ACT Research Report Series 87-11

Effects of Item Difficulty Heterogeneity on the Estimation of True-Score and Classification Consistency

Judith A. Spray
Catherine J. Welch

September 1987

For additional copies write:
ACT Research Report Series
P.O. Box 168
Iowa City, Iowa 52243

EFFECTS OF ITEM DIFFICULTY HETEROGENEITY ON THE ESTIMATION OF TRUE-SCORE AND CLASSIFICATION CONSISTENCY

Judith A. Spray

Catherine J. Welch

ABSTRACT

The purpose of this study was to examine the effect that large within-examinee item difficulty variability had on estimates of the proportion of consistent classification of examinees into mastery categories over two test administrations. The classification consistency estimate was based on a single test administration from an estimation procedure suggested by Subkoviak (1976). Analyses of both actual and simulated data revealed that the use of a single, overall-item difficulty estimate for an examinee's true-score, even when item difficulty varied greatly within an examinee, did not influence the estimation of the proportion of consistent classifications any more than homogeneous difficulty situations.

Effects of Item Difficulty Heterogeneity on the Estimation of True-Score and Classification Consistency

Methods of estimating the consistency of classification into two or more categories over two testing occasions but using information gained from only a single test administration have been proposed (Huynh, 1976; Marshall & Haertel, 1975; Subkoviak, 1976). All of these methods of estimating classification consistency (CC) were originally conceived to be used on criterion- or domain-referenced tests which were tests of fairly short length (i.e., 30 items or less) consisting of items assumed to measure somewhat narrowly defined content areas. The latter assumption is frequently thought to carry with it an assumption of approximate equal item difficulties or item "exchangeability" throughout the test for a given examinee. In fact all of the previously cited methods of estimating CC require this assumption.

However, tests which are not constructed to these specifications can still be used to classify individuals into categories. The ACT Proficiency Examination Program (ACT PEP) tests are of this type. These exams are designed to measure subject matter proficiency attained primarily outside of the classroom in "on-the-job" situations, such as nursing career experiences. Exams of this type cover a number of subject matter categories that are related more by job performance criteria than similarity of content. Because of the diversity of content included, the tests also report results on more homogeneous subcategories of items. The ACT PEP test called Fundamentals of Nursing is a typical representative of such proficiency tests. It consists of item sets from six subject matter categories ranging in length from only 8 items to 52 items with the average length around 20 items. The test specifications require that the items within a category item set fall under the broad content heading but the items within a set are not thought of as exchangeable, in the sense that item difficulty is expected to vary. Item sets greater than 10 items are used to classify examinees into one of three categories: below minimum, minimum, and above minimum proficiency.

When estimates of the consistency of classification into the three categories for these sets of items were required, test-retest administrations were considered but rejected because of several reasons. First, the learning effect, even over a fairly short period of time between administrations could contaminate the results. A second consideration was the expense of setting up such a dual administration. Finally, since there were many ACT PEP tests to consider, no one test could be totally representative of all the exams, and test-retest administrations for all of the ACT PEP tests were out of the question.

Therefore, an approximation method based on only one test administration was selected, and the method chosen is one proposed by Subkoviak (1976). This method provided a direct procedure to test the assumption that test items for a given examinee must be homogeneous in difficulty in order to use certain true-score estimates. Using this procedure also allowed direct comparisons with some of the results obtained in a simulation study by Algina and Noe (1978), in which the authors evaluated, among other things, the effects of using two different estimators of examinee true-score, or item difficulty on CC estimation. The simulated responses in their study were generated from homogeneous item difficulties for examinees of the same ability for n-item tests (n = 5, 10 and 20), and the CC estimation followed the Subkoviak (1976) procedure.

The following section outlines the Subkoviak procedure for estimating the proportion of examinees who would be classified consistently into two or more categories on two test administrations, given only one test administration. The section provides the assumptions and definitions of the procedure.

Subkoviak's Method of CC Estimation

The Subkoviak (1976) procedure approximates the coefficient of agreement or proportion of consistent classification (P_c) into **q** categories over two testing occasions for each of **N** examinees, $P_{c(i)}$, and then defines P_c by

$$P_c = \frac{1}{N} \sum_{i=1}^{N} P_{c(i)} \cdot \quad (1)$$

The ith examinee, with test-retest scores X_i and X_i' would consistently be classified into **q** categories over 2 occasions with the use of q-1 criterion or cutoff scores, C_1, C_2, ..., C_{q-1}, $P_{c(i)}$ X 100% of the time, where

$$P_{c(i)} = P(X_i \geq C_1, X_i' \geq C_1) + P(C_1 > X_i \geq C_2, C_1 > X_i' \geq C_2)$$

$$+ \ldots + P(X_i < C_{q-1}, X_i' < C_{q-1}). \quad (2)$$

In order to estimate $P_{c(i)}$, i = 1, ..., N and subsequently P_c, with only one test administration, it is necessary to impose two assumptions about the test yielding the scores, X_i and X_i'. These assumptions are that (1) test scores X_i and X_i' are independently distributed for each examinee and (2) X_i and X_i' are identically distributed random variables that follow the binomial distribution. These two assumptions allow (2) to be written as

$$P_{c(i)} = P(X_i \geq C_1) \cdot P(X_i' \geq C_1) + P(C_1 > X_i \geq C_2) \cdot P(C_1 > X_i' \geq C_2) + \ldots + P(X_i < C_{q-1}) \cdot P(X_i' < C_{q-1})$$

and

$$P_{c(i)} = \left[P(X_i \geq C_1)\right]^2 + \left[P(C_1 > X_i \geq C_2)\right]^2 + \ldots + \left[P(X_i < C_{q-1})\right]^2. \tag{3}$$

The binomial assumption provides the probability values in equation (3); for example,

$$P(X_i \geq C_j) = \sum_{X_i = C_j}^{n} \binom{n}{X_i} \pi_i^{X_i} (1 - \pi_i)^{n - X_i} \tag{4}$$

where, for an n-item test, person i has probability π_i of getting each item correct. This requires an assumption of equal item difficulty for an examinee for all n items.

The primary obstacle in obtaining approximations of $P_{c(i)}$ is the estimation of the item true-score, π_i. Subkoviak (1976) originally suggested two estimators of π_i: (1) $\hat{p}_i = X_i/n$, the proportion of test items correctly answered by the ith examinee and (2) $\tilde{p}_{21_i}$, a linear regression estimate, given by

$$\tilde{p}_{21_i} = \hat{\alpha}_{21}\left(\frac{X_i}{n}\right) + \left(1 - \hat{\alpha}_{21}\right)\frac{\bar{X}.}{n} \tag{5}$$

where $\hat{\alpha}_{21}$ is the Kuder-Richardson formula 21 reliability estimate. Subkoviak (1976) suggested that $\bar{p}_{21_i}$ is probably better for shorter tests (e.g., $n < 40$ items) since it uses "collateral information" to supplement the estimate when n is small. However, he also suggested using $\hat{\alpha}_{20}$ in place of $\hat{\alpha}_{21}$ in equation (5) for those instances where the assumption of equal item difficulties for an examinee may be violated. This estimator of π_i is labeled $\tilde{p}_{20_i}$ in the remainder of this paper. Algina and Noe (1978) determined that $\tilde{p}_{20}$ performed better than $\hat{p}$ under many simulated testing conditions.

Estimation of π_i Under a Compound Binomial Assumption

Although the employment of $\tilde{p}_{20_i}$ makes use of the possibility that the n item difficulties could be unequal, a <u>single</u> estimate for all item difficulties still must be used in equation (4). However, if it were possible to obtain item difficulties for each item for an examinee, π_{ij}, j = 1, ..., n, then equation (4) could be written using a compound binomial expression. We can define a random variable, U_{ij}, for the ith examinee and the jth item such that

$$u_{ij} = \begin{cases} 1, \text{ if the item is correctly answered} \\ 0, \text{ otherwise.} \end{cases}$$

A response vector, $\underset{\sim}{v}_i$ for the ith examinee is

$$\underset{\sim}{v}_i = (u_{i1}, u_{i2}, \ldots, u_{in})'$$

with success probability vector,

$$\underset{\sim}{\pi}_i = (\pi_{i1}, \pi_{i2}, \ldots, \pi_{in})',$$

and

$$P(\underset{\sim}{v}_i | \underset{\sim}{\pi}_i) = \prod_{j=1}^{n} (\pi_{ij})^{u_{ij}} (1 - \pi_{ij})^{1-u_{ij}}. \tag{6}$$

Equation (6), however, is only the probability of observing one response vector, $\underset{\sim}{v}_i$, and in using equation (6) to evaluate $P_{c(i)}$ in equation (3), one would have to evaluate (6) for all response vectors that belonged to sets S_1, S_2, ..., S_{q-1} such that $S_1 = \{X_i | X_i \geq C_1\}$, $S_2 = \{X_i | C_1 > X_i \geq C_2\}$, ..., $S_{q-1} = \{X_i | X_i < C_{q-1}\}$

$$\text{where } \sum_{j=1}^{n} u_{ij} = X_i .$$

For tests of any length n, there are 2^n possible response vectors, and this could prove to be a formidable task, even for fairly short tests.

Fortunately, Lord and Novick (1968) have provided a compound binomial expansion approximation for the evaluation of $P(X_i = x_i | \pi_{ij})$ (see p. 525). Therefore, if estimates of π_{ij} can be obtained, it is possible to evaluate equation (3) from these estimates and this expansion approximation, truncated to some reasonable number of terms.

Logistic Estimates of π_{ij}

Estimates of π_{ij} can be obtained by fitting the u_{ij} responses to a logistic function of the form

$$P(U_{ij} = u_{ij} = 1 | \beta_{oj}, \beta_{1j}, X_i) =$$

$$\exp(\beta_{oj} + \beta_{1j}X_i)/\{1 + \exp(\beta_{oj} + \beta_{1j}X_i)\} \quad (7)$$

and obtaining estimates of β_o and β_1. Such a procedure is usually referred to as a logistic regression analysis. Maximum likelihood procedures can be used to estimate the coefficients β_{oj} and β_{1j} for each test item and the overall goodness-of-fit of the model usually can be checked via a likelihood-ratio chi-square procedure. Such an analysis is provided by the statistical computing package, BMDP, by program LR (BMDP, 1983).

The logistic regression (LR) coefficients act as "pseudo-IRT" parameters in the sense that if the exponential term in (7) is rewritten as $\beta_{1j}(X_i + \beta_{oj}/\beta_{1j})$, the coefficients estimate item difficulty and item discrimination, albeit on the raw score scale, X, rather than on some trait scale. However, since the ultimate goal of this estimation procedure is to estimate P_c, and since P_c is based on raw test scores, this is not viewed as a serious problem.[1]

The purpose of this study was to apply the LR procedure to item responses from tests that were originally constructed to produce a wide range of within-examinee item difficulty and to use these estimates of π_{ij} to estimate P_c via the Subkoviak procedure modified to fit a compound binomial model. These estimates of P_c could then be compared to those obtained using $\hat{p}$ and $\tilde{p}_{20}$ estimates. A secondary purpose was to simulate test data that followed the degree of item difficulty heterogeneity exhibited in the real test data and to observe the behavior of P_c as estimated using $\hat{p}$ and $\tilde{p}_{20}$, much as those simulations performed by Algina and Noe (1978).

Method - Part 1

We selected an ACT PEP exam that provided the widest possible range in item set size, from 8 to 52 items. The total test length is 124 items. Estimates of P_c based on $\hat{p}$ and $\tilde{p}_{20}$ were obtained for all six item sets for 509 examinees. However, for purposes of this study, only the smallest (8 items) and the largest (52 items) sets were used for the LR analysis. This was done primarily because of the cost of fitting the LR to all six content sets. The 8- and 52-item sets were chosen in order to study the effect of set length on the estimates.

Observed overall proportion correct values ranged from .55 to .97 for the 8-item set and .50 to .96 for the 52-item set. The mean difficulty and standard deviation, in terms of overall proportion correct values were .74 (S.D. = .15) and .79 (S.D. = .12) for the 8- and 52-item sets, respectively.

In practice the 8-item set is considered too short to use for classification purposes, and an examinee's performance on this set only serves to help make an overall assignment to one of two categories, pass or fail, based on the total test score. For purposes of this study, we arbitrarily looked at a single, hypothetical cutoff score of 5 for this set. The actual cutoff scores of 41 and 37 were used on the 52-item set (i.e., $X \geq 41$ implied above minimum proficiency; $37 \leq X < 41$ implied minimum proficiency; and $X < 37$ implied below minimum proficiency).

The two estimators, $\hat{p}$ and $\tilde{p}_{20}$, were used for the 509 examinees on each set to estimate the CC cell probabilities from equation (3). The sum of the cells along the diagonal gave the estimates of P_c.

For the LR analysis, the actual item responses to the 8- and 52-item sets were used to fit the logistic function given by equation (7). In order to obtain a wider range of test scores for the regression, we used an examinee's total test score (out of 124 items) minus that item score being "fit" to the LR model (i.e., we used a transformed test score, $X_i^* = X_i - u_{ij}$). This was especially important for the LR analysis of the smaller set, as an 8-item test score gave a rather restricted range for the independent variable.

For each item response set, the goodness-of-fit of equation (7) was checked using a likelihood-ratio chi-square statistic (see BMDP, program LR, 1983). Lack of fit was determined to be any LR analysis that produced a significant chi-square value of $p < .05/60$ or $< .0008$, since 60 items, in total, were fit. The smallest p-value from the 60 items was .004.

The coefficient estimates, $\hat{\beta}_{oj}$ and $\hat{\beta}_{ij}$, were then used in equation (7) to obtain an LR estimate of π_{ij}, given X_i^*. These estimates were used in the Lord and Novick (1968) approximation to the compound binomial expansion, truncated after 5 terms. These estimates were then used in equation (3) to calculate the cell probabilities as before.

Results - Part 1

Tables 1 and 2 give the results of the three different estimates of the CC cell probabilities for the 8- and 52-item subtests, respectively. These preliminary results indicated that, if the P_c estimates from the LR analysis procedure were considered as the "true" parameters, then the estimates of P_c from $\hat{p}$ and $\tilde{p}_{20}$ (abbreviated as $\hat{P}_c$ and $\tilde{P}_c$, respectively) were not that much different from the results reported by Algina and Noe (1978) on test data with much smaller item difficulty variability. Their findings indicated that, for 5-, 10- and 20-item tests, $\tilde{p}_{20}$ yielded estimates of P_c that were less biased than $\hat{p}$, especially near cutoff scores close to the mean true-score of the test. Furthermore, near the mean true-score, $\hat{P}_c > P_c > \tilde{P}_c$, which occurred in the analysis of the 52-item set of the present study. As Table 1 shows, all three estimates of P_c were fairly close for the 8-item set, the greatest difference occurring between the LR estimate and the $\tilde{p}_{20}$ estimate (.028). The greatest difference for the 52-item set was .073 ($\hat{P}_c - \tilde{P}_c$). In neither case were the differences so great as to warrant the use of $\hat{\pi}_{ij}$ in estimating P_c. However, other cutoff scores had not been investigated. Therefore, a second part of this study was undertaken to see which estimator, $\hat{p}$ versus $\tilde{p}_{20}$, would perform better (i.e., have smaller bias and/or smaller variance) under conditions of large item difficulty variability, for a variety of cutoff scores on simulated responses. If neither one performed well under more varied conditions, this might indicate that the only way to achieve good estimates of P_c would be to use LR procedures on each test item.

Method - Part II

Since the results of the real data comparison seemed to indicate that the use of a single true-score estimate for all items might not make a difference in P_c estimation, we set up a data simulation to study the effect of true-score variability across test items on P_c estimation.

In order to simulate conditions of item difficulty or true-score variability for an examinee, we first analyzed the LR coefficients estimated in Part I of this study. We formed a "difficulty-like" parameter, $-\hat{\beta}_o/\hat{\beta}_1$, for each of the 60 items (sets 8 and 52), and selected those items which yielded the 10 largest and 10 smallest difficulty estimates. This gave us 20 LR models for a 20-item test with fairly large variability in difficulty. On the original sample of 509 examinees, these 20 items had a mean proportion correct value of .73 with standard deviation of .16. Table 3 shows this item difficulty variability for 5 examinees. The 20 items selected are listed by their original item numbers (1-60). The 5 examinees were those scoring (1) the minimum observed total test score, 57; (2) -1 S.D. from the total test scores mean, $\bar{X}$ - SD, 83; (3) $\bar{X}$, 94; (4) $\bar{X}$ + SD, 105; and (5) the maximum test score observed, 117. Table 3 gives each of these examinee's true probability correct score, π_{ij}, based on the LR models.

A random sample of size N = 100 was drawn from the original empirical distribution of 509 total test scores. The sample size was chosen so that the unimodal shape and other characteristics of the original empirical distribution would be retained in the sample. This sample of 100 scores remained constant throughout all replications. The sample mean was 94.7 with SD = 10.3, max (X) = 115 and min (X) = 64. The sample empirical distribution was also unimodal.

Using the LR models for the 20 items, it was possible to determine true-scores, π_{ij}, i = 1, 2, ., 100, j = 1, 2, ..., 20 from equation (7). Then, for a two-category classification problem, cutoff scores were varied from 9 to 19 (i.e., 9, 10, 11, ..., 19). Using the compound binomial approximation mentioned previously, equation (3) could then be used to give true P_c values at each of these cutoff scores.

Response data (0's and 1's) were generated from a uniform number generation subroutine (IMSL, 1985) to simulate a 20-item test for each of the 100 examinees for each cutoff score evaluation. Then estimates, $\hat{p}_i$ and $\tilde{p}_{20_i}$, were calculated for each examinee and P_c was estimated for all 11 possible cutoff scores.

Results - Part II

The results of the estimation of P_c using $\hat{p}_i$ and $\bar{p}_{20_i}$ are given in Table 4, in terms of mean deviation and standard deviation for each estimator. Again, these results were quite similar to those of Algina and Noe (1978) for a 20-item test. The $\tilde{P}_c$ estimator, in general, had smaller bias but greater variability than $\hat{P}_c$. The direction of the bias for $\tilde{P}_c$ was negative over the entire cutoff score range used. Algina and Noe also found that $\tilde{P}_c$ underestimated P_c, but only for cutoff scores near the mean true-score. The $\tilde{P}_c$ estimator had its greatest amount of negative bias at cutoff scores near the mean true-score in both studies. In the present study, however, $\tilde{P}_c$ had much larger bias, over twice that found by Algina and Noe. In contrast, the $\hat{P}_c$ estimator had positive bias at cutoff scores near the mean true-score (14.6), again a similar finding to the previous study, but the amount of bias for $\hat{P}_c$ was smaller than that found by Algina and Noe.

Summary and Conclusions

The violation of the assumption of equal item difficulty in the estimation of CC via the Subkoviak procedure does not appear to be serious. The amount and direction of bias in $\hat{P}_c$ and $\tilde{P}_c$ when items are not homogeneous appear to be quite similar to situations where the assumption is less seriously violated. When we compare our results to those of Algina and Noe (1978), we note that the main effect of increasing within-examinee item difficulty variability is to increase the size of the P_c coefficient, but only slightly. We deduce this from the amounts of bias exhibited by $\hat{P}_c$ and $\tilde{P}_c$: $\tilde{P}_c$ underestimated P_c for all cutoff scores analyzed and the amount of underestimation increased, especially near cutoff scores close to the mean true-score, while $\hat{P}_c$ showed smaller overestimation in this range. The conclusion must be that the true value of P_c is slightly larger when item difficulty varies within examinees, but that the value of P_c does <u>not</u> increase so much that estimates from $\tilde{P}_c$ and $\hat{P}_c$ are inaccurate. The use of $\tilde{P}_c$ still seems to be prudent since it will usually act as a good "lower bound" for the true parameter value. However, Algina and Noe's suggestion of averaging $\hat{P}_c$ and $\tilde{P}_c$ would appear to make sense in light of the findings presented in Table 4. In only one instance would the average value seriously increase the amount of bias (see Table 4, cutoff score = 13), and in the cutoff score range from 14 to 17, the averaging of approximately equal amounts of positive and negative bias would produce fairly accurate estimates. Certainly the use of $\hat{P}_c$ and/or $\tilde{P}_c$ is easier and less expensive in terms of time and money than fitting each item to an item-response function such as the logistic curve.

REFERENCES

Algina, J., & Noe, M. J. (1978). A study of the accuracy of Subkoviak's single-administration estimate of the coefficient of agreement using two true-score estimates. Journal of Educational Measurement, 15, 101-110.

BMDP Statistical Software. (1983). Stepwise logistic regression [Computer program manual]. Berkley: BMDP, University of California Press. (BMDPLR, pp. 330-346).

Huynh, H. (1976). On the reliability of decisions in domain-referenced testing. Journal of Educational Measurement, 13, 253-264.

IMSL User's Manual. (1984). GGUBS: Basic uniform (0, 1) pseudo-random number generator. [Computer program manual]. Houston: IMSL.

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Marshall, J. L., & Haertel, E. H. (1975). A single-administration reliability index for criterion-referenced tests: The mean split-half coefficient of agreement. Paper presented at the annual meeting of the American Educational Research Association, Washington, D.C.

Subkoviak, M. (1976). Estimating reliability from a single administration of a criterion-referenced test. Journal of Educational Measurement, 13, 265-276.

Footnote

[1]We acknowledge that perhaps the best estimation procedure would be to estimate π_{ij} on a latent trait scale and then to evaluate equation (3) by converting $\sum_{j=1}^{n} \hat{\pi}_{ij}$ to an estimate of a number-correct true-score.

TABLE 1

CC Estimation for the 8-Item Set with $\bar{X}$ = 5.9, SD = 1.4, and $\hat{\alpha}_{20}$ = .31

Estimate used: $\hat{p}$, P_c = .790

	Category	'Retest' 1	'Retest' 2	Marginals
Test	1	.680	.105	.785
	2	.105	.110	.215

Estimate used: $\tilde{p}_{20}$, P_c = .777

	Category	'Retest' 1	'Retest' 2	Marginals
Test	1	.751	.112	.863
	2	.112	.026	.138

Estimate used: LR ($\hat{\pi}_{ij}$), P_c = .805

	Category	'Retest' 1	'Retest' 2	Marginals
Test	1	.770	.097	.867
	2	.097	.035	.132

TABLE 2

CC Estimation from the 52-Item Set with $\bar{X}$ = 41.0, SD = 6.0, and $\hat{\alpha}_{20}$ = .67

Estimate used: $\hat{p}$, P_c = .670

	Category	'Retest' 1	2	3	Marginals
	1	.470	.086	.025	.581
Test	2	.086	.074	.054	.214
	3	.025	.054	.126	.205

Estimate used: $\bar{p}_{20}$, P_c = .597

	Category	'Retest' 1	2	3	Marginals
	1	.437	.118	.031	.586
Test	2	.118	.094	.053	.265
	3	.031	.053	.066	.150

Estimate used: LR ($\hat{\pi}_{ij}$), P_c = .660

	Category	'Retest' 1	2	3	Marginals
	1	.503	.109	.018	.630
Test	2	.109	.081	.043	.233
	3	.018	.043	.076	.137

TABLE 3

Examples of π_{ij} scores for Five Examinees in the Simulation of a 20-Item Test

Item #	Min 57	X-SD 83	X 94	X+SD 105	Max 117
		Examinee's Total Test Score			
4	.15	.38	.51	.64	.76
7	.89	.92	.93	.94	.95
8	.18	.45	.59	.72	.82
10	.81	.89	.91	.93	.94
14	.31	.49	.57	.64	.72
15	.20	.51	.66	.78	.87
17	.37	.48	.52	.57	.61
20	.14	.50	.68	.82	.92
25	.69	.73	.75	.76	.78
27	.84	.91	.92	.94	.95
28	.74	.82	.85	.87	.90
30	.82	.85	.87	.88	.89
31	.68	.79	.82	.85	.88
37	.81	.88	.90	.92	.93
40	.92	.95	.96	.97	.98
46	.28	.51	.61	.70	.79
53	.15	.47	.63	.77	.87
54	.86	.96	.97	.99	.99
55	.36	.50	.55	.61	.67
57	.26	.49	.59	.68	.77
MEAN	.52	.67	.74	.80	.85
SD	(.29)	(.20)	(.16)	(.13)	(.10)

TABLE 4

Bias and Variability for Two Estimates of P_c: N = 100, n = 20

Parameter or Estimator	Cutoff Scores										
	9	10	11	12	13	14	15	16	17	18	19
True P_c	.995	.993	.961	.899	.804	.697	.619	.619	.709	.845	.953
$E\{\hat{P}_c - P_c\}$	-.031	-.055	-.067	-.057	-.024	.036	.082	.080	.031	-.040	-.067
$E\{\tilde{P}_c - P_c\}$	-.000	-.008	-.008	-.007	-.031	-.051	-.071	-.065	-.026	-.002	-.002
$\hat{\sigma}(\hat{P}_c)$	.008	.010	.013	.015	.016	.015	.014	.015	.012	.016	.012
$\hat{\sigma}(\tilde{P}_c)$	.003	.005	.012	.021	.026	.027	.022	.019	.022	.024	.011
$E(\hat{\alpha}_{20})$	.280	.255	.278	.272	.296	.279	.283	.267	.293	.271	.287