DOCUMENT RESUME

ED 390 926 TM 024 373

AUTHOR Smith, Robert L.; Carlson, Alfred B.
TITLE Using Judgmental Estimates of Item Difficulty To Assemble Test Forms with Equivalent Cut Scores. Research Memorandum.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RM-95-2
PUB DATE Aug 95
NOTE 24p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Cutting Scores; Decision Making; *Difficulty Level; *Equated Scores; *Estimation (Mathematics); Evaluation Methods; Evaluators; *Judges; Licensing Examinations (Professions); Raw Scores; Sample Size; Test Construction; Test Format; *Test Items
IDENTIFIERS Test Specifications

ABSTRACT

The feasibility of constructing test forms with practically equivalent cut scores using judges' estimates of item difficulty as target "statistical" specifications was investigated. Test forms with equivalent judgmental cut scores (based on judgments of item difficulty) were assembled using items from six operational forms of the Multi-State Insurance Licensure Program examination. Nineteen judges took part in the standard setting session. Comparisons between the judgmental and equated cut scores showed the judgmental cut scores to differ by one or two raw score points from cut scores obtained through equating procedures. Comparisons of equated cut scores for test forms constructed using judgmental estimates of item difficulty and those constructed using classical statistics suggested that judgmental estimates of item difficulty may be suitable for use as target "statistical" specifications when empirical item difficulties are not available or are unstable due to small sample size. (Contains 2 tables, 6 figures, and 2 references.) (SLD)

RM-95-2

RESEARCH MEMORANDUM

USING JUDGMENTAL ESTIMATES OF ITEM DIFFICULTY TO ASSEMBLE TEST FORMS WITH EQUIVALENT CUT SCORES

Robert L. Smith
Alfred B. Carlson

Educational Testing Service
Princeton, New Jersey
August 1995

USING JUDGMENTAL ESTIMATES OF ITEM DIFFICULTY TO ASSEMBLE TEST FORMS WITH EQUIVALENT CUT SCORES

Robert L. Smith

and

Alfred B. Carlson

Acknowledgments

We would like to express our thanks to Brendon Hickey for his contribution to the development of the proposal and as consultant on the item response theory equating; Anna Kubiak who developed the specifications from the judges' ratings used to construct the test forms; Yvonne Hopson for coordinating the cut score meeting; Sheila Connolly and Duane Heinrichs for conducting the cut score meeting that yielded the judges' estimates of item difficulty; Sheila Connolly for assembling the test forms according to the specifications; Mei Su for her assistance in data analysis; and Debbie Slovinsky for assistance in data analysis and for producing the graphs. We would also like to thank William Cowell, David Wright, and Anne Ninneman for many helpful comments on earlier drafts of the paper.

Abstract

The feasibility of constructing test forms with practically equivalent cut scores using judges' estimates of item difficulty as target "statistical" specifications was investigated. Test forms with equivalent judgmental cut scores (based on judgments of item difficulty) were assembled. Comparisons between the judgmental and equated cut scores showed the judgmental cut scores to differ by one or two raw score points from cut scores obtained through equating procedures. Comparisons of equated cut scores for test forms constructed using judgmental estimates of item difficulty and those constructed using classical statistics suggested that judgmental estimates of item difficulty may be suitable for use as target "statistical" specifications when empirical item difficulties are not available or are unstable due to small sample size.

Using Judgmental Estimates of Item Difficulty to Assemble Test Forms with Equivalent Cut Scores

Introduction

When multiple forms of a test are to be used interchangeably, it is essential that scores from the various forms be comparable. However, very small candidate volumes, a characteristic of some certification and licensing tests, often do not support traditional equating procedures. As sample sizes approach zero they not only become too small for stable empirical equating, but the usual item statistics also become unstable. In these circumstances it is essential that the test development procedure be enhanced --- ideally, to the point that parallel forms are developed so that equating becomes unnecessary. In practice, of course, parallel forms are rarely achieved even when using statistics based on reasonably sized samples. Therefore, the focus of this research is only on achieving *equivalent cut scores*, which are the primary scores of interest in licensing and certification tests.

In order to ascertain whether it is possible to assemble forms with equivalent cut scores using judgmental data, the research was conducted using judgmental estimates of item difficulty obtained using the Angoff (1984) procedure, a test with a fairly large item pool, and a large group of judges. The judges were trained by familiarizing them with the total group candidate performance on representative items using standard item analysis data. Such data

are at least occasionally available for programs with small candidate volumes, whereas the optimum training data, performance of candidates in the region of the cut, would be based on small sample sizes. Test forms were assembled to have equivalent cut scores using the judgmental estimates of item difficulty. The criterion used to assess the efficacy of the method was the equated cut score for each of the respective test form.

Method

Item Pool. Items from six operational forms (A-F, 50 items per form), constructed according to test specifications of the Multi-State Insurance Licensure Program (MILP), Life (Part 1) Examination formed the item pool. Due to item redundancy in these forms the entire pool consisted of only 278 items. All items were four-option multiple-choice items previously calibrated for a three-parameter logistic model using LOGIST (Wingersky, Patrick, & Lord, 1987). Item calibrations were based on approximately 2000 examinees.

Panel of Judges. Nineteen judges took part in the standard setting session. The panel was composed of insurance educators, state insurance representatives, insurance managers, and newly licensed agents. This mixture of judges provided broad experience in the insurance field including, individuals with many years of experience and those who had more recent experience with the

licensing procedure. Judges were randomly divided into three groups (n=7, n=6, n=6). Items were presented to the judges in blocks of 50 (i.e., by intact test form) with the blocks presented in a different random order for each group of judges. Each judge rated the entire pool of items.

The judges' ratings were investigated for mean differences by gender, ethnic group, the background of the judge (e.g., state insurance representatives versus new agents, educators, etc.), and the order in which the different groups of judges saw the test forms. A split-plot analysis of variance design was used to analyze the judges' ratings, where the variables described above (gender, ethnic group, background, and order of presentation) served as the between-judges factors and test item served as the within-judges factor. No between-judges factors, the primary interest of the investigation, were found to be significant. Thus, the judges' ratings did not appear to be influenced by their gender, ethnic background, experience or the order of item presentation.

Judgmental Item Rating. All judges were convened in one room for the rating session. The panel first defined the minimally licensable candidate. This definition was available to the panel throughout the rating process. Once the "minimally licensable candidate" was defined, the judges rated twenty representative items, for practice, from outside the item pool used in the study.

During the practice session, the correct answer and the percentage of the total group of candidates answering an item correctly (P+) from a previous administration were made available to the judges. The items were rated using the Angoff (1984) procedure.

During the actual rating session judges first rated the items without collateral information. Following this initial rating, the keys and the other judges' ratings were revealed. The judges with the highest and lowest ratings then gave the rationale for their ratings. General discussion of the items was permitted, but rarely occurred. Judges were given an opportunity to revise their ratings during the discussion.

Construction of New Test Forms Using the Distribution of Judges' Ratings as Specifications. The distribution of the judges' ratings for each of the six operational test forms (Forms A-F) was compared to the distribution of ratings in the entire pool to ascertain which of the operational forms most closely approximated the distribution of ratings in the pool (see Table 1, Forms A-F and Figures 1-1 to 1-6). Form D was selected. The cut score and the distribution of ratings for Form D then served as the "statistical" specifications for the construction of the experimental forms.

Insert Table 1 and Figures 1-1 to 1-6 about here

Three experimental forms (Forms G, H, & I) were constructed from

the pool of 228 remaining items after the target form had been removed. Each experimental form was constructed by selecting items from the pool according to the specified content plan and by targeting the Form D passing score and the distribution of Angoff ratings. Statistics based on past examinee performance and IRT parameters were not considered in the construction of the experimental test forms. The forms were constructed to have no overlap since this would provide the most stringent test for the construction of test forms.

The three experimental forms (Forms G, H & I) were then equated to Form D using IRT concurrent calibration methodology. In addition to the experimental forms, three randomly selected operational forms (Forms A, B, & F) were also equated to Form D to serve as a baseline for the amount of variation in cut scores that might be expected for forms constructed according to classical statistical specifications when item statistics were available for all items.

Results

The descriptive statistics reported in Table 2 indicate that test forms with approximately the same cut score as the target form could be constructed using the judges' estimates of item difficulty (Angoff estimates). The unrounded cut score for form D (target)

Insert Table 2 about here

was 38.55. Cut scores for the three experimental forms were 38.26 (Form G), 38.06 (Form H), and 38.49 (Form I). The judgmental cut scores for the experimental forms are within half a raw score point of the target form judgmental cut score.

Equivalence of judgmental cut scores and equated cut scores. Table 2 also presents the results from the equating of the test forms to the target form. The difference (equated - rated) observed for the *experimental* form cut scores based on item judgments are about one and a half raw score points below their corresponding equated values (Mean = 1.63, SD = 0.41), on average. This is also true for the *operational* test forms (Mean = 1.56, SD = 1.28) which had been assembled to classical statistical specifications --- a difficulty distribution and mean r-biserial correlation.

Consistency of Equated Cut Scores. The variation of the differences between the cut scores based on judges' estimates of item difficulty and their respective equated cut scores is less for the experimental test forms than that for the operational test forms (standard deviations of 0.41 and 1.28, respectively). The variation among the equated cut scores for the experimental and the operational forms is very similar. The standard deviation of the equated cut scores for the three experimental forms is 0.61 and that for the three operational forms is 0.47.

Discussion

The goal of this study was to determine whether test forms could be constructed from judgmental estimates of item difficulty to have practically equivalent cut scores to a target form (Form D). Whether this goal was met depends upon one's definition of "practical" equivalence.

Test forms were assembled to approximate the cut score of a target form using judgmental estimates of item difficulty; the judgmental cut score fell within about half a raw score point of the cut on the target form. This is probably about as close to equivalent as could be achieved in a typical licensure program; that is, a program that does not have a very large item pool.

When experimental form cut scores based on the judgmental ratings were compared to cut scores determined by equating, the judgmental cuts were found to be about one and a half raw score points below their corresponding equated cut scores (mean difference = 1.63 raw score points). Thus, the judgmental cut scores were only roughly equivalent, with the difference in one case exceeding two raw score points.

A control set of operational forms were also equated to the target form to examine whether the judgmental estimates of item difficulty might serve as proxies for empirical item difficulties when the

latter are not available during test construction. The size of the average difference between cut scores based on judgments and their corresponding equated values was about the same for the experimental and operational forms (mean difference = 1.63 and 1.56 raw score points, respectively). The variation among the equated cut scores was also about the same for the experimental and operational forms (standard deviations of 0.61 and 0.47, respectively). These results suggest that the judgmental estimates of item difficulty may serve as proxies for empirical item difficulties if the latter are not available or are unstable due to small sample size.

It should be noted that all of the items used in this study had acceptable classical and IRT statistics. If the pool had contained some items with unacceptable statistics, which would be expected if the items had not been pretested on a reasonable sample, test forms assembled on the basis of judgmental difficulty might be less acceptable.

Conclusions

The present study sought to determine the feasibility of constructing test forms with equivalent cut scores using the judgmental estimates of item difficulty from a large number of judges as target "statistical" specifications.

The use of judgmental ratings as "statistical" specifications proved to be as successful as classical statistics in the construction of parallel forms given content sampling constraints. However, comparisons between judgmental and equated cut scores showed the judgmental cut scores to differ from cut scores obtained through equating the test forms to a degree that would be acceptable only if sample sizes for equating are extremely small.

References

Angoff, W. H. (1984). Scales, Norms, and Equivalent Scores. Princeton: Educational Testing Service.

Wingersky, M. S., Patrick, R. & Lord, F. M. (1987). LOGIST User's Guide, LOGIST 6 [computer program]. Princeton: Educational Testing Service.

Table 1
Descriptive Statistics for
Judges' Ratings
by Form

Form Type	Form	N	Mean	Std Dev	Skew	Kurtosis
Operational	A	50	74.13	4.92	-0.57	-0.18
	B	50	74.74	6.00	-1.13	2.14
	C	50	75.02	4.31	0.26	0.30
	D	50	77.10	4.18	-0.48	0.77
	E	50	76.03	4.35	-0.58	0.04
	F	50	77.34	4.52	-0.22	-0.02
Experimental	G	50	76.52	5.71	-0.68	-0.14
	H	50	76.17	3.90	0.18	-0.49
	I	50	76.98	4.10	0.24	-0.79
All Items in Pool		278	75.70	4.95	-0.63	1.14
Items Remaining with Form D Items Removed		228	75.39	5.06	-0.61	1.10

Table 2
Descriptive Statistics for
Judgmental Cut scores and Equated cut scores
by Form

Form Type	Form	N	Mean Rating	Judgmental Cut score	Equated+ Cut score	Difference (E - J)
Operational	A	50	74.13	37.07	39.23	2.16
	B	50	74.74	37.37	39.71	2.43
	D	50	77.10	38.55	38.55	---
	F	50	77.34	38.67	38.76	0.09
Experimental	G	50	76.52	38.26	39.69	1.43
	H	50	76.17	38.06	39.43	1.37
	I	50	76.98	38.49	40.59	2.10
All Items in the Pool		278	75.70	37.85		
Items Remaining with Form D Items Removed		228	75.39	37.69		

+ Form D is the base form. The equated cut score value for Form D is computed directly from the ratings.

Figure 1-1
Plot of Ratings by Test Form

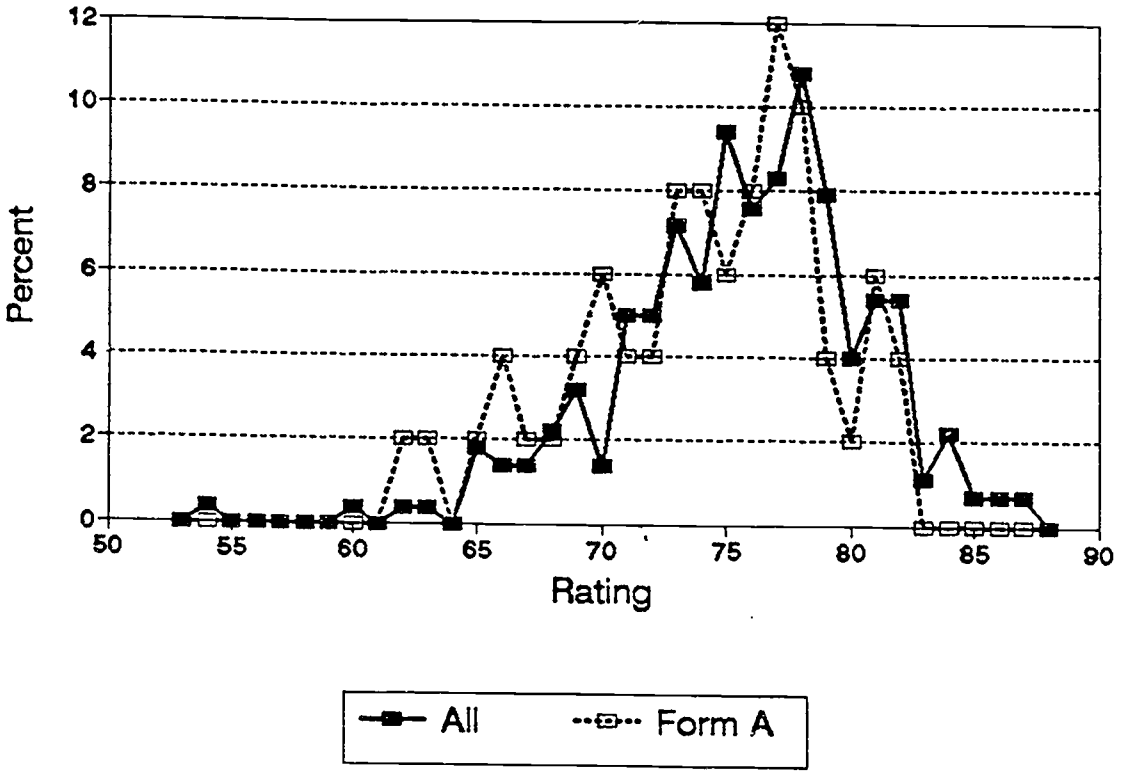

Figure 1-?

Plot of Ratings by ' st Form

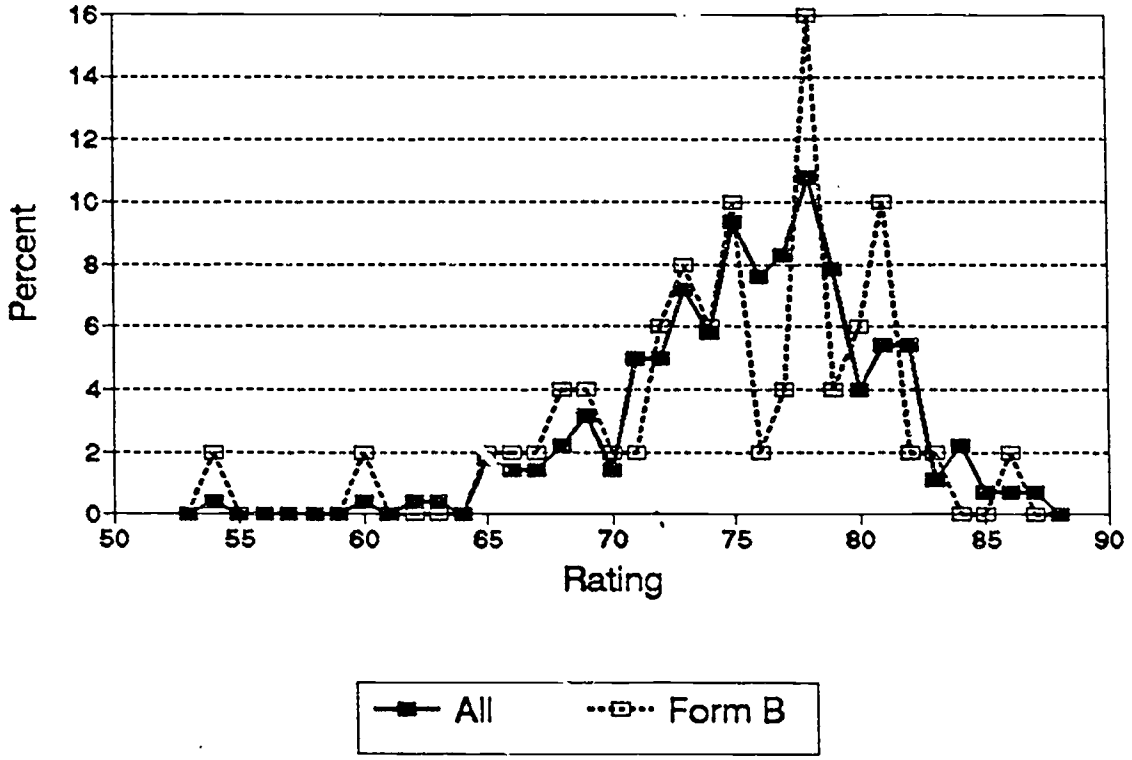

Figure 1-3
Plot of Ratings by Test Form

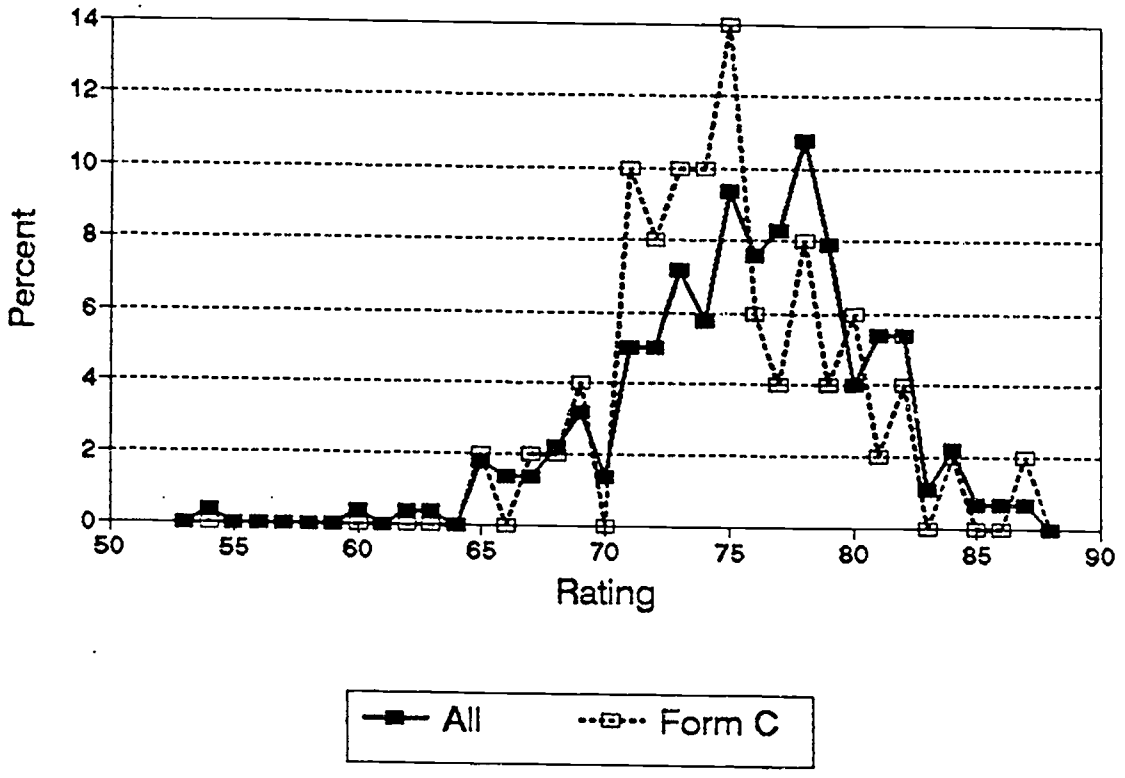

Figure 1-4
Plot of Ratings by Test Form

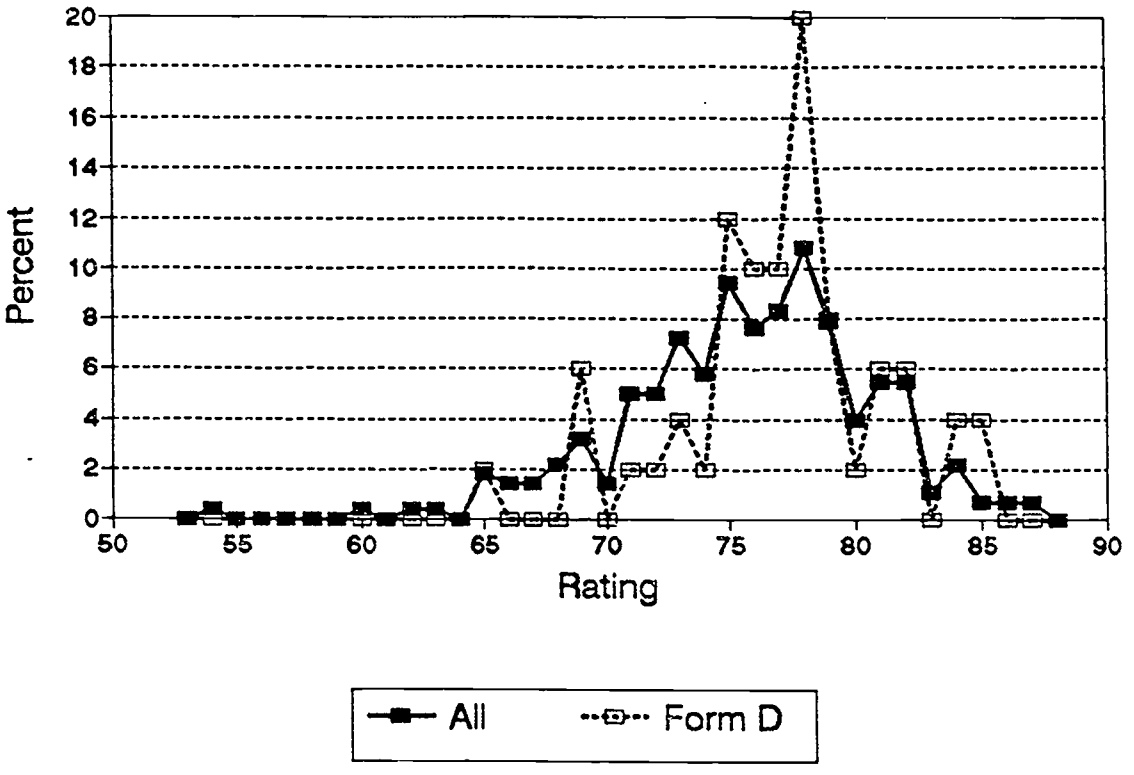

Figure 1-5
Plot of Ratings by Test Form

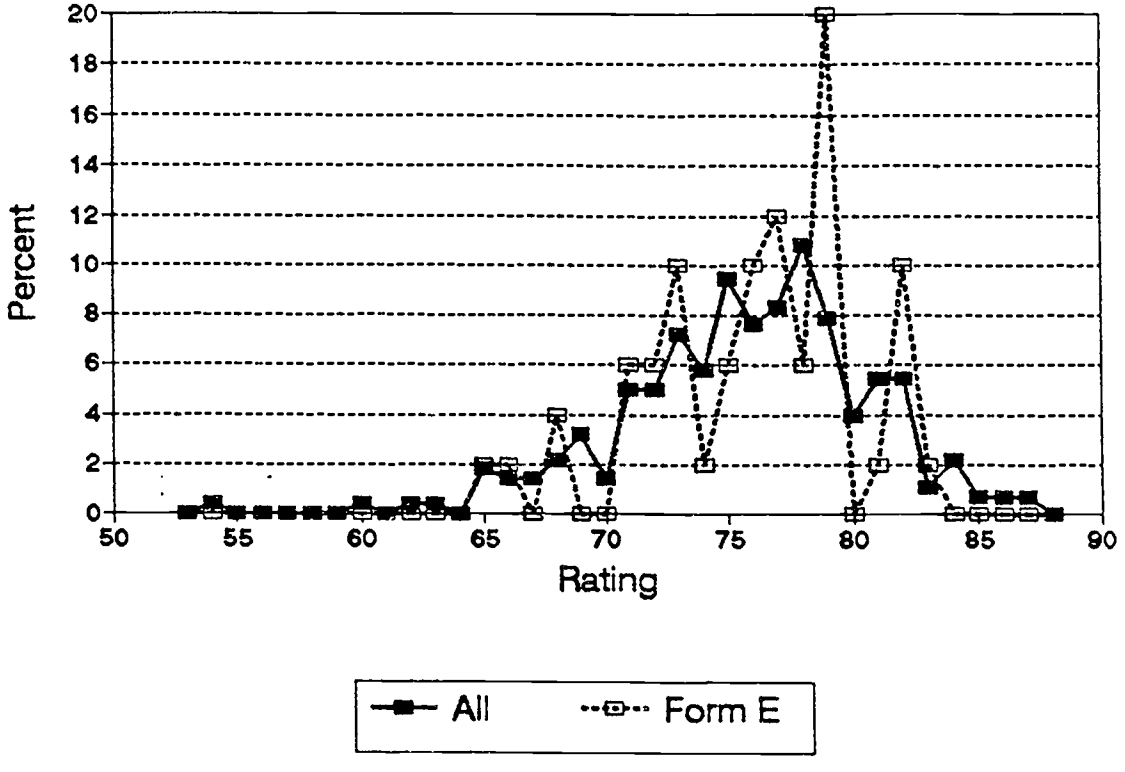

Figure 1-6

Plot of Ratings by Test Form