A FIVE-YEAR QUEST: IS COMPUTERIZED ADAPTIVE TESTING FEASIBLE?

VERN W. URRY
U.S. Civil Service Commission

Five years of research on the feasibility of computer assisted testing has attempted to answer four extremely significant questions: (1) What types of items are required for effective computerized adaptive testing? (2) Do these types of items exist in sufficient number to measure important abilities adequately? (3) Can estimates of the item parameters be obtained that are sufficiently reliable to be used successfully in a computerized adaptive testing algorithm? and (4) Is there an efficient and accurate adaptive algorithm for computerized testing?

In answer to the first question, "What types of items are required for effective computerized adaptive testing?", the development of specifications for effective item banks or item pools for computerized adaptive testing was begun about five years ago (Urry, 1970). These specifications were written with reference to the three parameters of the normal ogive model (Lord & Novick, 1968) and the logistic model (Birnbaum, 1968). At that time, they included requirements for a minimum of 100 items with item discriminatory powers (the a_i) of at least .80, with item difficulties (the b_i) evenly distributed on the interval from −2.00 to 2.00, and with item coefficients of guessing (the c_i) of .25 as a maximum. Some research was later completed (Jensema, 1974; Urry, 1974*b*) indicating that the maximum value for the c_i could be set as high as .30 with item bank effectiveness still maintained.

In these studies, an item bank was adjudged effective when computerized adaptive testing required fewer items than conventional paper and pencil testing to attain the same level of reliability. The specifications were arrived at through model sampling and simulation techniques. The concern was the capability of the 3-parameter models for the specific purpose of computerized adaptive testing. After model capabilities were adequately explored, there remained the empirical question, "Do these types of items exist in sufficient number to measure important abilities adequately?"

At first glance, it might have appeared that the requirement for item discriminatory powers of .8 or greater was unreasonably high given the usual test item because an item discriminatory power of .8 corresponds to a biserial correlation of .62 between the item and latent ability. In the experience of most psychometricians this would seem an impossible specification to meet, because the usual item-test biserial correlations tend to be much lower than this specified value. However, the impossibility exists only if the attenuating effects of guessing on conventional indicants of item discriminatory power are not fully understood. These effects mask the true discriminatory power of multiple-choice items to a marked degree, and they are still largely unappreciated.

In order to illustrate these effects, equations were derived for the point-biserial (Urry, 1974*a*) and the biserial (Urry, 1975) correlations between multiple-choice items and latent ability. The equation for the point-biserial correlation was derived as

$$\rho_{i'\theta} = \frac{(1 - c_i)\,\rho_{I\theta}\,\phi(\gamma_i)}{\sqrt{P_i'\,Q_i'}}$$

(Urry, 1974*a*, eq. 15); (1)

and the derivation of the biserial correlation resulted in

$$\rho_{I\theta}' = \frac{(1 - c_i)\,\rho_{I\theta}\,\phi(\gamma_i)}{\phi(\gamma_i')}$$

(Urry, 1975, eq. 6). (2)

In these equations, a prime was used to indicate that the given term was affected by guessing. Definitions were as follows:

- c_i the item coefficient of guessing, is the lower asymptote of the regression of the binary item on latent ability;
- $\rho_{I\theta}$ is the biserial correlation, unaffected by guessing, between the binary item and latent ability;
- γ_i is the baseline value of the item distribution $N(0,1)$ above which the probability of (or proportion) knowing the correct response occurs;
- $\phi(\gamma_i)$ is the height of the ordinate at γ_i;
- P_i' is the probability of (or proportion) passing a multiple-choice item;
- Q_i' or $1 - P_i'$, is the probability of (or proportion) missing a multiple-choice item;

γ_i' is the baseline value on the distribution $N(0,1)$ above which the probability of (or proportion) passing, viz. P_i', occurs:

$\phi(\gamma_i')$ is the height of the ordinate at γ_i'.

The difference between the probability of (or proportion) knowing the correct response to an item, viz.,

$$P_i = \frac{1}{\sqrt{2\pi}} \int_{\gamma_i}^{\infty} exp \left[\frac{-t^2}{2} \right] dt, \tag{3}$$

and the probability of (or proportion) passing a multiple-choice item, viz.,

$$P_i' = c_i + (1 - c_i) P_i, \tag{4}$$

is to be duly noted. As a consequence, it is known that γ_i is equal to γ_i' only when c_i is zero. When guessing is effective (or, synonomously, c_i is not zero), neither γ_i and γ_i' nor $\phi(\gamma_i)$ and $\phi(\gamma'_i)$ are equal. Further, when guessing is effective, γ_i', as a baseline value, is unlike γ_i which divides the item distribution meaningfully on the basis of success on the item. Notice that for c_i equal to zero, equation (2) indicates the equality of $\rho_{I\theta}'$ and $\rho_{I\theta}$. Otherwise the distinction between these two coefficients is to be kept clearly in mind. Since item discriminatory power is defined by the normal ogive model as

$$a_i \equiv \frac{\rho_{I\theta}}{\sqrt{1 - \rho_{I\theta}^2}}, \tag{5}$$

it is totally inappropriate to substitute estimates of $\rho_{I\theta}'$ for $\rho_{I\theta}$ in equation (5) to estimate a_i. When guessing is effective or when the items are of a multiple-choice variety, this procedural error adversely affects computerized adaptive testing.

The derived equations for the point-biserial and biserial correlations were used to illustrate the attenuating effects of guessing on these conventional indicants of item discriminatory power. In the procedure, the item coefficient of guessing is usually set at some meaningful value, say, the reciprocal of the number of alternatives for a multiple-choice question; and for this fixed value of c_i, the equations are evaluated to map the levels of a_i and b_i onto the planes defined by the coordinates, the point-biserial correlation and the p-value, or the biserial correlation and the p-value. In Figure 1, the levels of a, viz., .8, 1.0, 1.2, 1.4, 1.6, 2.0, and 3.0, and the levels of b, viz., 2.0, 1.6, . . . , -2.00, have been mapped onto the plane defined by the population point-biserial correlation and the population proportion passing or p-value for c equal to .20. When c is fixed at .20, the effectiveness of guessing is roughly equivalent to the level typical of 5-alternative items. Since the biserial correlation (unaffected by guessing) between the item and latent ability is defined as

$$\rho_{I\theta} \equiv \frac{a_i}{\sqrt{1 + a_i^2}} \tag{6}$$

in the normal ogive model, the levels of a portrayed in Figure 1, viz., .8, 1.0, 1.2, 1.4, 1.6, 2.0 and 3.0, correspond to item ability biserials of .62, .71, .77, .81, .85, .89, and .95. Notice then the apparent paradox. For example, an item which has an item-test point–biserial correlation of .11 with a p-value of .22 is indicated to have an item discriminatory power, a_i, of 3.00 or a $\rho_{I\theta}$ of .95. The astonishing paradox is due to the attenuating effect of guessing. In Figure 2, identical levels of a and b have been mapped onto the plane defined by the population biserial correlation and the population proportion passing or p-value, again, for c fixed at .20. While the attenuating effect is less pronounced for the biserial correlation relative to the point-biserial correlation, it is most severe for difficult items. For example, a five-alternative multiple-choice item with an item-test biserial correlation of .17 and a p-value of .22 is indicative of an item discriminatory power of 3.0 or an item-ability biserial of .95 and an item difficulty of 2.00. What would happen if the procedural

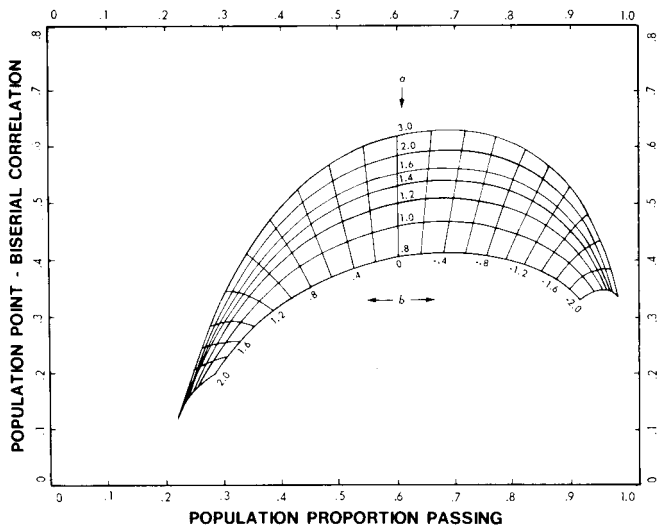

Figure 1. Relationship between conventional and normal ogive item parameters when the coefficient of guessing (c) equals .20.

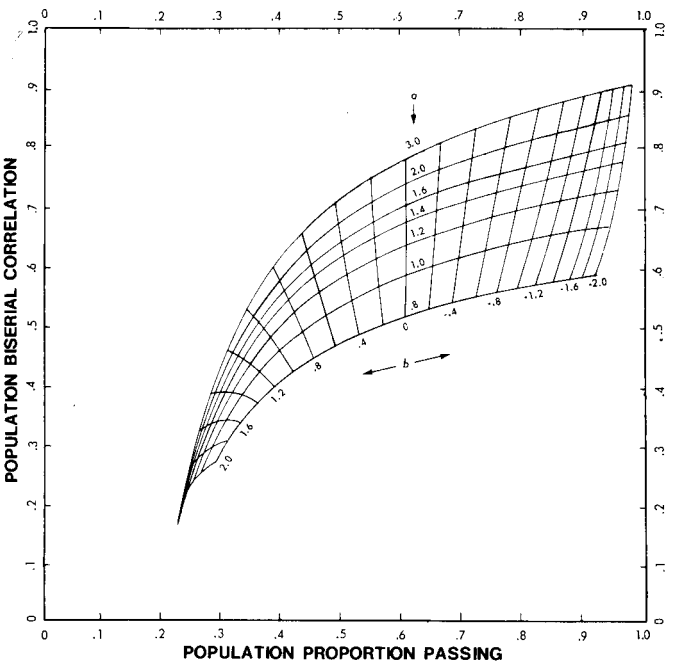

Figure 2. Relationship between conventional and normal ogive item parameters when the coefficient of guessing (c) equals .20.

error alluded to earlier were committed in connection with this interesting case? It will be recalled that the error involved the misuse of $\rho'_{I\theta}$ in equation (5). In this instance, a_i would have been erroneously estimated as .17 when the true value was 3.00. Obviously, gross errors of this nature render computerized adaptive testing less efficient than it should normally be. If the data point defined by the item-test point-biserial or biserial correlation and the p-value is plotted on one of these maps or charts, the corresponding values of a_i and b_i for the given item can be interpolated from the grid system that identifies the various levels of a_i and b_i. For reliable total tests[1] and large samples, the interpolated values of a_i and b_i approximate the true parameters and allow the researcher (1) to identify items appropriate for the purpose of computerized adaptive testing and (2) to assess the efficacy of a given set of appropriate items for the purpose of computerized adaptive testing by comparing the obtained interpolated values with the specifications for item bank effectiveness. When the specifications are met, improved reliability per item used is assured for computerized adaptive tests relative to conventional tests. However, the number of items required in computerized adaptive testing relative to conventional testing can be markedly reduced when the a_i appreciably exceed the minimum value of .80, the b_i are widely and evenly distributed, and the c_i are maintained at low values.

Experience has shown (Jensema, 1972; Urry, 1974*b*) that roughly one-third of the items in the usual aptitude or ability test survive this screening for appropriateness. Moreover, item discriminatory powers have been frequently found to exceed 2.0 in value.

After it was ascertained that sets of items could be found that would satisfy the specifications for effective item banks, there remained the important question, "Can estimates of the item parameters be obtained that are sufficiently reliable to be used successfully in a computerized adaptive testing algorithm?" In answer to this question, a relatively rapid and inexpensive item-analytic procedure was developed (Urry, in press–*a*). It has been programmed and is currently available for use on several computers. The output of the program is an item analysis yielding ancillary estimates for a_i, item discriminatory power; b_i, item difficulty; and c_i, item coefficient of guessing.

Estimates of the parameters a_i, b_i, and c_i are obtained by an iterative minimum χ-square procedure. The procedure consists of two stages that differ only with respect to the particular measure used for manifest ability. In the first stage, the distribution of manifest ability is represented by corrected raw scores where the item being parameterized is omitted from the scoring. In the second stage, the distribution of manifest ability is represented by Bayesian modal estimates of ability (Samejima, 1969). Generally, Bayesian modal estimates of ability more closely approximate the distribution of latent ability than does the distribution of corrected raw scores. Therefore, the second stage constitutes a refinement on the first stage. In both stages the procedure iterates item by item through values of c_i to obtain pairs of a_i and b_i consistent with large sample estimates of the item-manifest ability point-biserial correlation and the item p-value. This allows the generation of various item characteristic curves (ICC's). The ICC's are then compared with the regression of the binary item on manifest ability. The ICC that best fits this regression, as indicated by the minimum χ-square, is given by the set of approximations – $\hat{\hat{a}}_i$, $\hat{\hat{b}}_i$, and $\hat{\hat{c}}_i$. The approximations are then corrected for characteristics of the particular sample of items being parameterized to obtain "ancillary estimates" – $\hat{a}_i$, $\hat{b}_i$, and $\hat{c}_i$. Ancillary estimation as a generic method was developed by Fisher (1950). The ancillary corrections improve the efficiency of the estimates.

The procedure has been evaluated through model sampling and simulation techniques. In particular, two parameterization samples, one of 2,000 and one of 3,000 cases, were generated from the logistic model using specified, and hence known, item parameters. The data were then analyzed by the procedure, and the resulting estimates were compared to the known parameters for each

[1] As total test reliability decreases, the approximations for the parameters a_i systematically underestimate the true values of a_i.

of the samples. Specifically, root mean square errors (RMSE's), i.e.

$$\left\{\sum_{i=1}^{m}(\hat{a}_i - a_i)^2\, m^{-1}\right\}^{1/2}, \quad \left\{\sum_{i=1}^{m}(\hat{b}_i - b_i)^2\, m^{-1}\right\}^{1/2}, \text{ and}$$

$$\left\{\sum_{i=1}^{m}(\hat{c}_i - c_i)^2\, m^{-1}\right\}^{1/2},$$

were obtained. These measures of deviation are given in Table 1 for the two parameterization samples and stages. Notice that the particular RMSE indicated by a given equation tends to decrease with stages. This is an indication of improved efficiency due to ancillary corrections. For the final stage ancillary estimates, these deviation measures were .242, .123 and .056 for the 2000 case sample, and .228, .148, and .056 for the 3000 case sample. For 100-item parameterization tests, these data indicated that 2,000 cases were sufficient for the effective use of the procedure. Correlations were also computed between the estimates and the known parameters, i.e., $r_{\hat{a}a}$, $r_{\hat{b}b}$, and $r_{\hat{c}c}$. These correlations are provided in Table 2 for the two parameterization samples and stages. Notice that there is a tendency for each correlation to increase with stages as predicted given that the ancillary corrections improve efficiency of estimation. For the final stage ancillary estimates, the correlations were .915, .996, and .764 for the 2,000 case sample, and .918, .997, and .760 for the 3,000 case sample. Since the ranges of the a_i and c_i were somewhat restricted, these correlations are very respectable. The results of these comparisons between the estimates and the known parameters indicated the merit of the item-analytic procedure.

The ancillary estimation procedure was further evaluated using simulation techniques. In particular, testing was conducted using a Bayesian algorithm developed by Owen (1969). Samples of 100 cases each were generated for computerized adaptive testing using 100 items with known item parameters. In the generation process, values of θ, the ability parameter, are sampled randomly from $N(0,1)$ and are also known. As a result, estimates of the ability obtained under computerized adaptive testing could be correlated with known ability. Comparisons of correlations, $r_{\hat{\theta}\theta}$, were made across three conditions of computerized adaptive testing where (1) the known item parameters, (2) the ancillary estimates of the item parameters based on the 2,000 case sample, and (3) the ancillary estimates of item parameters based on the 3,000 case sample were used in the algorithm. The appropriateness of the use of the ancillary estimates could be evaluated, therefore, by comparing the results obtained for the last two conditions with those

TABLE 1

Root Mean Square Errors for Estimates by Parameterization Samples and Stages

Sample Size	Parameterization Stage	Root Mean Square Error		
		$\left(\sum_{i=1}^{m}\{\hat{a}_i - a_i\}^2 m^{-1}\right)^{1/2}$	$\left(\sum_{i=1}^{m}\{\hat{b}_i - b_i\}^2 m^{-1}\right)^{1/2}$	$\left(\sum_{i=1}^{m}\{\hat{c}_i - c_i\}^2 m^{-1}\right)^{1/2}$
2000	Corrected Raw Score:			
	Approximation	.309	.181	.077
	Ancillary Estimate	.283	.120	.067
	Bayesian Modal:			
	Approximation	.269	.150	.061
	Ancillary Estimate	.242	.123	.056
3000	Corrected Raw Score:			
	Approximation	.308	.139	.081
	Ancillary Estimate	.253	.135	.073
	Bayesian Modal:			
	Approximation	.252	.109	.059
	Ancillary Estimate	.228	.148	.056

TABLE 2

Correlations Between Estimates and Known Parameters by Parameterization Samples and Stages

Sample Size	Parameterization Stage	Correlation		
		$r_{\hat{a}a}$	$r_{\hat{b}b}$	$r_{\hat{c}c}$
2000	Corrected Raw Score:			
	Approximation	.876	.996	.651
	Ancillary Estimate	.873	.996	.668
	Bayesian Modal:			
	Approximation	.909	.996	.754
	Ancillary Estimate	.915	.996	.764
3000	Corrected Raw Score:			
	Approximation	.884	.996	.611
	Ancillary Estimate	.895	.996	.616
	Bayesian Modal:			
	Approximation	.914	.997	.752
	Ancillary Estimate	.918	.997	.760

obtained for the first. In Table 3, the results are summarized for each of the conditions of testing.

Further explanation, however, is in order before proceeding to an interpretation of these results. When compared with conventional testing procedures, computerized adaptive testing can lead to a substantial reduction in the number of items required to obtain a given degree of validity. Therefore, the concern was not only with the validity obtained but also with the economy in items observed in obtaining the given validity. Control over the validity of computerized adaptive testing is direct. When an individual is being evaluated, the standard error of the estimate of ability is available at any stage in the sequence. Validity, over individuals, is controlled by terminating the

TABLE 3

Validity Coefficients ($r_{\hat{\theta}\theta}$), and Average Number of Items ($\bar{n}$) Required for Tailored Testing to Various Termination Rules Where the Item Parameters Were Known or Estimated

Termination Rules				Item Parameters Estimated in a Sample of:					
#	σ_ϵ	$\rho^2_{\hat{\theta}\theta}$	$\rho_{\hat{\theta}\theta}$	Parameters Known		2,000 Cases		3,000 Cases	
				$r_{\hat{\theta}\theta}$	$\bar{n}$	$r_{\hat{\theta}\theta}$	$\bar{n}$	$r_{\hat{\theta}\theta}$	$\bar{n}$
1	.5477	.70	.84	.84	2.7	.83	2.0	.84	2.3
2	.5000	.75	.87	.85	3.2	.86	2.7	.86	2.6
3	.4472	.80	.89	.89	3.9	.89	3.4	.88	3.2
4	.3873	.85	.92	.91	4.7	.90	4.0	.90	4.0
5	.3162	.90	.95	.94	6.6	.92	5.4	.93	5.6
6	.2828	.92	.96	.96	8.2	.94	6.7	.93	7.1
7	.2449	.94	.97	.96	10.8	.95	9.1	.94	9.6
8	.2236	.95	.97	.96	13.3	.95	11.1	.95	11.9

individual sequences at a common value for the standard error of the estimate of ability. In the study, eight such termination rules were designated. These rules are identified in columns 1 and 2 of Table 3 and specify that the standard error of the estimate of ability, σ_ϵ, was equal to or less than (1) .5477, (2) .5000, (3) .4472 (4) .3873, (5) .3162, (6) .2828, (7) .2449 and (8) .2236, respectively, over all individuals. Given σ_ϵ for any termination rule, synonomous rules may be generated through

$$\rho^2_{\hat{\theta}\theta} = 1 - \sigma^2_\epsilon \qquad (7)$$

and

$$\rho_{\hat{\theta}\theta} = \sqrt{1 - \sigma^2_\epsilon} \qquad (8)$$

for the expected reliability and validity, respectively. These synonomous rules are given in column 3 and 4. The validities of column 4 may then be compared with obtained validities. Eight estimates of ability satisfying these rules were obtained for all cases. Obtained validities were indexed by the correlations between known ability and estimated ability $r_{\hat{\theta}\theta}$, for specified termination rules as appropriate to the testing condition. As the termination rule becomes more stringent, the obtained validities given in columns 5, 7, and 9 increase and compare very closely with expected validities given in column 4. Additionally, the average numbers of items required, the $\bar{n}$, given in columns 6, 8, and 10 also increase as the termination rule becomes more stringent. Notice that the $\bar{n}$ at each termination rule differ only slightly across testing conditions. Since the results were almost identical across testing conditions, the item-analytic procedure appeared very appropriate in computerized adaptive testing applications. Consequently, ancillary estimates of the item parameters based on more than 2,000 cases and 100 items were strongly recommended for use in computerized adaptive testing.

Further research in evaluating the item-analytic procedure has been accomplished for varying numbers of cases and items (Gugel et. al., 1975), and more detailed recommedations regarding the use of the procedure will be given later in the conference.

As it turned out, the last significant question, "Is there an efficient and accurate adaptive algorithm for computerized testing?" could have been answered in the affirmative as early as 1969. The important event was the publication of an Educational Testing Service research bulletin, "A Bayesian Approach to Tailored Testing", by Roger J. Owen. Subsequent research (Urry, 1971, 1974*b*, in press-*a*; Jensema, 1972, 1974, 1975) has shown the efficiency and accuracy of the algorithm. For example, it is possible to construct some 2,000 computerized adaptive tests in some 17 minutes of central processor unit time, and the precision of measurement can be accurately controlled with termination rules.

In summary, we now find that: (1) the specifications for effective item banks have been developed, (2) these specifications can be met for a number of significant abilities, (3) efficient procedures exist for the reliable estimation of parameters, and (4) an efficient computerized adaptive testing algorithm is available to conduct the actual testing. All the necessary prerequisites for the success of computerized adaptive testing are therefore now in evidence. At this juncture, the feasibility of computerized adaptive testing can be realistically assessed, and this realistic assessment is decidedly and resoundingly affirmative in nature. At present, computerized adaptive testing appears to have a future without parallel in the literature of psychological measurement.

REFERENCES

Birnbaum, A. Part 5. Some latent trait models and their use in inferring an examinee's ability. In Lord, F. M. & Novick, M. R. *Statistical theories of mental test scores.* Reading, Mass.: Addison-Wesley, 1968.

Fisher, R.A. Contributions to mathematical statistics. New York: John Wiley & Sons, 1950.

Gugel, J. F., Schmidt, F. L. & Urry, V. W. Effectiveness of the ancillary estimation procedure. Paper presented at the conference on Computerized Adaptive Testing, Washington, D.C.: June 1975.

Jensema, C. J. An application of latent trait mental test theory to the Washington pre-college test battery. Unpublished doctoral dissertation, University of Washington, 1972.

Jensema, C. J. The validity of Bayesian tailored testing. *Educational and Psychological Measurement,* 1974, 34, 757-766.

Jensema, C. J. Bayesian tailored testing and the influence of item bank characteristics. Paper presented at the conference on Computerized Adaptive Testing, Washington, D.C.: June 1975.

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

Owen, R. J. A Bayesian approach to tailored testing. Research Bulletin 69-92. Princeton, N.J.: Educational Testing Service, 1969.

Samejima, F. Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph,* 1969, No. 17.

Urry, V. W. A Monte Carlo investigation of logistic mental test models. (Doctoral dissertation, Purdue University, 1970). *Dissertation Abstracts International,* 1971, 31, 6319B. (University Microfilms No. 71-9475)

Urry, V. W. Individualized testing by Bayesian estimation. Seattle: Bureau of Testing, University of Washington, 1971. (Duplicated Report)

Urry, V. W. Approximations to item parameters of mental test models and their uses. *Educational and Psychological Measurement,* 1974, 34, 253-269. (a)

Urry, V. W. Computer-assisted testing: calibration and evaluation of the verbal ability bank (TS-74-3). Washington D.C.: Personnel Research and Development Center, U.S. Civil Service Commission, December 1974. (b)

Urry, V. W. Ancillary estimators for the item parameters of mental test models. In press-a.

Urry, V. W. The effects of guessing on parameters of item discriminatory power. (TN-75-2) Washington, D.C.: Personnel Research and Development Center, U.S. Civil Service Commission, May 1975.