DOCUMENT RESUME

ED 391 842 TM 024 658

AUTHOR Dirir, Mohamed A.
TITLE Construction of Parallel Test Forms Using Optimal Test Designs.
PUB DATE Apr 95
NOTE 22p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, CA, April 19-21, 1995).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Elementary School Students; Grade 4; Intermediate Grades; *Item Banks; Language Arts; Listening Comprehension Tests; Reading Comprehension; Scores; *Selection; Statistical Distributions; *Test Construction; Test Format; Test Reliability; Writing Tests
IDENTIFIERS *Parallel Test Forms

ABSTRACT

The effectiveness of an optimal item selection method in designing parallel test forms was studied during the development of two forms that were parallel to an existing form for each of three language arts tests for fourth graders used in the Connecticut Mastery Test. Two listening comprehension forms, two reading comprehension forms, and two written communication forms were developed using item-response-theory-based Optimal Test Design software. The tests included mixed item formats, cluster-based sections, and passage-related items. Three item banks were available for the study. For each of the 1993 forms, test and cluster information curves were computed. Restrictions on the number of passages that could be accessed for any test meant that relatively few items were available, but weakly parallel tests, as indicated by their relative information curves, were assembled using the software. The parallelism of the tests was also portrayed by similarities among the total score distribution indices and reliabilities of the tests. The small differences in the test information between the target and new forms were not significant threats to the parallelism of the forms. (Contains 2 tables, 4 figures, and 20 references.) (SLD)

Construction of Parallel Test Forms Using Optimal Test Designs

Mohamed A. Dirir

Connecticut State Department of Education

Paper Presented at the 1995 NCME Annual Meeting
San Francisco, California
April 21

Construction of Parallel Test Forms Using Optimal Test Design

In test development, the choice of an item selection method is important and may have an effect on the characteristics of the resulting test. The item selection method may affect a test's reliability, measurement precision, decision accuracy, and descriptive statistics (Ackerman, 1989; Hambleton, Dirir, & Lam, 1992; Dirir, 1993). It has been established that item response theory (IRT) provides useful and efficient item selection procedures (Birnbaum, 1968; Lord, 1980; Hambleton and de Gruijter, 1983; Green, Yen, & Burket, 1989). The IRT based item selection methods, sometimes referred to as optimal methods, are made possible by the fact that examinees and items are placed on the same scale.

In IRT, item selection methods entail selection of items that provide most information at desired targets along the ability scale. Because of the additive behavior of item information, this process would result in a test that provides the most information at the desired ability range. It has been shown in several studies that the optimal item selection procedures are superior to classical and random methods of item selection (Hambleton, Arrasmith, & Smith, 1987; Haladyna & Roid, 1983).

The IRT-based optimal test designs are implemented in computer programs. In most cases, optimization procedures implemented in linear programming or a heuristic approach are used in selecting items and minimizing or maximizing a test characteristic (most likely the test information) called the objective function. Optimal test development procedures are efficient, easy to use, and are available in computer software packages (see, for details, van der Linden & Boekkooi-Timminga, 1989; Theunissen, 1985; Adema, 1990; Baker, Cohen, & Barmish, 1988). Among the desirable features of the optimal test design procedures are the construction of parallel test forms (Ackerman, 1989; Boekkooi-Timminga, 1991). However, the few studies in this area were limited to multiple-choice tests, simulated data sets in some of the studies, synthetic target information curves, and few constraints in the item selection. The currently used performance measures, the increased development of passage-based tests, and the increased use of open-ended items all need to be tested with the IRT-based optimal test designs. Whether the optimal test designs can solve test development problems in performance assessments as effectively as in traditional assessments needs to be explored.

In large scale assessment, it is not unusual to construct parallel forms which would be administered in several successive testing years. The forms are often designed to be parallel (as

close as possible) in order to assess the different cohorts consistently. This could be accomplished by maintaining similar test content composition, test difficulty, and item quality across years and forms. The parallelism of the different forms is often assessed by examining different indicators such as the examinee score distributions, the reliabilities of the forms, the content composition of the forms, the test information functions, or the test characteristic curves.

The purpose of this study was to examine the effectiveness of optimal item selection method in designing parallel test forms. The efficiency of the procedure in developing two forms that are parallel to an existing form for each of three language arts tests was investigated. All items in the tests were passage-related, and two of the tests included both multiple-choice and open-ended items. Two listening comprehension forms, two reading comprehension forms, and two written communication forms, which were parallel to existing forms, were developed. The IRT-based Optimal Test Design (OTD) software (Verschoor, 1991) was used in developing the new test forms.

Data.

The target and constructed forms were language arts tests intended for fourth grade students. The anchor tests were listening comprehension, written communication, and reading

comprehension tests which were administered to Connecticut's fourth grade students in the 1993 administration (Form F) of the Connecticut Mastery Test (CMT). The CMT is a census criterion-referenced test which is administered every year to the State's fourth, sixth, and eighth graders.

The language arts sections of the CMT were developed on the premises of the National Assessment of Educational Progress (NAEP) Frameworks--Aspects of Reading Literacy. Each of the studied tests consists of two or more clusters, and there are two or more objectives within each of the clusters. The reading comprehension consists of three clusters: A, B, and C; the listening comprehension consists of two clusters: A and B; the written communication consists of three clusters: P, C, and E. Furthermore, each of these clusters consists of two or more objectives. The listening and reading forms contain a mixture of open-ended and multiple choice items, while the written communication form consists of multiple-choice items only. All CMT language arts tests are reported in clusters, and each cluster has a cut-off score on the raw score scale against which students are classified as masters or nonmasters.

Three item banks, which were pretested and pre-calibrated with the Rasch model, were available for the study. Item statistics and descriptives such as Rasch difficulty, item format, item

cluster (A, B, and so on), item objective, the passage which the item belongs to, and the passage type (inferential or narrative) were all available in the item banks. Three hundred and five listening comprehension items, 345 reading comprehension items, and 529 written communication items were available. However, the number of items which could be selected for any form depended upon the three passages used to develop that particular form. Hence small numbers of items were accessible in assembling each form. The number of passages in each test were 10 in reading comprehension, 17 in written communication, and 22 in listening comprehension.

Although the effect of the dimensionality of the item bank on optimally designed test is relatively smaller (Dirir, 1993; Ackerman, 1991), it is desirable to document the factorial structure of any measure before the test is assembled. It was assumed that each item bank will have dimensionality similar to that present in the corresponding Form F test. Hence, the dimensionality of the Form F data sets was assessed prior to this study. Two analyses were carried out in each data set. One was to examine whether each test was essentially undimensional, and was used with the multiple-choice portions of each form. The DIMTEST program (Stout, Nandakumar, Junker, Chang, & Steidinger, 1991) was used for these analyses. It was found that each of the tests (listening comprehension, reading comprehension, and written

communication) was essentially unidimensional. The second dimensionality assessment was confirmatory factor analyses using LISREL 7 (Joreskog & Sorbom, 1989). The intention was to examine if the clusters in the same test measure different traits. Items belonging to the same cluster were grouped into item parcels in each test, and each cluster was treated as a factor. Three factors were fitted to each the reading comprehension and the written communication, while two factors were fitted to the listening comprehension. The correlations among the cluster-based factors were then examined. In each of the data sets, the correlations among the factors were substantially high (from 0.9 to 1.0), an indication that the clusters are measuring the same trait. As both dimensionality assessment approaches have indicated, each of Form F tests was unidimensional, and any test developed from the banks could be presumed unidimensional.

Method.

For each of the 1993 forms (target forms), the test and cluster information curves were computed. The cluster information curves were needed to compare since the CMT language arts test forms are reported in clusters. The obtained information curves were used as target in assembling the new forms. Using Form F information curves as a target seemed appropriate for the purpose of this study. The OTD software (Verschoor, 1991) was used in constructing tests with information similar to the target tests,

and under the constraints proposed by the test development committee for the 1993 form. The constraints included the number of passages in each test, the number of items from each cluster, the number of items from each objective, and the number of open-ended items in each cluster. Table 1 summarizes the blueprint used in developing Form F for each test.

Table 1

Number of Items per Cluster in Form F

Test	Cluster	Multiple Choice	Open Ended	Total
Reading Comprehension	A	8	2	10
	B	7	1	8
	C	4	3	7
Listening Comprehension	A	8	3	11
	C	7	2	9
Written Communication	P	15	-	15
	C	15	-	15
	E	15	-	15

As mentioned in the preceding section, objectives are nested within the clusters, and almost all objectives were included in the test in each area. The number of items which were selected from each objective varied, ranging from one item in several

objectives in the listening test to nine items in each of two objectives in the written communication. Each of the three clusters in the Form F written communication comprised 15 multiple-choice items. In Form F listening comprehension, cluster A consisted of 11 items, and cluster B consisted of nine items. Furthermore, cluster A comprised three open ended items and eight multiple-choice items, while cluster B comprised two open-ended items and seven multiple-choice items. In the reading comprehension, cluster A comprised 10 items, cluster B comprised eight items, and cluster C comprised seven items. Also, there were two open-ended items in Cluster A, one open-ended item in cluster B, and three Open-ended items in cluster C.

Following the composition of and constraints in Form F, the new forms, called Form G and Form H, were assembled sequentially using OTD. The sequential process is more appropriate than the simultaneous procedure when the Rasch model fits the data (see, for comparison of the two methods, Boekkooi-Timminga, 1991). The number of passages in each form was fixed at three for each test. Hence, three passages were randomly selected from the available passages (10 to 22 per test) in developing each form. Separate passages were used for each form; that is, the three passages used to develop Form G were not reused in developing Form H. In selecting items, an ability range between -1 and +1 was given a priority. This ability range was chosen because that was the area

most of Form F tests provided highest information. Moreover, this ability range is where the cut-off score would most likely be located.

In exploring the extent the new forms were parallel to the target form, several indicators were examined. First, the test information curves of the new forms and the old form were compared. The information provided by the clusters were also compared. Second, the test characteristic curves of the forms were examined as well as the characteristic curves of the clusters. Finally, the distribution properties of the simulated examinee scores on the forms, and the reliabilities of the forms were examined.

The final comparison was accomplished by using simulated data. Abilities were generated for 1000 examinees from a normal distribution (zero mean and unit variance). In each of the Forms F, G, and H, the probability of each examinee getting a score of 1 on each of the multiple-choice items was computed using the one-parameter logistic model (Hambleton & Swaminathan, 1985). The probabilities of each examinee getting a score 1 and score of 2 on each of the open ended items were then computed using Muraki's generalized partial credit model (Muraki, 1992). For the open-ended items, the discrimination indices were fixed at one since the item banks were originally calibrated with the one-parameter

model. Uniform random numbers in the interval [0,1] were generated and compared with each examinee's probability (probabilities for the open-ended items). In any item, the examinee was then assigned to a score of 0 or 1 in the multiple-choice items, and a score of 0, 1, or 2 in the open-ended items. The total score distribution and Cronbach's alpha were then computed for each of the resulting test data sets.

Results.

Information Curves. The first comparison that was made among the forms was on the information curves. This was the criterion used to assemble Forms G and H, and the criterion in which the forms were supposed to be evaluated. Figure 1 shows the information curves of the written communication test. At the total test level, the new forms provided more information at an ability range between -1 and 1. Also, the new forms were closer to each other than to Form F. At the cluster level, a trend similar to that of the test level was found. That is, Forms G and H provided more information at the center of the distribution, and were closer to each other than to Form F.

In the listening comprehension test, Form G provided information equivalent to the one provided by Form F at all ability levels as can be seen in Figure 2. Form H, on the other hand, differed at ability range between -1.5 to .2 in which it provided more

information than the other two forms. At the cluster level, Form H provided more information between -1 and .3 in cluster A, while Form G provided information similar to Form F at most of the prioritized ability range. In cluster B, both new forms provided approximately same information as Form F at all ability levels. Figure 3 shows the information curves of the three forms in the reading comprehension. The new forms provided more information at the peak of the distribution at test level and in all clusters. Moreover, the new forms seem to be more parallel in the information provided at all ability levels.

In all three tests, the information provided by the new forms is greater at the focused ability range [-1,1] at the test level except one form in the listening comprehension. That was also true for the clusters in the reading comprehension and the written communication. Hence it was possible to reach the target information, and may be beyond at the critical ability range, in assembling forms that are closely parallel to Form F. The slightly more information at the prioritized ability ranges were not unexpected; similar results were reported in other studies (for example, Ackerman, 1989).

Characteristic Curves. The characteristic curves of the forms were also compared at test level and at cluster level. Figure 4 shows the characteristic curves of the written communication

test. As in the comparisons of information curves, all three forms were quite close, and the two new forms were even closer together. That is especially true for the whole test and cluster C. Similar results were found in comparing the characteristic curves of the listening and reading tests. In all cases, all three forms had relatively close curves at both test level and cluster level.

Summary statistics. Using simulated examinees, the total score distribution statistics and the reliability of each of the studied tests were examined. Table 2 summarizes the statistics of the first four moments of the total scores and the reliability indices of the three forms by test. As can be seen in the table, the four moments were all close among the forms in all tests. The reliabilities of the forms were also close. In fact the new forms have alpha coefficients higher than that of Form F by 0.01 in almost all tests (this difference was 0.02 in Form H reading). Obviously, that is in agreement with the findings in the information function analyses where slightly more information were provided by the new forms at the prioritized ability ranges.

Discussion

The purpose of the paper was to examine the feasibility of assembling tests which are parallel to existing forms. Language arts tests which included mixed item formats, cluster-based

Table 2

Summary of the Generated Test Score Distributions for the Three Forms

Test	Form	Mean	STD Dev.	Skewness	kurtosis	Alpha
Reading	F	17.6	5.14	-0.16	-0.70	0.76
	G	17.8	5.45	-0.22	-0.71	0.77
	H	18.0	5.61	-0.21	-0.75	0.78
Listening	F	13.9	4.20	-0.10	-0.43	0.70
	G	14.5	4.34	-0.14	-0.57	0.71
	H	13.9	4.24	-0.06	-0.54	0.71
W. Commun.	F	22.6	7.26	0.02	-0.64	0.82
	G	23.1	7.53	0.02	-0.75	0.83
	H	22.5	7.76	0.07	-0.78	0.83

sections, and passage-related items were investigated. Although fewer items were available due to restriction on the number of passages that could be accessed for any test, weakly parallel tests, as indicated by their relative information curves, were assembled using the IRT-based OTD software. The parallelism of the tests was also portrayed from a classical perspective by the similarities among the total score distribution indices and reliabilities of the tests. It has been shown that large scale assessment programs could benefit from using the currently available test construction software, and that most of the

constraints used in test development could be incorporated in the optimal test designs.

The small differences in the test information between the target and the new form were not a significant threat to the parallelism of the forms. The form that seemed different in few cases was the target form, and all three forms might have been close had they been developed at the same time. Moreover, the small differences between the information curves could be controlled in some optimal test assembly procedures (see, for discussion, van der Linden & Boekkooi-Timminga, 1989; Boekkooi-Timminga, 1991). The objective function in the test development could be formulated such that the distance between the target information curve and the new test's information curve is minimal.

REFERENCES

Ackerman, T. A. (1989, April). An alternative methodology of creating parallel test forms using IRT information function. Paper presented at the meeting of the NCME, San Francisco.

Ackerman, T. A. (1991, April). An examination of effect of multidimensionality on parallel forms construction. Paper presented at the meeting of NCME, Chicago

Adema, J. J. (1990). The construction of customized two-stage tests. Journal of Educational Measurement, 27, 241-253.

Baker, F., Cohen, A., & Barmish, B. R. (1988). Item characteristics of tests constructed by linear programming. Applied Psychological Measurement, 12, 189-199.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, Statistical theories of mental test scores (pp. 397-479). Reading, MA: Addison-Wesley.

Boekkooi-Timminga, E. (1991). The construction of parallel tests from IRT-based item banks. Journal of Educational Statistics, 15, 129-145.

Dirir, M. A. (1993). The effects of dimensionality and item selection methods on the validity of criterion-referenced test scores and decisions. Unpublished Doctoral Dissertation, University of Massachusetts at Amherst.

Green, D. R. , Yen. W., & Burket, G. R. (1989). Experiences in the application of item response theory in test construction. Applied Measurement in Education, 2, 297-312.

Haladyna, T., & Roid, G. (1983). A comparison of two approaches to criterion-referenced test construction. Journal of Educational Measurement, 20, 271-281.

Hambleton, R. K., Arrasmith, D., & Smith, I. L. (1987). Optimal item selection with credentialing examinations (Laboratory of Psychometric and Evaluative Research report No. 157). Amherst, MA: University of Massachusetts, School of Education.

Hambleton, R. K., & de Gruijter, D. N. (1983). Applications of item response models to criterion-referenced test item selection. Journal of Educational Measurement, 20, 355-367.

Hambleton, R. K., Dirir, M. A., & Lam, P. (1992, April). Effects of optimal test designs on measurement precision and decision accuracy. paper presented at the meeting of the AERA, San Francisco.

Hambleton, R. K., & Swaminathan, H. (1985). Item response theory: Principles and applications. Boston, MA: Kluwer Academic Publishers.

Joreskog, K., & Sorbom, D. (1989). LISREL 7 (a software package). Mooresville, Indiana: Scientific Software.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. Applied Psychological Measurement, 16, 159-176.

Stout, W., Nandakumar, R., Junker, B., Chang, H, & Steidinger, D. (1991). DIMTEST and TESTSIM (a software package). Champaign, IL: University of Illinois.

Theunissen, T. J. J. M. (1985). Binary programming and test design. Psychometrika, 50, 41-420.

van der Linden, W., & Boekkooi-Timminga, W. (1989). A maximin model for test design with practical constraints. Psychometrika, 54, 237-247.

Verschoor, A. (1991). Optimal test design (a software package). Arnhem, The Netherlands: CITO.

Figure 1. Comparions of Information Curves for the Written Communication

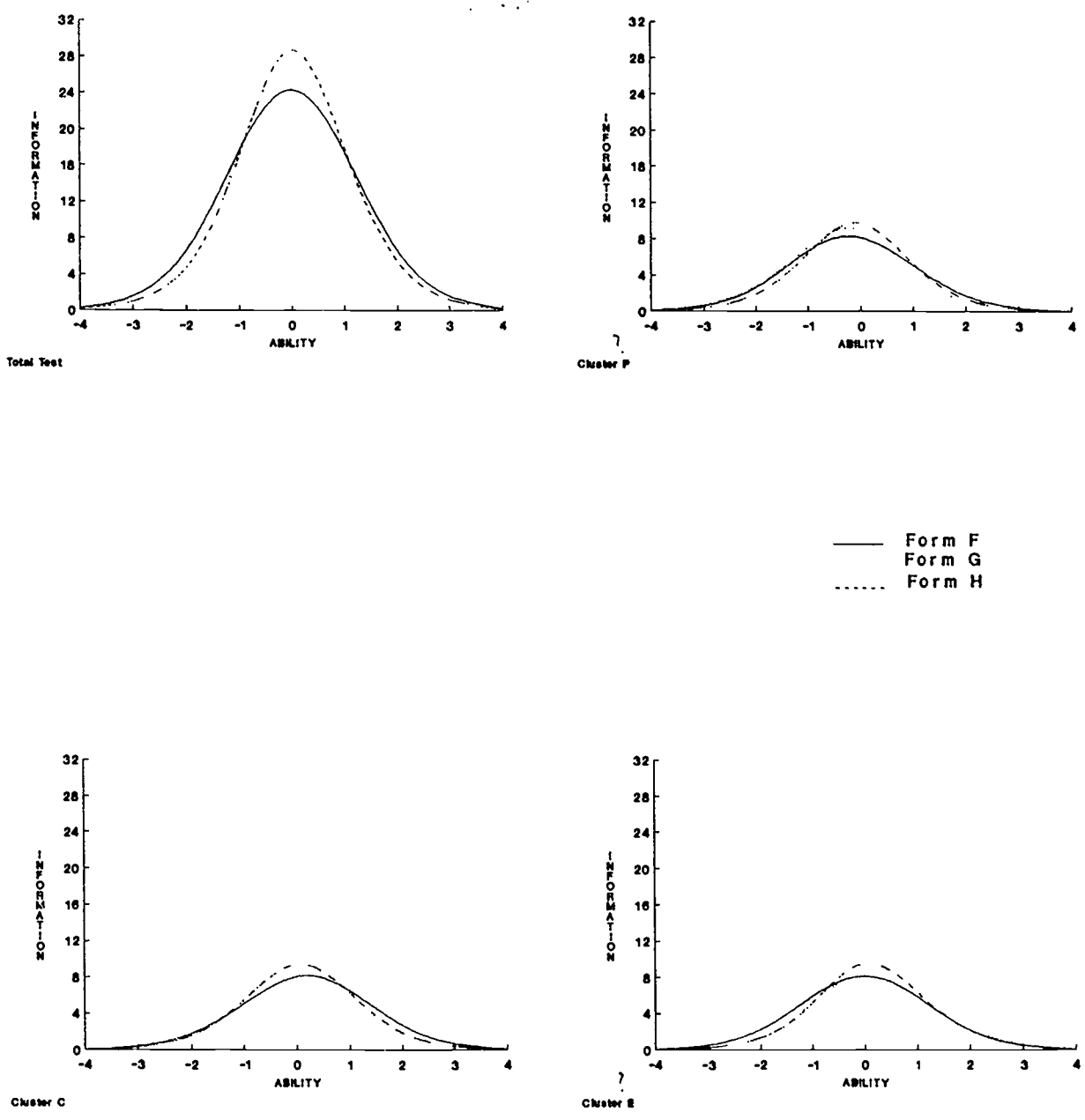
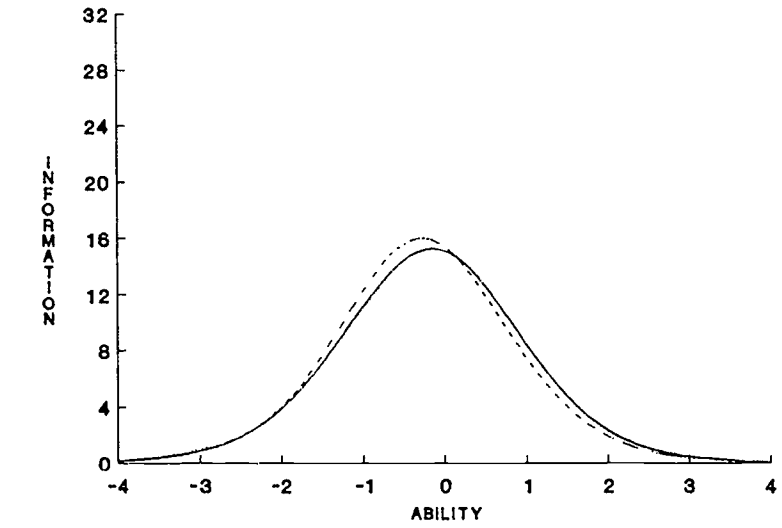
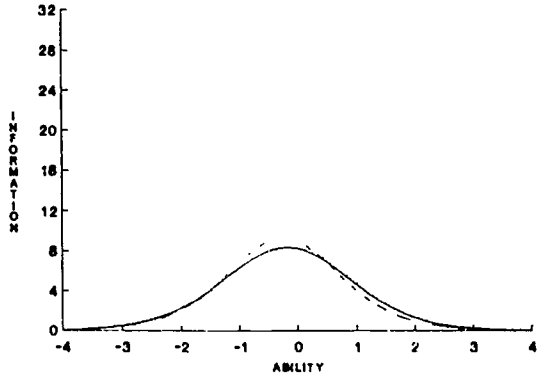


Figure 2. Comparison of Information Curves for the Listening Comprehension

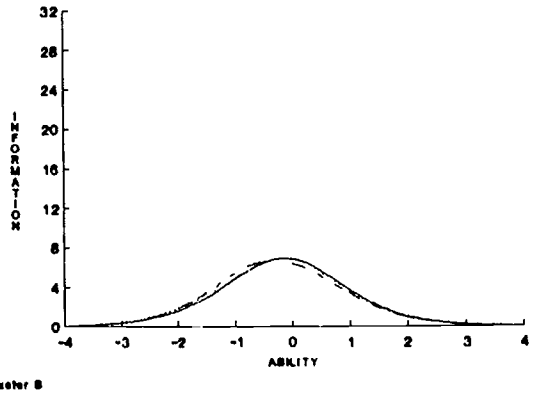


TOTAL TEST

Form F
Form G
Form H



Cluster A



Cluster B

Figure 3. Comparions of Information Curves for the Reading Comprehension

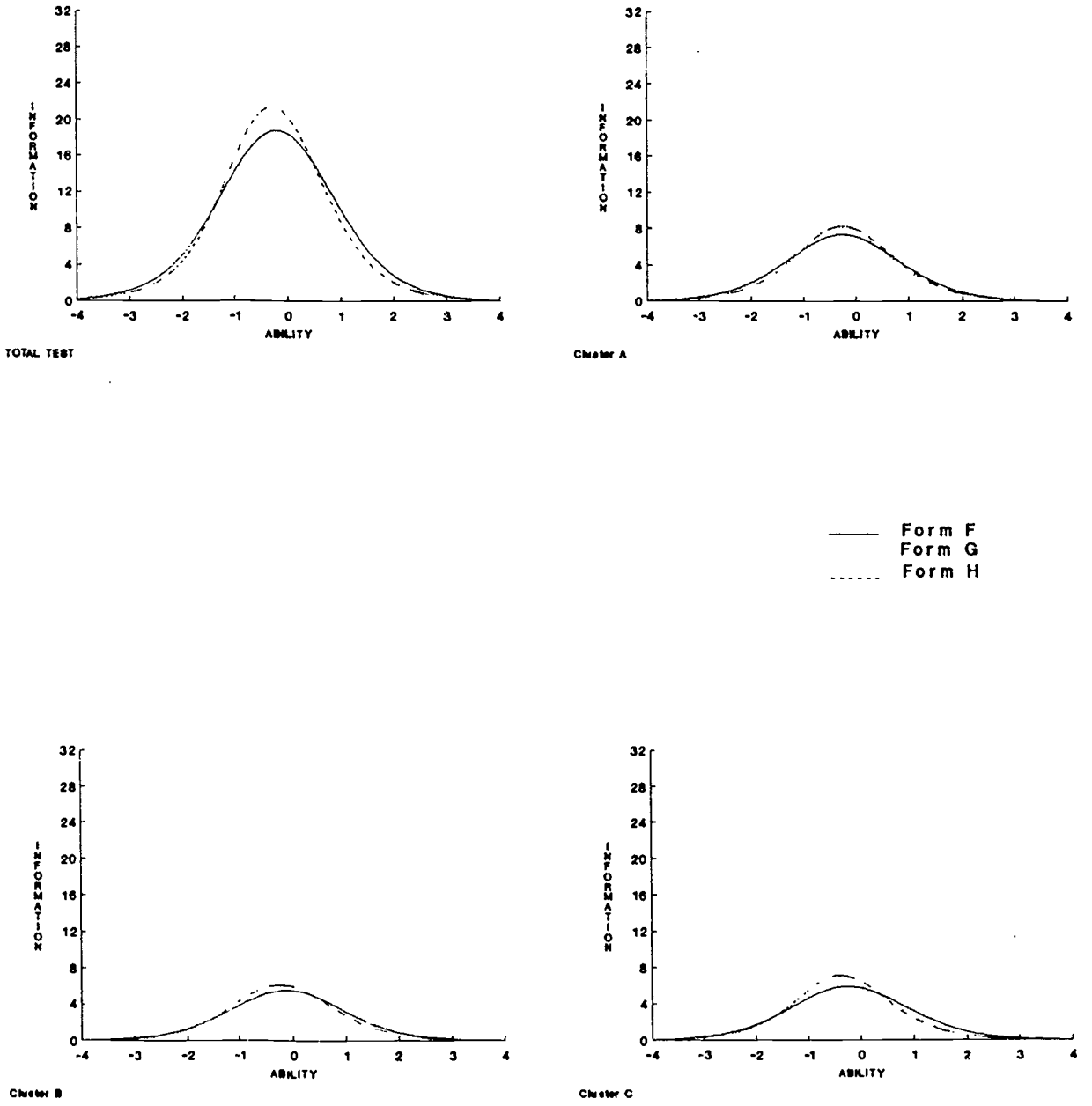


Figure 4. Comparions of Characteristic Curves for the Written Communication