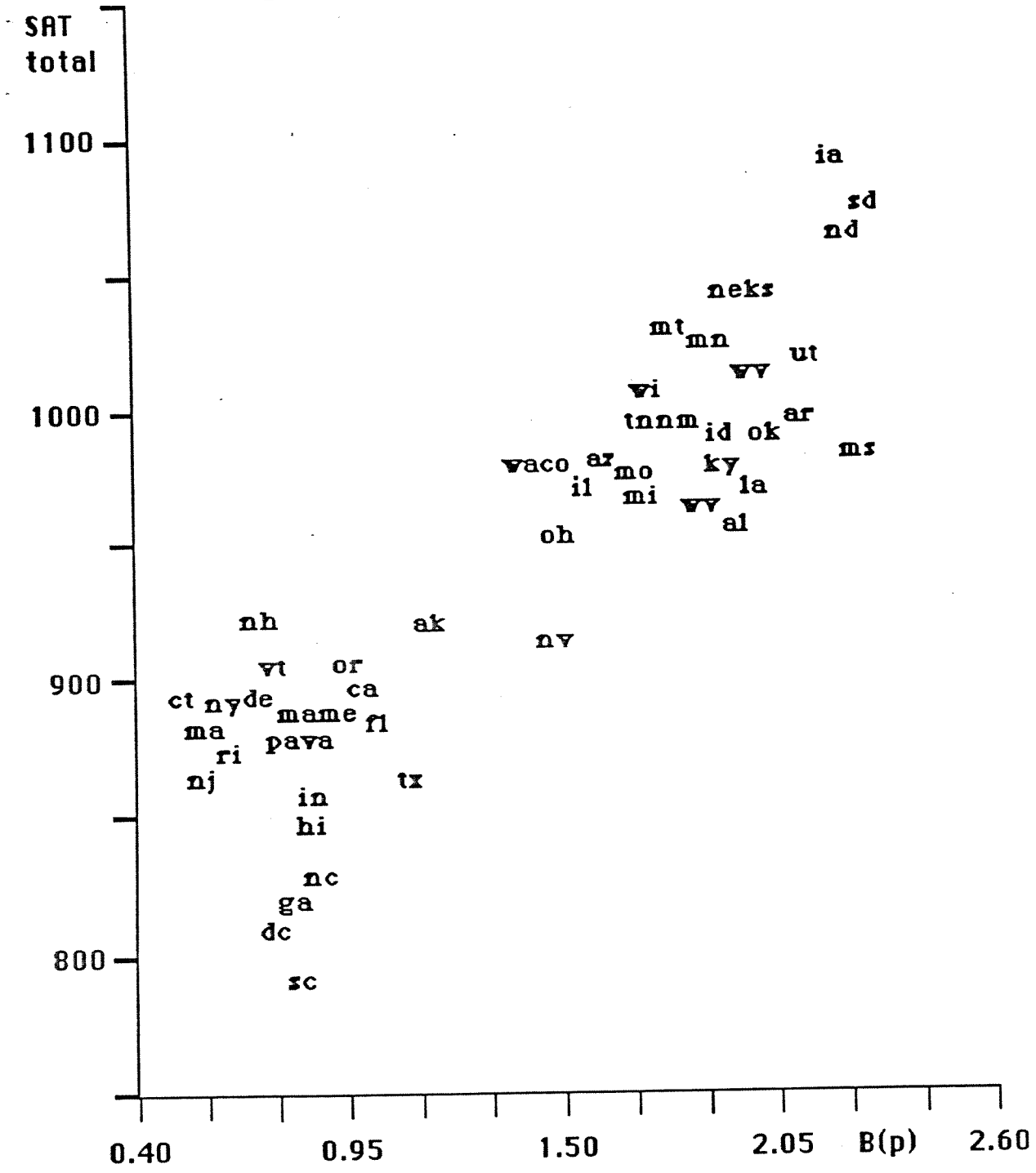


FIGURE 6

State SAT averages vs. B(p), by region

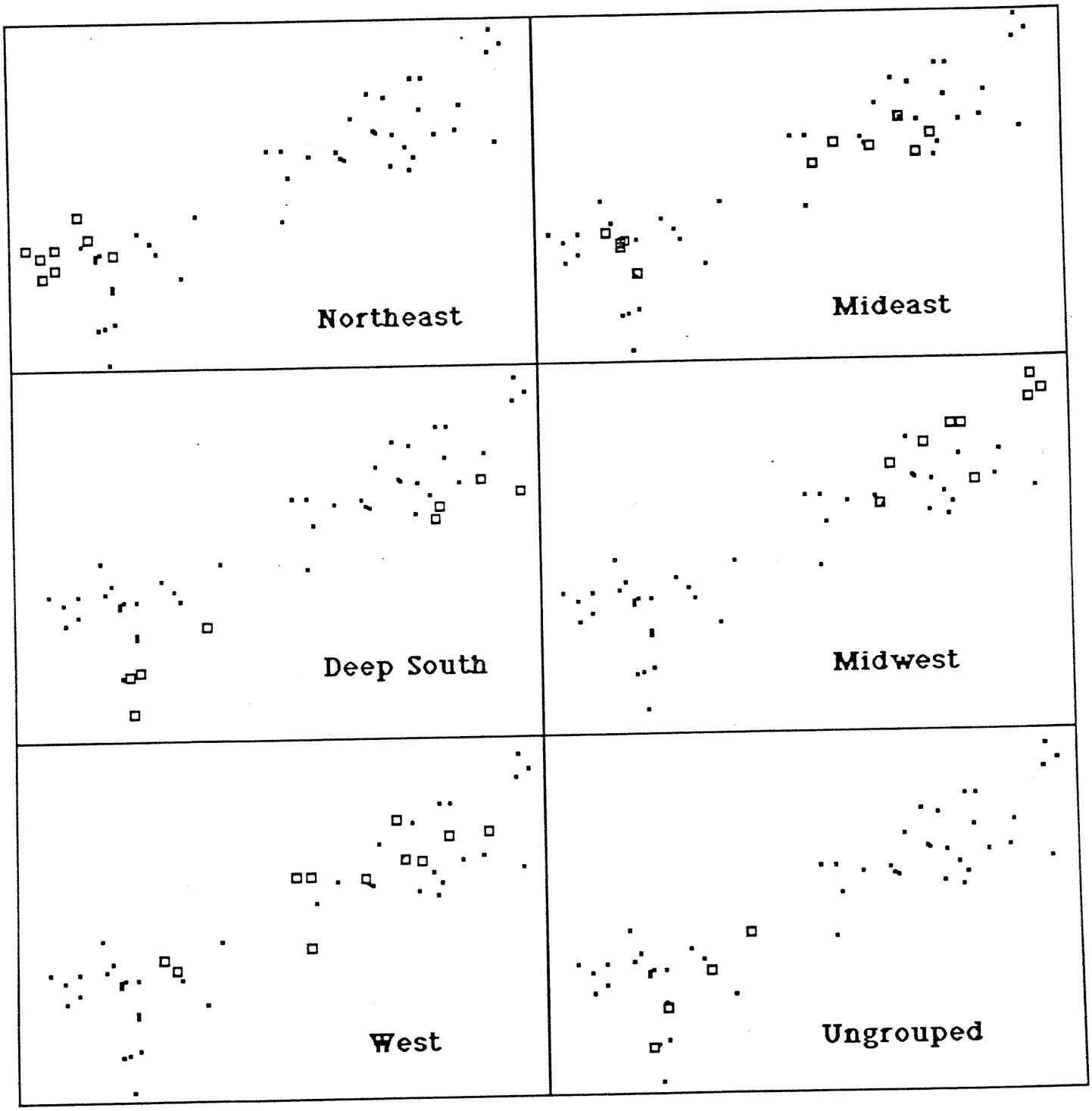


FIGURE 7

Raw and adjusted state SAT's, by region, 1982

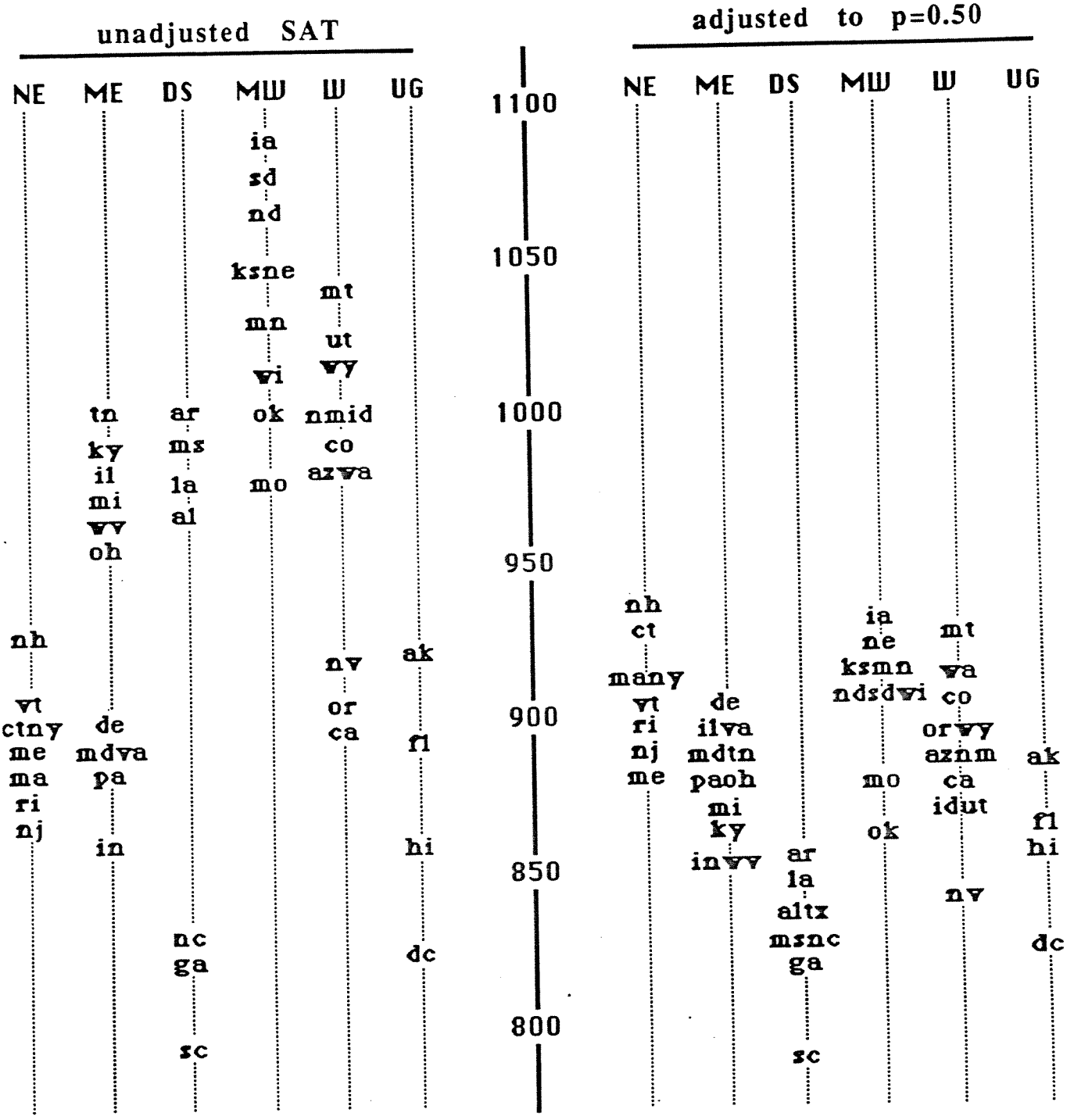
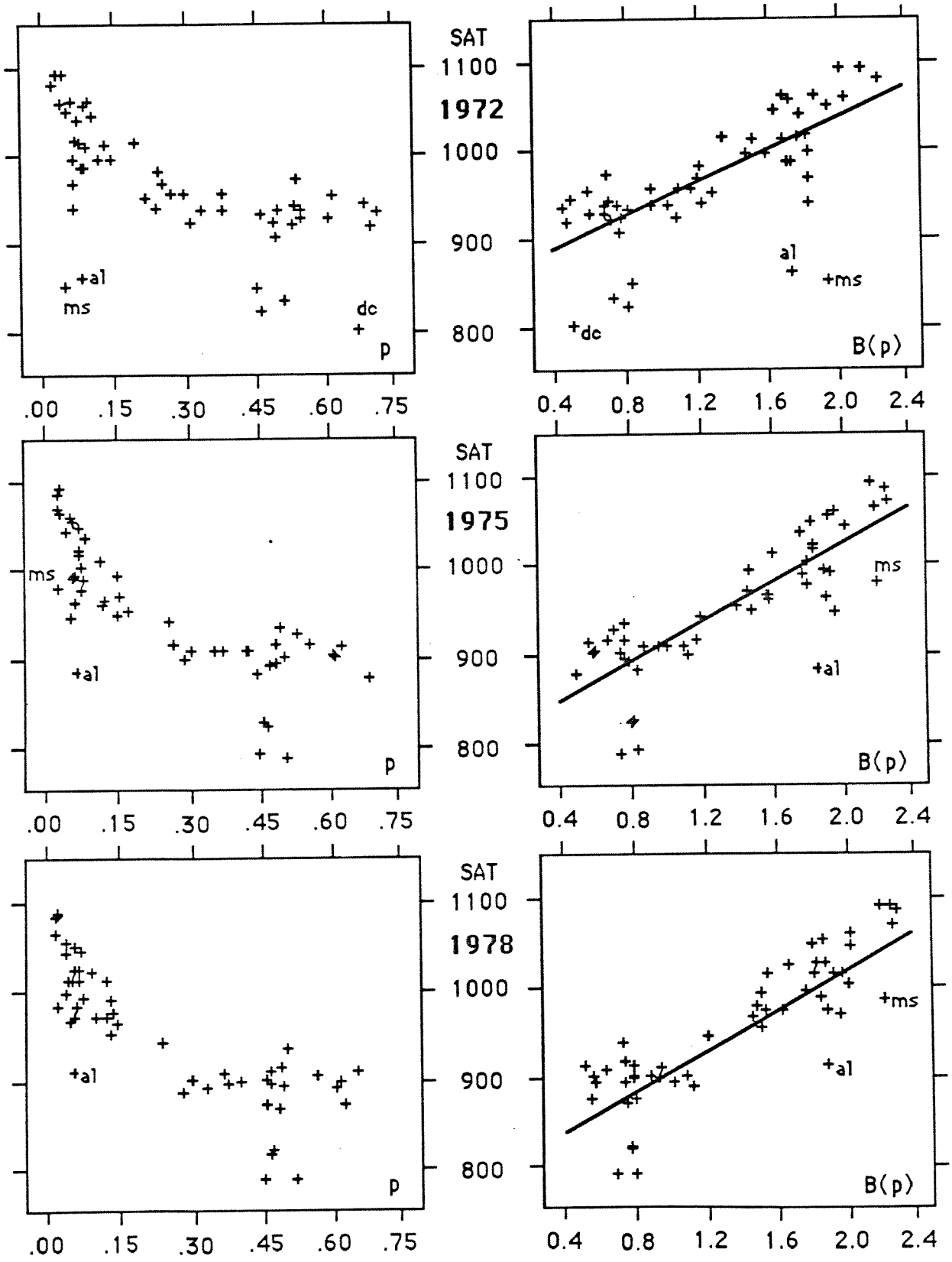


FIGURE 8

State SAT averages vs. p and B(p), 1972-84, with robust regression lines



(continu
on next
page....)

(Figure 8 continued)

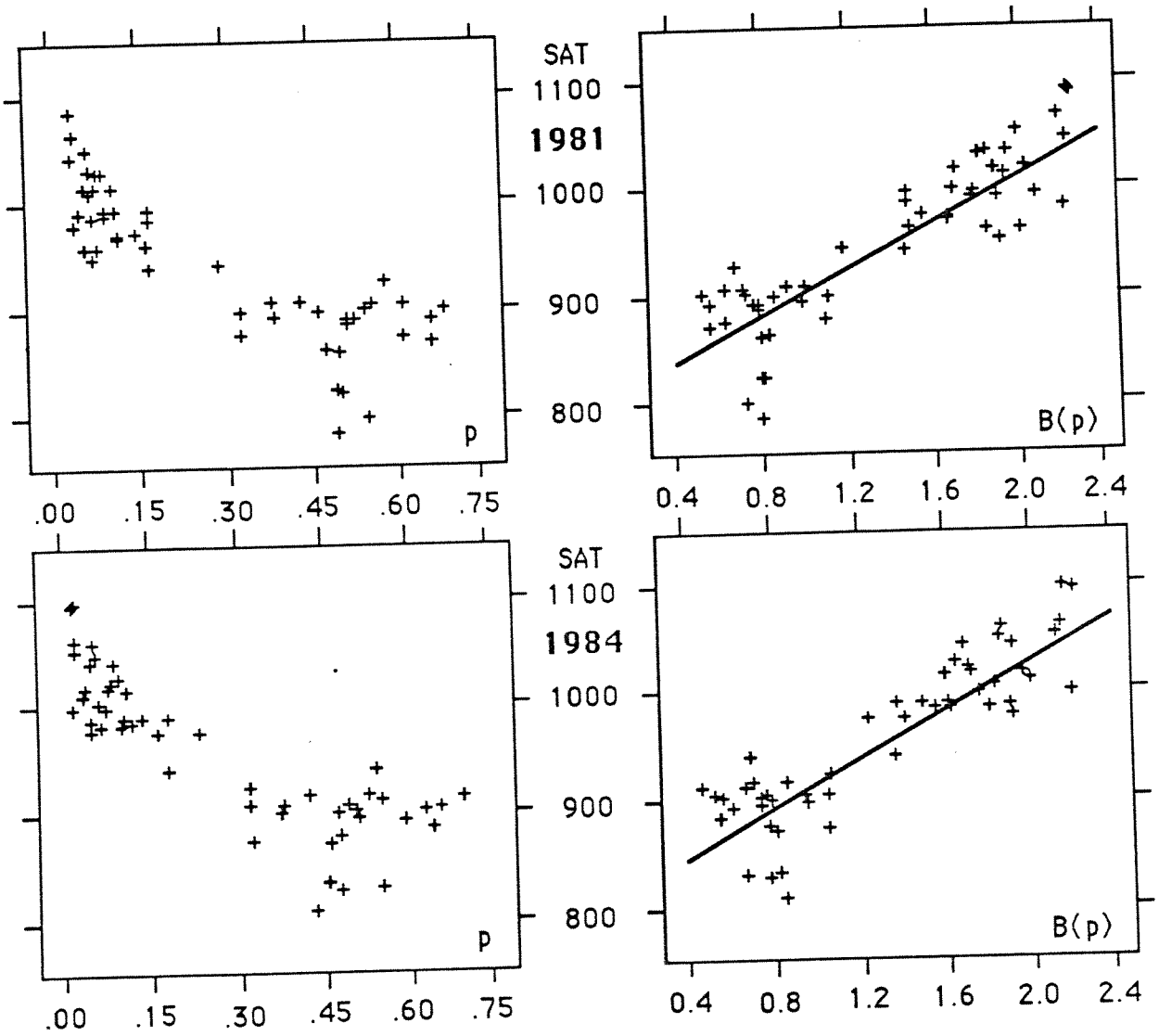
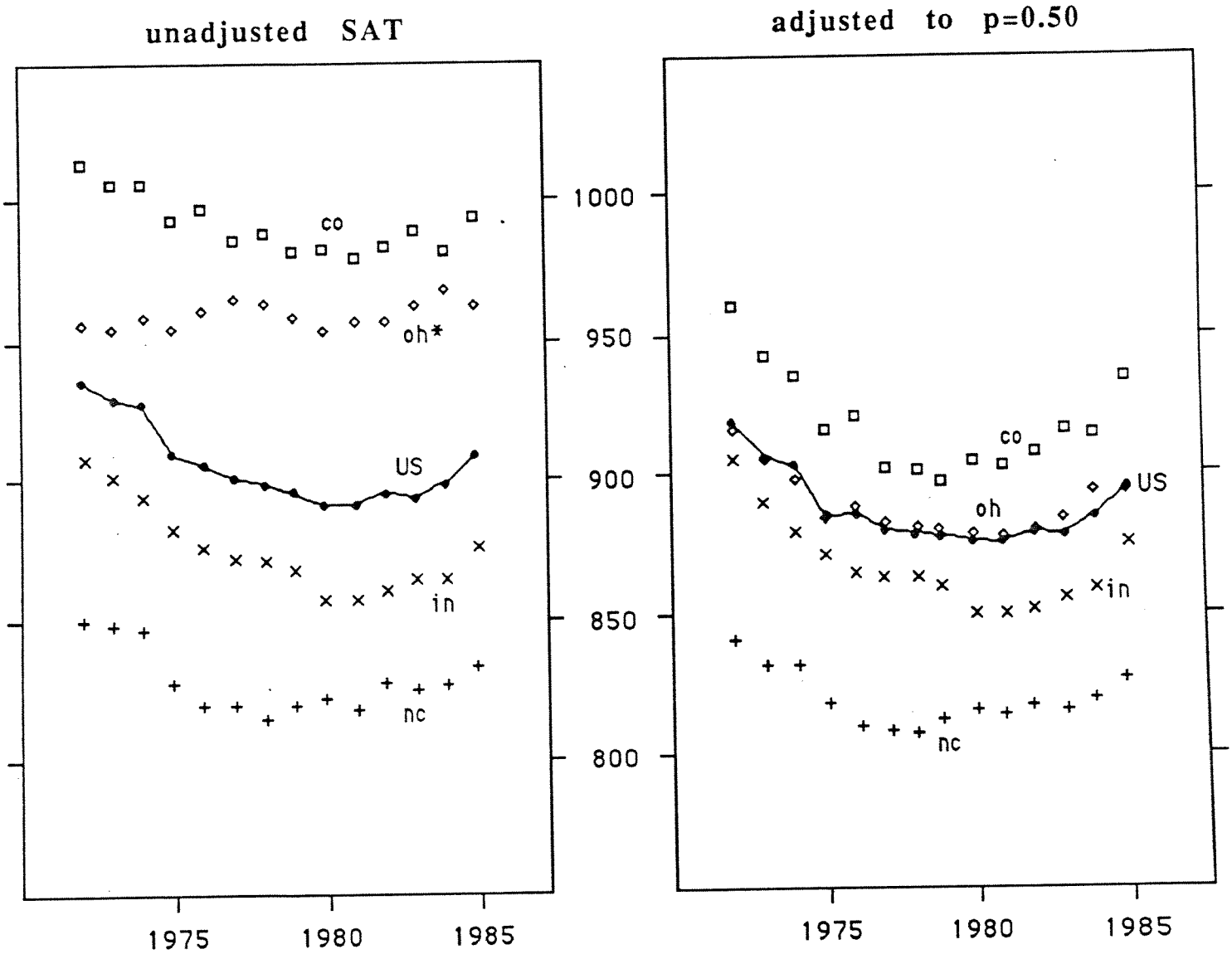


FIGURE 9

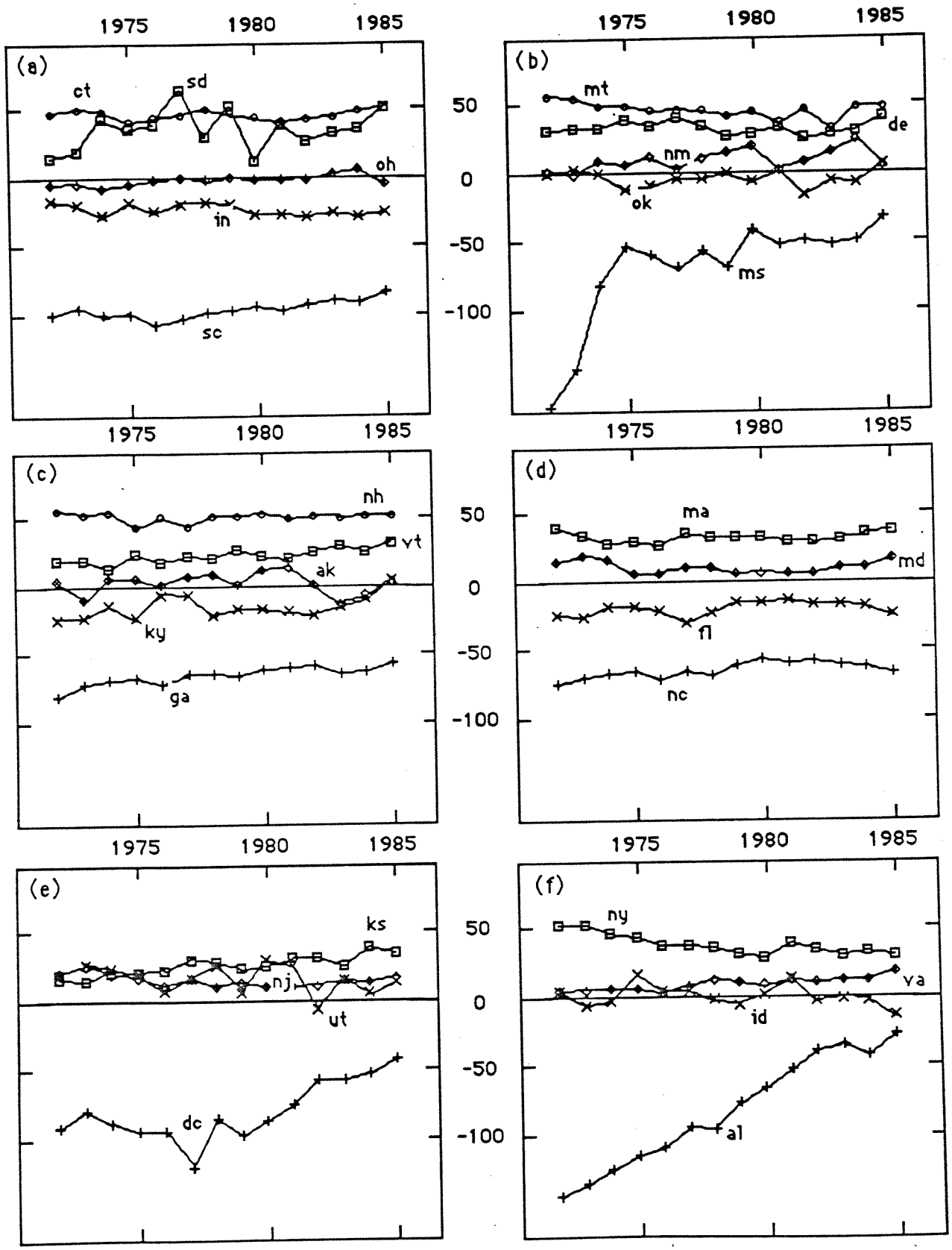
Unadjusted and adjusted mean SAT, 1972-85, for the U.S. and four selected states



*Ohio's participation rate decreased fairly steadily from 0.28 in 1972 to 0.16 in 1977

FIGURE 10

State deviations from national mean, 1972-85 (adjusted scores)



(Figure 10 part 2)

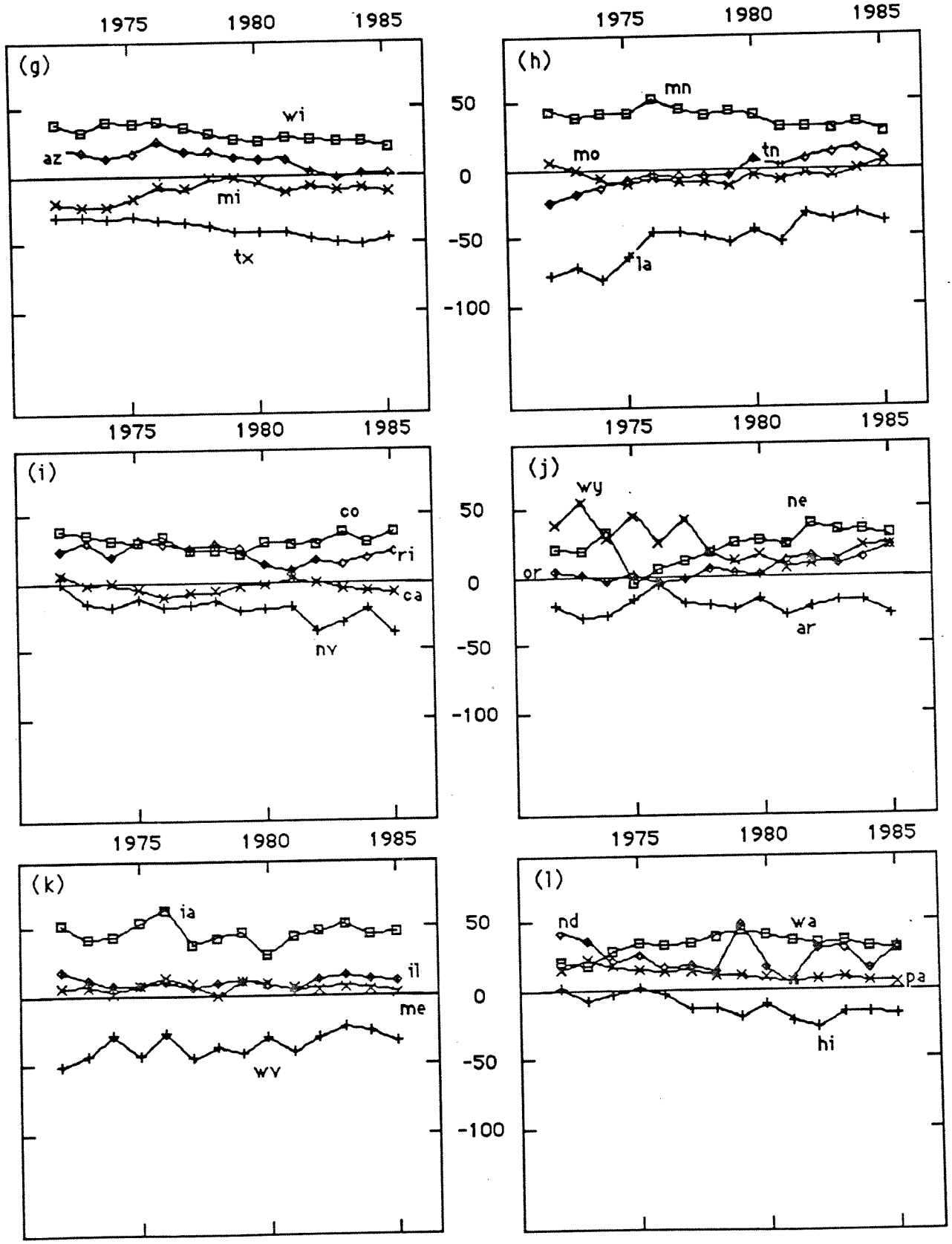


FIGURE 11

Conditional densities of test-taker abilities under fuzzy boundaries: example for p=.15