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

TECHNICAL REPORT NO. 876

November 1989

**Comments on: Holland and Wainer's,
"Sources of uncertainty often
ignored in adjusting state mean
SAT scores for differential
participation rates: the rules of the game"**

by

Don Edwards and Cynthia B. Cummings*

To appear, Applied Measurement in Education.

*Don Edwards is Associate Professor, Department of Statistics, and
Cynthia Beckworth Cummings is Statistical Computing Consultant, College of
Business Administration, University of South Carolina, Columbia, SC 29208.
This work was completed while the first author was Visiting Scientist, School of
Operations Research and Industrial Engineering, Cornell University.

I. Introduction: take it easy, we're on your side

Howard Wainer and Paul Holland, with colleagues at the Educational Testing Service (ETS) and elsewhere, have in the eighties assumed the difficult and sometimes unpopular role of statistical "police". The educational literature has allowed several articles using multiple regression in the absence of any kind of sampling model to "correct" state SAT averages. Should the educational community accept seniors taken from the local high school parking lot as representative of all of a state's seniors? Certainly it would not, but it would probably be safer to do that than to consider the SAT test-takers as representative in the sense of a random sample. Simply throwing participation rate p (or p^2 , or $\sqrt{p}$, or any function of p designed just to provide a good empirical fit) into the multiple regression as a kind of "self-selection band-aid" doesn't make it correct, even if a good R-square results and the observed patterns make sense. In the absence of a decent approximation to a random sample, it is the scientists' responsibility to determine, explicitly state, and to some extent justify a selection model and the statistical methods used under that model. If they do not, it is the referee's job to reject the paper. Since a number of educational scientists and referees have apparently been falling short in these respects, the responsibility has been appropriately assumed by ETS statisticians.

The paper of Taube and Linden, 1989 (T&L) has come closer to satisfying the above ideals than most. They assume and state a reasonable truncated normal selection model, summarized by the indented statements 1 and 2 of §III in Holland and Wainer, 1989 (H&W). They do not, however, state an implicit assumption or two,

falling short of a formal justification of their initial corrections (those based on a simple linear regression of state test-taker means SAT_i on transformed participation rates $z^*(p_i)$; the multiple regression on demographics, if done at all, should be done using the adjusted means as dependent variable). H&W have done a difficult job of filling in some missing pieces, identifying and criticizing *one* set of assumptions that would justify this regression: that the quantities they call $(\sigma_i - \sigma)z^*(p_i)$ are negligibly close to 0, and that the differences they call $\delta_i = (\mu_i - \mu)$ are random variables uncorrelated with the $z^*(p_i)$'s (this latter is verbally discussed by T&L also).

However, the model so convincingly rejected by H&W is only one of a family of truncated normal models. It is not the one discussed in the cited 15-minute presentation (Edwards and Beckworth, 1989: E&B), nor in the master's thesis on which the presentation was based (Beckworth, 1988), nor is it any of the several models defined and explored in the technical report supporting the discussion here (Edwards and Cummings, 1989: E&C). The single most important difference can be quickly stated in terms of the classical testing theory model: fixing a state and year, if a senior could be chosen at random, and if he/she takes the test with appropriate motivation,

$$(\text{measured SAT score}) = (\text{true score } T) + (\text{test error } \varepsilon).$$

The model which H&W discuss and reject assumes a normal distribution and sharply-truncated-normal selection on the *measured SAT scores*; the models discussed in E&C and its predecessors (E&B and Beckworth, 1988) assume a normal distribution and some sort of truncated normal selection for *true*

scores T (called "ability" in E&C). Moreover, these models assume nothing about test error except that it has mean 0 ($E(\epsilon)=0$). Consequently, a wide variety of test-taker score distributions and variance patterns are possible under these models simply by considering different distributions on the ϵ 's. In addition, the models have evolved (in part as a reaction to H&W, which was completed before E&C) from the one discussed in E&B to relax the sharp cutoff of the truncated normal distribution. In §II below we describe our (currently) favorite selection model and show that the patterns observed in H&W are not really inconsistent with it.

II. A more flexible model

This is equivalent to the fuzzily-truncated normal model of §6 in E&C. For a randomly selected student in a given state and year, let T and ϵ be as above. The unknowns μ and σ are now the mean and standard deviation of the true score T , which is assumed normally distributed, and the goal at present is simply to consistently estimate μ . This student must decide whether to take the test; let $R=1$ if he/she does, 0 otherwise. The probability that the test is taken will increase with the student's ability (i.e. true score), $P(R=1|T=t) = r(t)$ for a nondecreasing selection function $r(t)$. A natural parametric form for $r(t)$ in this setting, similar to the logistic function of §VI of H&W, but handier mathematically, is the Probit function

$$r(t) = \Phi[(t-\alpha)/\beta], \quad -\infty < t < \infty,$$

where Φ is the standard normal distribution function and β , α are unknown constants. It will be convenient to reexpress β relative to σ as $\rho = [\sigma^2/(\beta^2+\sigma^2)]^{1/2}$, $0 \leq \rho \leq 1$. Also, if a statewide proportion p of all

students take the test, this implies $E[R]=p$, which results in a constraint: $\alpha = \mu + z_p \sigma / \rho$.

The quantity ρ (not to be confused with p) is a "fuzzy boundary" parameter. Taking $\rho \rightarrow 1$ (i.e. $\beta \rightarrow 0$) under the above constraint on α , the selection function $r(t)$ reduces to the "sharp-boundary" selection function studied in Beckworth, 1988, E&B, and §3 of E&C:

$$r(t) = \begin{cases} 1 & t > \mu + \sigma z_p \\ 0 & \text{otherwise.} \end{cases}$$

This would say that the test-takers are exactly the top $p \times 100\%$ students in terms of true score T . If $\rho < 1$, the selection is fuzzy in the sense that some students with true score somewhat less than the boundary (the upper p th quantile of the true score distribution, $\mu + \sigma z_p$) are taking the test, and some with true score somewhat above this boundary are not. This is more realistic than sharp boundaries (but also more complicated). When ρ decreases to 0, the selection becomes so fuzzy that it is independent of ability: the selection model is equivalent to a $p \times 100\%$ random sample of all seniors. The quantity $(1-p) \times 100\%$ is referred to here as "percent fuzziness" in the ability selection model.

If the test-takers select themselves according to such a model, the mean of the state test takers' scores, SAT, estimates not the target μ but something larger:

$$E(\text{SAT}) \equiv \mu + \rho \sigma z^*(p) = E(T | R=1) \quad (1)$$

(here and below, the " $\equiv$ " stems entirely from the annoyance that the large "n" in the denominator of the state average is technically a random variable; note: E&C use the notation $B(p)$ for $z^*(p)$). One's first impulse from (1) is to try to find an estimate for $(\rho\sigma)$, inserting this into (1) and solving for an estimate of μ . Thus far, limited attempts at this have not been successful (perhaps due to violation of (i)-(iv) below).

Alternatively, we could write, for states $i=1,\dots,51$ (each having its own ρ_i, μ_i , etc.),

$$SAT_i \equiv \mu + \sigma_i \rho_i z^*(p_i) + \delta_i + \varepsilon_i^* \quad (2)$$

where $\delta_i = \mu_i - \mu$, and the ε_i^* have mean 0. *If* it is the case that the products $\sigma_i \rho_i$ are approximately constant over i (this means that states with more variable true scores have fuzzier selection boundaries; it would also of course be true if all σ_i 's, and all ρ_i 's, were approximately equal), *and* that the differences δ_i considered as random variables have mean 0 given p (which would hold if the δ_i 's were uncorrelated with the p_i 's), then a simple linear regression of the state SAT's on the transformed proportions $z^*(p)$ gives residuals which are unbiased predictors of the quantities $\delta_i = \mu_i - \mu$.

We are not completely comfortable with the last two assumptions; hence in E&C the corrections (using robust regression, and adjusting to 50% participation) are referred to as "first-order bias corrections" for the state SAT means. There are some reasons, though, to believe that the adjusted means are close to the truth in most cases, considering: (a) The clean linear relationship of the state SAT's to the transformed proportions $z^*(p)$, as seen in Figure 1 for 1982; (b) regional consistency in the adjusted means (notice e.g. from

Figure 1 that the "Deep South" states will have nearly equal residuals, as will the "Mideast" states; put the border states Virginia, Kentucky, and Tennessee in the Mideast); (c) stability of states' adjusted means over time (1972-85; see §5 of E&C), despite changes in participation rate p and corresponding changes in unadjusted means, with a few exceptions; (d) logical patterns in the adjusted means vs. demographics like state expenditures per student, literacy rates of parents, etc; these patterns are absent from the unadjusted means.

We do speculate that this method slightly undercorrects the very lowest-participation-rate states. There may also be some exceptional states. For example, in California, the state universities require the SAT for admission only if high school GPA is under 2.0; we would not be surprised to find a bimodal true score distribution in these test takers, which would be impossible under even the fuzzy boundaries selection model. It is also important to note that for the dozen or so states with smallest numbers of test takers (these are also smallest in participation rates), the adjusted scores may be nearly unbiased but probably have high variance since they are somewhat unstable over time. Twenty point swings from year to year are not uncommon for these (this would be the case for these states under any attempt at bias correction, and could be statistically repaired in a number of reasonable ways).

There are several important non-assumptions in the model above. For example, it does not assume:

- (i) test taker true scores $T|R=1$ are independent (this could be dangerous; test takers come in clusters in space and time)
- (ii) test errors ϵ have equal variance σ_ϵ^2 (on any test involving some guessing, the test error variance would probably decrease with increasing true score: there will be fewer questions requiring guessing. Also, test error variance must decrease as true score approaches either of the truncation points 400 and 1600).
- (iii) test errors are independent, or even uncorrelated, with each other and with true scores.
- (iv) test errors are normally distributed.

Assumptions (i)-(iii), or variations on them, are necessary to make any model predictions about patterns in test taker variances (which in normal theory are known to be extremely unrobust, anyway- much less robust than inferences on means). Assumptions (i)-(iv), or variations on them, are necessary to make any model predictions about patterns in the frequency distributions of observed scores, or to do careful estimation of σ and ρ . Thus, even under the model criticized by H&W (which could be considered a special case of the one above with $\sigma_\epsilon^2 = 0$ and $\rho = 1$) the patterns found in the test taker variances and score distributions do not technically reject the model as it applies to means; these peculiarities could be due to errors in assumptions (i)-(iv).

Finally, even if (i)-(iv) are acceptable, it's not hard to find cases under the more general model which are consistent with the H&W variance and histogram patterns. Assuming the model described

above, and (i)-(iii), the test taker variance from state i , S_i^2 , has expectation

$$E(S_i^2) \equiv \sigma_\epsilon^2 + \sigma_i^2 - \sigma_i^2 \rho_i^2 B_2(p_i) \quad (3)$$

where $B_2(p) = z^*(p)[z^*(p) - z_p]$. This checks with (9) of H&W if $\sigma_\epsilon^2 = 0$ and $\rho_i = 1$, taking the square root. From this (and the fact that $0.1 < B_2(p) < 0.6$ for $0.02 < p < 0.8$) we can see that fuzzy boundaries (small ρ_i 's) dampen the dependence of the variances S_i^2 on participation rates. Also, even if the $\sigma_i \rho_i$ are constant (at approximately 100, as all 14 years of regressions suggest), the test-taker variance pattern is undetermined. Even giving the further reasonable assumptions that the state variances are approximately equal (to a common σ^2), and that the error variance is relatively small, say $\sigma_\epsilon^2 = (0.1)\sigma^2$, the reasonable choice $\sigma = 200$, $\rho = 0.5$ is consistent with the standard deviation patterns observed by H&W; it would predict test-taker standard deviations S_i essentially constant, in the range [195,207], whereas the observed S_i range on [175,223]. The wider actual range would be due to the variability in these S_i 's about their expectations.

As for entire test score distributions, the lower truncation point at 400 is an annoyance that still needs to be dealt with, but the rounded, approximately symmetrical appearance of the distributions in H&W's Figure 2 can occur with even small amounts of test error and/or fuzziness in the selection function. For truly masochistic formula buffs, under the general truncated normal selection model, and adding (i)-(iv), the density function of an observed test score can be seen to be:

$$h(s) = (1/p)\Phi\{c(s)\}(\psi/\sigma)\phi[\psi(s-\mu)/\sigma], \quad -\infty < s < \infty,$$

where

$$c(s) = [\psi^2(s-\mu) - z_p\sigma/\rho]/\sigma[\rho^{-2} - \psi^2]^{1/2},$$

and ϕ the standard normal density, and ψ^2 is the "true" reliability, $\psi^2 = \sigma^2/(\sigma^2 + \sigma_\epsilon^2)$. Figure 2 shows examples of these functions, for a 15% participation-rate state whose true score mean and standard deviation are $\mu=800$, $\sigma=200$. The left column assumes no test error, with percent fuzziness in the selection boundary 0%, 5%, and 50%. The right column shows the corresponding distributions with 10% test error. Of course, (a) corresponds to the special case discussed by H&W. Note how far to the right of the parent mean $\mu=800$ each of these score distributions lies. It's also interesting to note the similarity between cases (b) and (c) and cases (e) and (f). To some this may suggest that these models are "too flexible" and that the true parameter values will be difficult to determine. Certainly we anticipate the equivalent of collinearity between ρ and ψ^2 , but what these visual similarities really mean is that eyeballing frequency distributions is not a very efficient way to evaluate goodness of fit of a model. After all, the naked eye sees little difference in separately plotted normal and Cauchy density functions (the Cauchy distribution is a t-distribution with 1 df, notoriously unstable).

A passing note: The true reliability ψ^2 should be somewhat smaller than reliability as measured by Cronbach's method; the above test error variance σ_ϵ^2 includes "low frequency" sources of variation like illnesses, lack of sleep, and emotional trauma whose disturbances should be considered noise but will not be measured by one-test-per-subject, one-day-per-test methods.

III. Where to from here?

These models are interesting, and they are promising, but they are not yet ready for use as anything but a very crude adjustment. The global assumptions of equality of the products $\sigma_i \rho_i$ and uncorrelated δ_i 's and $z^*(p_i)$'s, though they may be nearly correct, are "cross your fingers and pray" assumptions that should not be necessary with so much state-level data. They are surely at least a little wrong, and since the state-level μ_i 's and σ_i 's change very slowly, they will stay "a little wrong" for a long time. Standard errors for the estimates are needed, as well.

The arguments put forth by H&W are however not sufficient to reject the more general form of the truncated normal model discussed above; the motivation for our comments here is that it is extremely important that these ideas not be dismissed as useless based on the H&W criticisms. In respect to them, it should be added that our own understanding of the problem and search for a selection model was largely initiated by the "five pitfalls" paper (Wainer, 1986b). Also, innumerable gaps in our knowledge have been filled through phone conversations with Howard Wainer and a visit to ETS, and enhancements to the model have resulted.

Even if NAEP eventually nullifies the application of this methodology to SAT comparisons between states and over time (it is, after all, a random sample survey, and will provide a more complete measure of school system performance), if improved, the methodology could have broad application in other settings where

self-selection occurs in a similar way. It is very important that these ideas be studied further.

IV. Additional references

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.

Edwards, Don and Cummings, Cynthia B. (1989). State SAT scores: a model-based exploration. Technical Report #875, School of Operations Research and Industrial Engineering, Cornell University.

FIGURE 1
State SAT averages vs. transformed
participation rates, 1982

FIGURE 2
Examples of test score density functions,
 $p=0.15$, $\mu=800$, $\sigma=200$

**a) sharp selection
boundary, no test error**

**b) sharp selection
boundary, 10% test error**

**c) 5% fuzzy selection
boundary, no test error**

**d) 5% fuzzy selection
boundary, 10% test error**

**e) 50% fuzzy selection
boundary, no test error**

**f) 50% fuzzy selection
boundary, 10% test error**

400 800 1200 SAT 1600

400 800 1200 SAT 1600