ACT Research Report Series 93-9

The Comparability of the Statistical Characteristics of Test Items Generated by Computer Algorithms

Richard Meisner
Richard Luecht
Mark Reckase

December 1993

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

The Comparability of the Statistical Characteristics of Test Items Generated by Computer Algorithms

Richard Meisner, Richard Luecht, and Mark Reckase

American College Testing

Please send requests for reprints to:

Richard Meisner
American College Testing
2201 North Dodge Street
Iowa City, IA 52243

Running head: The Comparability of the Statistical Characteristics

Abstract

This paper presents a study on the generation of mathematics test items using algorithmic methods. The history of this approach is briefly reviewed, followed by a survey of research done to date on the statistical parallelism of algorithmically generated mathematics items. Results are presented for 8 parallel test forms generated using 16 algorithms covering a variety of mathematics content and cognitive categories. The majority of the algorithms yielded items that were very homogeneous in their statistical characteristics. Those algorithms that did not yield homogeneous items were analyzed to determine if causes for differences could be determined. Possible innovative applications of the algorithms include computer generation of new test forms with specific content and statistical specifications, without the need for a pre-existing item bank.

The Comparability of the Statistical Characteristics of Test Items Generated by Computer Algorithms

Motivation for algorithmic item generation

Early suggestions for the algorithmic generation of achievement test items were in response to at least two interrelated weaknesses perceived in educational measurement: one at the test level, the other at the item level. An early discussion of both weaknesses was provided by Ebel (1962).

At the test level, there has been persistent dissatisfaction with the lack of meaningful information resulting from a purely norm-referenced interpretation of test results. At the item level, it has been argued that a more informative, criterion-referenced approach to testing would require a more objective, precise approach to item writing. Ebel's early comments on this subject are especially relevant:

> "The processes of test construction often appear to have more in common with artistic creation than with scientific measurement! In this respect, educational tests are distinctly different from most physical, chemical, or biological tests and measurements. In those more scientific fields, carefully specified measurement operations are designed to yield highly consistent results, almost regardless of the operator. The quantitative sophistication of many specialists in educational measurement is displayed, not in the precision and elegance of their procedures for obtaining initial measurements, but rather in the statistical transformations, elaborations, and analyses they are prepared to perform on almost any raw data given them. The term "raw" may be particularly appropriate when applied to the original data yielded by many educational tests. What we often overlook is the limited power of statistical transformations to refine these raw data and make them more precisely meaningful. If more systematic and standardized processes of test production could be developed and used, our educational measurements should become not only more consistently reproducible, but what is perhaps even more important, they should become more meaningful." (Ebel, 1962, p.22).

This theme has since been elaborated by many others, including Bormuth (1970):

> "To those with some scientific sophistication it will be clear that what is to be advocated is the very basic and almost self-evident idea that achievement test items, like any other measuring device, ought to be operationally defined." (p.3)

Many measurement specialists have been critical of what they see as an attempt to found an objective science upon individual measurement instruments which are largely the product of subjective creativity and judgements. More precise and replicable items or measurement tasks resulting from a more scientific approach to item writing would provide a more solid foundation for educational measurement as a meaningful science.

In an attempt to meet these concerns, Wells Hively in 1963 developed the concept of the "item form" as a practical means for effectively defining criterion behaviors in mathematics. A good definition of "item form" was provided by Osburn (1968):

> An item form has the following characteristics: (1) it generates items with a fixed syntactical structure; (2) it contains one or more variable elements; and (3) it defines a class of item sentences by specifying the replacement sets for the variable elements. (p.97)

As a specific example:

If _x + _ = _, then x = ?

The underlines (_) signify the variable elements. If we specify the "replacement set" for all three variable elements as integers from 2 to 20, the item form defines a universe of 19^3 = 6,859 unique items testing the ability to solve a certain type of linear equation.

Item forms have many advantages. They clearly define a domain; they allow substantial savings in test development time; and they are easily incorporated into computer software for rapid and automatic generation of items (Millman, 1989). Theoretically, from a few lines of computer code, a virtually infinite item pool may be obtained.

Durnin and Scandura (1973) criticized item forms on the grounds that they group items only on the basis of their observable appearance, and not on the skills needed to solve the items. The two do not always coincide. Items which are *produced* by the same algorithms may not necessarily all be *solved* by the same algorithms. Later overviews of item-writing technologies (e.g. Berk (1980), Roid (1984)), differentiate Durnin and Scandura's solution-oriented item algorithms from Hively's item forms.

Durnin and Scandura's approach to algorithmic item generation has distinct benefits. For example, it becomes possible to order items hierarchically based on analyses of the facts and skills required in their solution paths. This suggests a possible application to a computer-adaptive diagnostic testing system, in which correct responses might indicate branching to more complex solution paths until a fact or skill is added that is beyond the student's ability. Conversely, incorrect responses might indicate branching to less complex solution paths until the cause of difficulty is isolated.

Other researchers taking a solution-oriented approach have focused on anticipating and analyzing incorrect solution algorithms (with obvious promise for efficient computerized diagnostic testing). This line of investigation has been pursued, for example, by Brown and Burton (1978) and Tatsuoka (1990).

It seems reasonable to assume that items requiring the same knowledge and skills for their solution would be likely to exhibit parallel statistical characteristics for a given population of examinees. An interesting and important question is the extent to which items generated from the same algorithms actually turn out to be statistically parallel. The ability to reliably predict item statistics would have obvious implications for efficient parallel forms construction and computer adaptive testing. Research in this area has been surprisingly sparse.

Research on the statistical parallelism of algorithmically-generated items

Hively et. al. (1968) were the first to use item forms in research. They analyzed the performance of items based on item forms in the context of generalizability theory, and found a high degree of equivalence in total test scores:

> In general, the tests in every family satisfied very well the classical assumptions for parallel tests: equal means, variances, intercorrelations, and independence of "true" score and "error". (p.285)

Extending their analysis to questions of homogeneity of item operating characteristics, the results were less encouraging:

> The foregoing data lead one to place only moderate faith in the item forms as categories which represent distinct, homogeneous classes of behavior and which thus provide the foundation for detailed diagnosis and remediation. (p.289)

Macready & Merwin (1973) followed up on the homogeneity question, and found that item forms generating items of moderate difficulty could usually be used to obtain relatively homogeneous sets of items that are of approximately equivalent difficulty for a defined population of subjects.

Macready (1983) arrived at a similar conclusion for precisely defined arithmetic domains, finding that for many (but not all) of the domains, "it would appear to be possible to obtain accurate estimation of how students can be expected to perform on an entire domain of items based on their performance on a small sample of items." (p.156).

Scandura (1973) refined the item form concept by producing "equivalence classes" of items involving the same path of "atomic" skills in their solution, and found a significant improvement in coefficients of generalizability over the item forms used by Hively.

The above studies all focused primarily on low-level arithmetic operations. In this study, we look at the statistical parallelism of algorithmically-generated items having more advanced content.

Methods

Sixteen different item-generation algorithms were used in conjunction with the Math Item Creator (MIC) software (Meisner, 1993) to produce pretest items for eight forms of the ACT Assessment Program (AAP) Mathematics Test. One MIC pretest item was generated by each of the 16 algorithms for each of 8 test forms, producing a total of 16 × 8 = 128 items.

Design of the algorithms

The 16 algorithms used in this study were designed to cover a variety of skills from pre-algebra through intermediate algebra, and from basic skills through higher-order analysis. Each "item algorithm" actually consisted of a stem algorithm and 5 foil algorithms (one representing the correct solution method, the others representing anticipated incorrect solution methods). Brief descriptions of the 16 item-generating algorithms are given in Table 1.

Insert Table 1 here

The range of random numbers that could theoretically be generated for each algorithm was restricted as needed to assure that the items generated from a particular algorithm would be not just superficially similar to one another, but would require nearly identical skills for their solution.

As an illustrative example of the approach taken, an algorithm like the following could be used for generating stems:

For all x, $(ax + b)(c) = ?$

where a was a randomly generated integer chosen from the set $\{2,\ldots,9\}$ and b and c are randomly generated integers chosen from the set $\{1,\ldots,9\}$.

The key and four alternatives reflecting plausible solution errors could then be derived from the values of a, b, and c in the following way:

A. $acx + b$

B. $ax + b + c$

C. $ax + bc$

D. $acx + bc$

E. $abcx$

Pretest Samples

The MIC pretest item units were administered to examinee samples during the Spring, 1992 administration of the ACT Assessment Program (AAP). Each of the eight AAP Mathematics Test forms (labeled consecutively, Form 239 to Form 246) contained 16 MIC items, each representing a different generating algorithm (see Table 1).

The spiralled administration design of the AAP test appeared sufficient to assume that the eight test forms were assigned to randomly equivalent examinee sample groups. The sample sizes per pretest unit test form ranged from $N = 426$ to $N = 430$ (mode = 426). In four of the samples, one to four examinees were randomly eliminated to balance the data sets to the minimum, modal sample size of $N = 426$.

Results

The 16 Mathematics Item Creator (MIC) algorithms used in this pretest study constituted a primary facet of interest. A split plot design (e.g. Kirk, 1982) was used where "items" were considered to be domain sampled to match the treatment condition for each of the 16 MIC algorithms. This allowed examinees to be nested in forms, where algorithms and forms were crossed. The split plot design was forced to a full-rank, balanced model by random elimination of examinees ($n_{\mathrm{jk(min)}} = 426$), as noted earlier.

The split-plot ANOVA summary is provided in Table 2. The within examinee *residual* term is actually *algorithms x examinees-within-forms* plus the error component.

Insert Table 2 About Here

The nonsignificant main effect for *form* tends to support the assumption of random assignment of pretest units among the AAP Mathematics Test forms. That is, there were no systematic differences among the examinee samples that were administered the eight test forms.

However, there was a significant *form x algorithm* interaction. The mean p-values (proportion of correct responses per item) are plotted by algorithm (01 to 16) for each of the eight test forms in Figure 1.

Insert Figure 1 About Here

Two considerations relative to the interaction of test forms and algorithms are suggested by Figure 1. First, the general patterns of p-values are quite consistent across the 16 algorithms. No single test form or algorithm seems to demonstrate a systematic ordinal or disordinal trend that could adequately explain the significant interaction. Rather, different patterns of both ordinal and disordinal interactions occur for different algorithms in subtle ways. For example, Form 239 appears to show the greatest degree of aberrance relative to other forms, but only for certain algorithms (e.g. algorithms 07, 12, 13, 14, and 15). For the rest of the algorithms, the p-values for Form 239 are very consistent with the p-values for other test forms. Similar statements could be made regarding Form 240 and Form 241. In general, the lack of any systematic patterns makes it difficult to develop a rationale for the combinations of forms and algorithms that likely contributed to the overall significance of the interactions.

The second consideration suggested by Figure 1 concerns the apparent clusters of p-values (mean item scores) for some of the individual algorithms. For example, Forms 240, 241, and 242 indicate one clear cluster for algorithm 01 at approximately 0.46, with the remaining test forms producing a second cluster at approximately 0.56. This clustering also appears to occur for a number of other algorithms.

Post hoc contrast hypothesis tests for combined means in these clusters were not conducted. The standard errors averaged 0.024 for each of the mean proportions shown in Figure 1 with only nominal sampling variances (of the standard errors). Since the original observations were dichotomous scores, $u \in [0,1]$, standard errors were computed from the mean proportions, i.e., $\mathrm{SE} = \{[\mathrm{p}_{\mathrm{fa}}(1-\mathrm{p}_{\mathrm{fa}})]/N\}^{1/2}$. Given this small average standard error, differences between the p-values did not need to be very large to contribute to the significant interaction. However, in place of post hoc t-tests, independent χ^2 analyses were run on the dichotomous scores by form frequencies (2 x 8 tables) for each of the 16 algorithms, where $\alpha = 0.01$, the error rate, was adjusted for the 16 simultaneous significance tests ($\alpha_{\mathrm{adj}} = 0.0007$). Algorithms 1, 4, 10, 13, 14, and 16 produced significant χ^2 values at the adjusted level of α. The corresponding p-values for all the algorithms by form, with corresponding χ^2 values, are shown in Table 3.

Insert Table 3 About Here

In addition to looking at item difficulty, the algorithms were compared across the eight forms in terms of item-test biserial correlations[1]. Figure 2 provides a plot of the biserials for each pretest unit form.

Insert Figure 2 About Here

Although there are marked peaks (e.g., Form 239 at Algorithms 01 and 11, Form 240 at Algorithm 06, and Form 241 at Algorithm 10), the item discriminations appear rather consistent across forms. A no-interaction randomized block ANOVA conducted using the biserial correlations failed to detect any differences across forms. (It should be noted that, in general, biserial correlations tend to be less stable than item difficulties.) There was a main effect for algorithms, as might be expected.

The aggregate discriminating power of the pretest unit forms was also evaluated in another manner, using item response theory (IRT, e.g. Lord, 1980). The 16 items on each of the 8 forms were calibrated using a one-parameter logistic IRT model, i.e.

$$P(u_i|\theta;\xi_i) \equiv P_i(\theta) = \{1 + \exp[a(\theta - b_i)]\}^{-1} \qquad (1)$$

[1] The biserial correlations were computed by conditioning on the total number right scores for the operational ACT Assessment Mathematics test forms associated with each pretest unit.

where: θ is the latent trait, θ;

a is a constant slope parameter for all calibrated items denoting the discrimination or sensitivity of all the items to the latent trait, θ; and

b_i is an item difficulty parameter.

Note that the discrimination parameter, a, was estimated as a constant for all items within each form calibration, but was allowed to vary across forms.

BILOG (Mislevy & Bock, 1991) was used to perform these calibrations. The small sample sizes obtained for these pretest units precluded using the two- or three-parameter models. The 8 forms were scaled to a common metric under the assumption of equivalent sample groups. The *test information function* (e.g. Lord, 1980),

$$T(\theta) = a^2 \sum_{i=1}^{16} P_i(\theta)[1-P_i(\theta)] \qquad (2)$$

was then computed for all 16 items comprising each form. The function, $P_i(\theta)$ in (2), is the IRT probability given in (1). The test information, $T(\theta)$, indicates the measurement precision as a function of the latent trait, θ. Where $T(\theta)$ is equivalent for all forms of a test and that equivalence holds for all values of θ, the tests are said to be *parallel* (Lord, 1980). Figure 3 shows the test information plots for all 8 forms (239 to 246).

Insert Figure 3 About Here

The rather clear implication from Figure 3 is that Form 239 is more discriminating (i.e., has more information at the peak of the curve) than the other forms. Conversely, Form 242 is slightly less discriminating at its peak than the other forms. These differences in test information or discriminating power indicate that Forms 239 and 242 have slightly irregular characteristics in comparison to the other forms. Those same two forms were also previously noted as being more aberrant with respect to the reported differences in p-values (see Figure 1). However, we cannot be certain whether the irregularities arise from different ability distributions for the examinee samples, from characteristics of the generated algorithms, or some combination of both factors.

A more promising aspect of Figure 3 is that the majority of test information curves are actually quite similar. That similarity strongly suggests that the pretest forms, considered as aggregate units, tended to operate in a fairly parallel manner for the sampled examinees.

Discussion

The purpose of the research reported in this paper was to evaluate the characteristics of items that had been generated using algorithmic methods to determine if the items were comparable in the way that they functioned. A secondary goal was to determine if there were identifiable differences in the features of the items that would allow the statistical characteristics to be predicted. If both of these conditions held, than test forms could be custom created by computer to match desired content and statistical characteristics.

The content domain for this study was limited to that of mathematics achievement as assessed by the ACT Assessment Program. A detailed description of that content domain is presented in ACT (1992). That document describes the skills that are assessed by the test and the procedures used to select those skills. Some of the skills in that domain were carefully selected for use in this study, but no attempt was made to obtain a representative sample of skills. Rather, content areas were selected because they could be relatively easily assessed using an algorithmic approach.

The results of this study show that many algorithms for mathematics items yield items that are very homogeneous in their statistical characteristics. Items from these algorithms can be generated with high confidence that they will have certain statistical characteristics for populations of examinees that are consistent with those used in this study. Other algorithms seemed to yield items that varied in their characteristics more than would be expected by chance.

Some of the algorithms that resulted in items that varied in their characteristics yielded ready explanations for the differences. In these cases, it may be possible to refine the algorithms to provide items that were more uniform in their characteristics, or they could be kept as they currently are and they could be used to generate items with desired characteristics on demand. The discovery of these algorithms was a very exciting part of this study because it provided some evidence that test forms with specific statistical specifications could be written by computer using an algorithmic approach.

Some other algorithms did not reveal the source of the variation in the item characteristics. For these algorithms, a replication study is being conducted to determine whether the results were chance findings.

This study reports some initial efforts to understand the characteristics of items produced according to an algorithmic approach. Further studies are being conducted to gain additional knowledge of the advantages and disadvantages of this approach.

Future Directions for Algorithms Yielding Significant χ^2 Results

A followup study currently in progress will investigate more closely the algorithms from this study that resulted in significant χ^2 results. Following is a brief discussion of possible reasons for the variations in p-values among items produced with these algorithms.

Algorithm 1

Example:

What is the sum of the two solutions of the equation $x^2 + x - 12 = 0$?

For the four items in the low p-value cluster for this item form, the value of the middle coefficient of the quadratic equation was 1. The quadratic equations in the other four items had middle coefficients of 2, 3, or 4.

In the followup study we will look at this difference more systematically to see if this effect persists.

Algorithm 4

Example:

What is the smallest integer greater than $\sqrt{53}$?

The three items in the low p-value cluster had keys of 9 or 10. The keys of the remaining five items were 7, 8, 8, 8, and 11.

In the followup study we will further investigate the possibility of a correlation between key value and p-value.

Algorithm 10

Example:

To keep up with rising expenses, a motel manager needs to raise the $60.00 room rate by 22%. What will be the new rate?

No plausible explanation could be found for the variation in p-values with items of this type. Another set of 8 such items will be used in the followup study in the hope that a source of systematic variation will become apparent.

Algorithm 13

Example:

What is the slope of any line parallel to the line $2x + 3y = 3$?

This algorithm produced only one item with a prominently different (lower) p-value than the others, and this happened to be the item with the stem shown above. Interestingly, this outlying item was the only one of the eight to have a repeated digit in the stem (3).

The followup study will look further at the effect of repeated digits in the stem on item p-value.

Algorithm 14

Example:

If $(x + k)^2 = x^2 + 48x + k^2$, then $k = ?$

As with Algorithm 13, this algorithm produced only one item which departed significantly from the others, this time having an unusually *high* p-value. Of all the items generated, the item with the unusually high p-value had the lowest middle coefficient (22). The values of the middle coefficient in the other items were 36, 48, 52, 64, 76, 88, and 94.

In the followup study no middle coefficients less than 30 will be used.

Algorithm 16

Example:

For what value of a would the following system of equations have an infinite number of solutions?

$$4x - 2y = 15$$

$$12x - 6y = 5a$$

No plausible explanation could be found for the variation in p-values with items of this type. Another set of 8 such items will be used in the followup study to further investigate possible sources of systematic variation.

References

American College Testing (1992). *EPAS Content Validity of ACT's Educational Achievement Tests.* Brochure

Berk, R. A. (1980, September). A comparison of six content domain specification strategies for criterion-referenced tests. *Educational Technology*, 49-52.

Bormuth, J. R. (1970). *On the theory of achievement test items.* Chicago, Illinois: Univ. of Chicago Press.

Brown, J. S., & Burton, R. B. (1978). Diagnostic models for procedural bugs in basic mathematical skills. *Cognitive Science, 2*, 155-192.

Durnin, J. H. & Scandura, J. M. (1973). An algorithmic approach to assessing behavior potential: comparison with item forms and hierarchical technologies. *Journal of Educational Psychology*, 1973, *65*, 262-272.

Ebel, R. L. (1962). Content standard test scores. *Educational and Psychological Measurement*, *22*, 15—25.

Hively, W., Patterson, H. L., & Page, S. H. (1968). A "universe-defined" system of arithmetic achievement tests. *Journal of Educational Measurement*, *5*, 275-290.

Lord, F. M. (1980). Applications of Item Response Theory to Practical Testing Problems. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Macready, G. B., & Merwin, J. C. (1973). Homogeneity within item forms in domain referenced testing. *Educational and Psychological Measurement*, *33*, 351-360.

Macready, G. B. (1983). The use of generalizability theory for assessing relations among items within domains in diagnostic testing. *Applied Psychological Measurement*, *7*, 149-157.

Meisner, R. (1993). *Math Item Creator (MIC).* Software.

Mislevy, R. J. & Bock, R. D. (1991). PC-BILOG 3.04. Chicago, IL: Scientific Software.

Millman, J. (1984). Individualizing Test Construction and Administration by Computer. In R. A. Berk (Ed.), *A guide to criterion-referenced test construction* (pp. 78-96) Baltimore, MD: Johns Hopkins University Press.

Osburn, H. G. (1968). Item sampling for achievement testing. *Educational and Psychological Measurement*, *28*, 95-104.

Roid, G. H. (1984). Generating the test items. In R. A. Berk (Ed.), *A guide to criterion-referenced test construction* (pp. 49-77) Baltimore, MD: Johns Hopkins University Press.

Tatsuoka, K. K. (1990). Toward an integration of item-response theory and cognitive error diagnosis. In Frederiksen, N., Glaser, R., Lesgold, A., and Shafto, M., *Diagnostic Monitoring of Skill and Knowledge Acquisition* (pp. 453-488). Hillsdale, NJ: Erlbaum.

Figure Captions

Figure 1 Plot of Item P-values by Algorithms for 8 Test Forms
Figure 2 Biserial Correlations for 8 Test Forms by 16 MIC Algorithms
Figure 3 Test Information Curves for 8 MIC Pretest Units

Table 1. Description of item algorithms

Item Algorithm	Skill	Cognitive level
1	Solving a quadratic equation by factoring	Basic skills
2	Squaring a binomial	Basic skills
3	Computing the area of a triangle	Basic skills
4	Finding the smallest integer greater than $\sqrt{x}$	Basic skills
5	Conversion from scientific notation to decimal form	Basic skills
6	Solving a simple equation in two variables for one of the variables	Basic skills
7	Converting an equation to slope-intercept form	Basic skills
8	Dividing one fraction by another	Basic skills
9	Substitution of values into an algebraic expression	Basic skills
10	Calculating a percentage of a dollar amount	Application
11	Solving a quadratic equation by factoring	Application
12	Solving an absolute value equation	Basic skills
13	Calculating the slope of any line parallel to a given line	Basic skills
14	Calculating the last term in a polynomial perfect square from the value of the middle term	Analysis
15	Calculating the largest possible product for two integers having a given sum	Analysis
16	Calculating a coefficient that would give a system of linear equations in 2 variables an infinite number of solutions	Analysis

Table 2. Split Plot ANOVA Summary Table (N = 426, n_{forms} = 8, $n_{\text{algor.}}$ = 16).

Source	SS	df	MS	F
Between Examinees	3644.12	3407	1.07	-
Forms	12.77	7	1.82	1.71
Examinees w.Forms	3631.35	3400	1.07	-
Within Examinees	9626.19	51120	0.19	-
Algorithms	592.61	15	39.51	224.57*
Forms x Algorithms	61.41	105	0.58	3.32*
Residual	8972.16	51000	0.18	-
Total	13270.31	54527	0.24	-

* $p \leq 0.0001$

Table 3. P-Values and χ^2 Values for Form by Algorithm Interactions

Algorithm	Forms								χ^2 Values
	239	240	241	242	243	244	245	246	
01	0.531	0.472	0.458	0.448	0.577	0.533	0.554	0.585	34.51*
02	0.674	0.592	0.617	0.613	0.613	0.620	0.636	0.671	10.95
03	0.538	0.594	0.554	0.582	0.568	0.545	0.582	0.559	4.79
04	0.610	0.568	0.509	0.495	0.617	0.561	0.498	0.580	28.81*
05	0.660	0.723	0.681	0.688	0.746	0.707	0.728	0.683	12.03
06	0.526	0.509	0.514	0.585	0.566	0.561	0.570	0.545	9.38
07	0.653	0.587	0.585	0.615	0.592	0.610	0.594	0.577	7.30
08	0.653	0.671	0.573	0.669	0.615	0.643	0.660	0.711	22.27
09	0.742	0.779	0.768	0.803	0.742	0.714	0.784	0.777	13.80
10	0.739	0.758	0.667	0.732	0.796	0.763	0.803	0.805	34.88*
11	0.498	0.474	0.479	0.423	0.479	0.491	0.493	0.531	10.91
12	0.549	0.596	0.608	0.587	0.634	0.601	0.627	0.624	9.23
13	0.474	0.608	0.589	0.603	0.624	0.582	0.617	0.566	28.23*
14	0.678	0.531	0.491	0.493	0.561	0.484	0.573	0.538	48.95*
15	0.505	0.486	0.439	0.415	0.420	0.423	0.472	0.413	15.48
16	0.369	0.404	0.298	0.369	0.392	0.270	0.376	0.350	27.99*

* Prob($\chi^2 \geq \chi$) $\leq$ 0.0007 with df = 5.

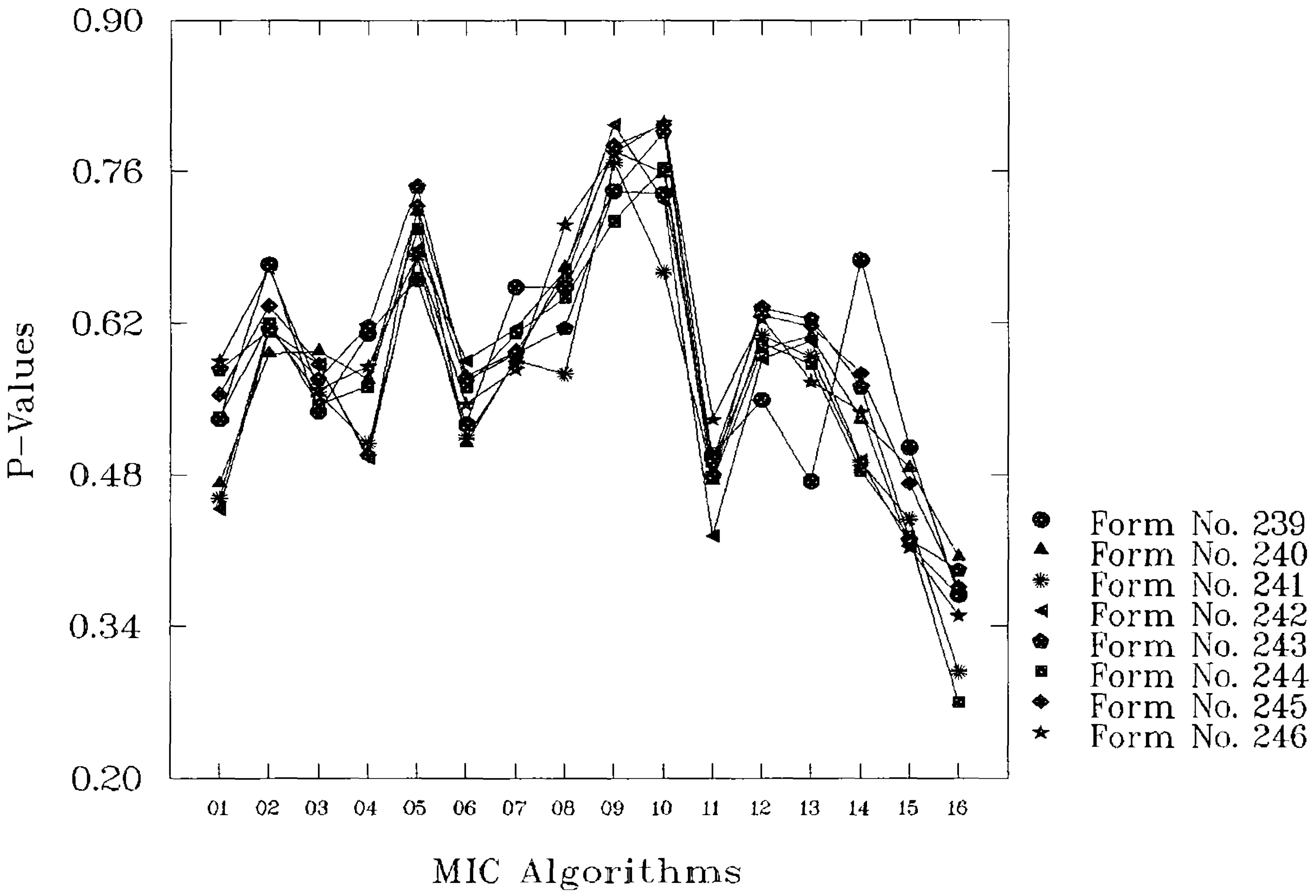

Figure 1. Plot of Item P-values by Algorithms for 8 Test Forms

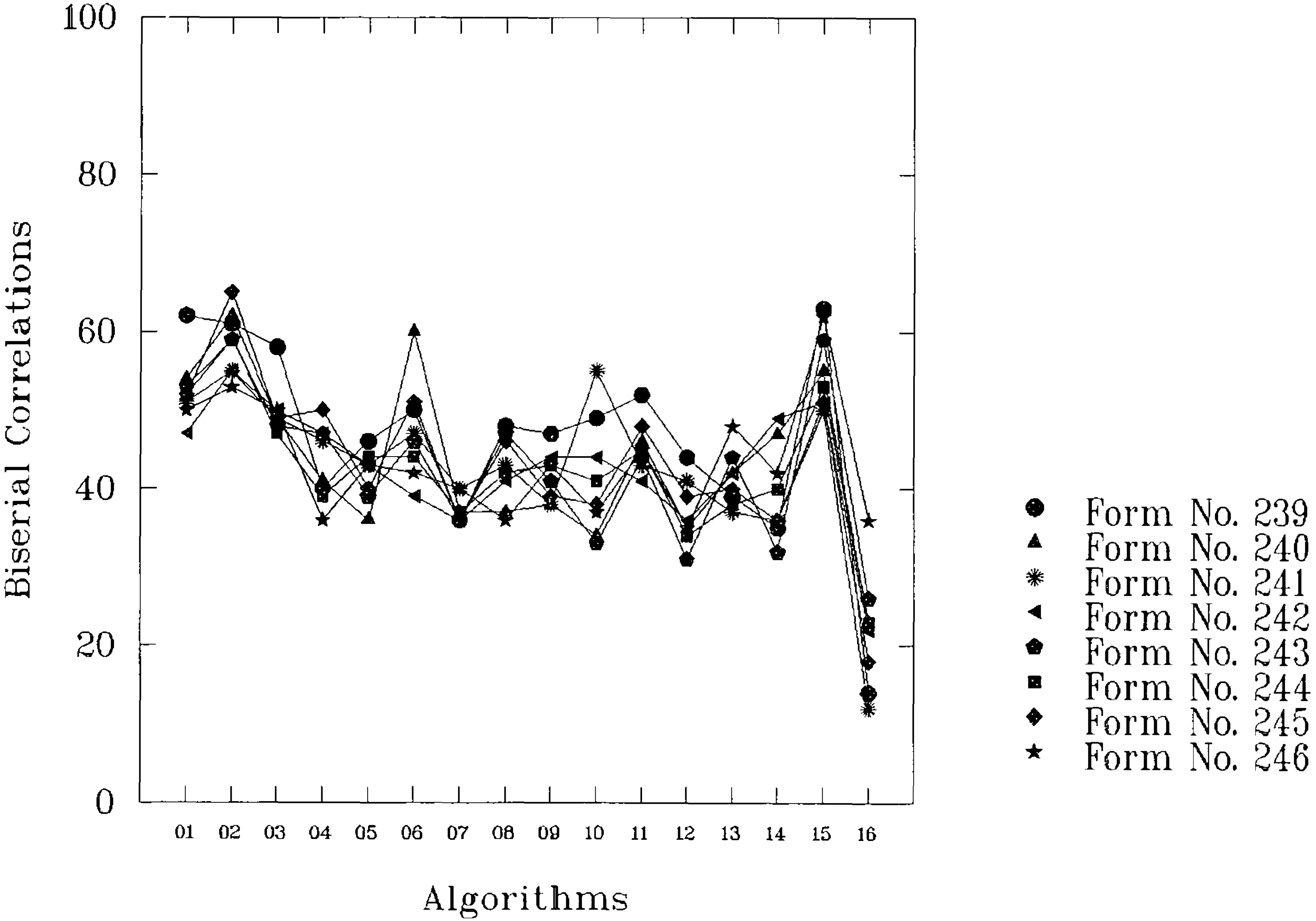

Figure 2. Biserial Correlations for 8 Test Forms by 16 MIC Algorithms

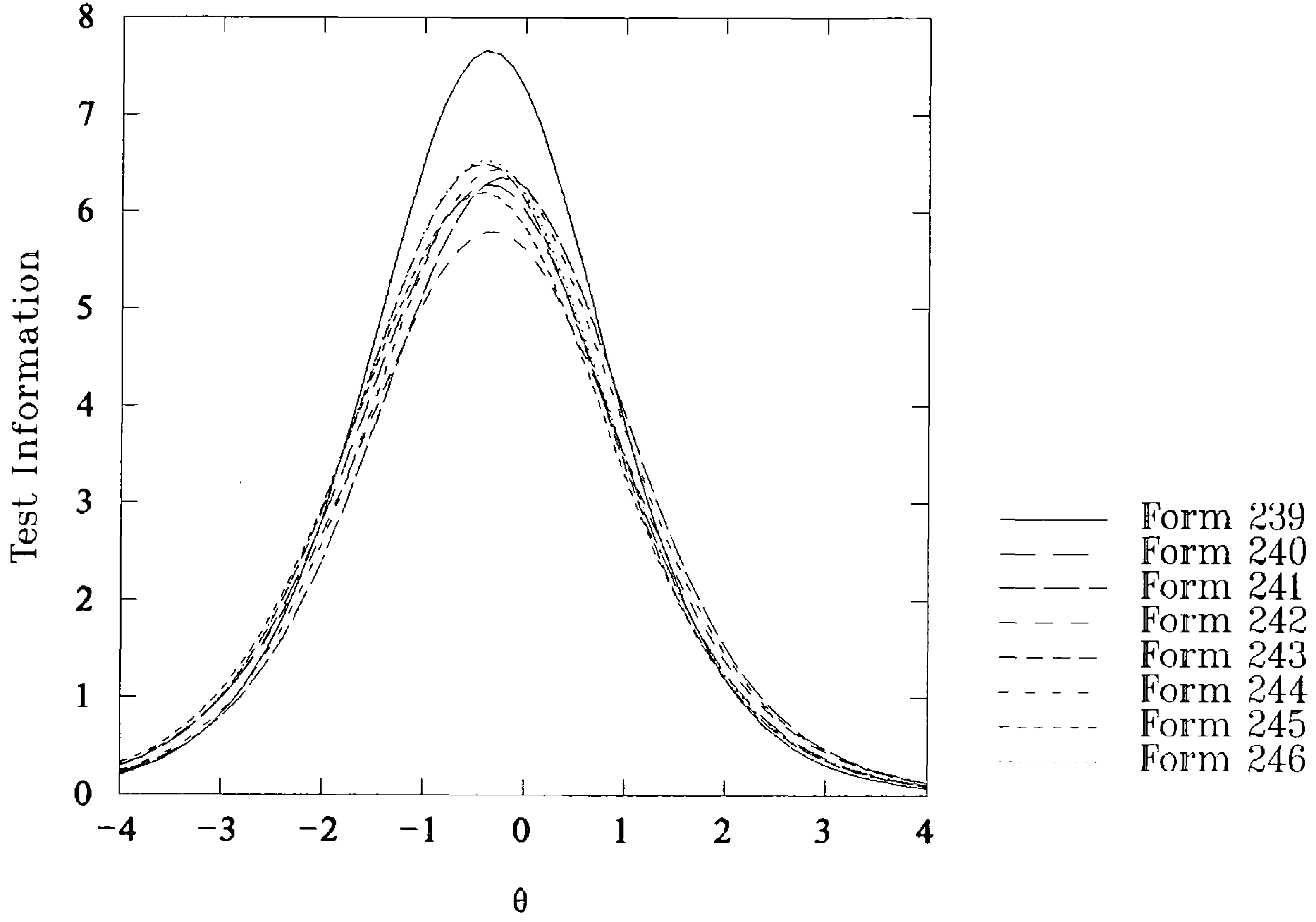

Figure 3. Test Information Curves for 8 MIC Pretest Units