DOCUMENT RESUME

ED 320 944 TM 015 207

AUTHOR	Cramer, Stephen E.
TITLE	Some Practical Solutions to Standard-Setting Problems: The Georgia Teacher Certification Test Experience.
PUB DATE	Apr 90
NOTE	21p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Boston, MA, April 17-19, 1990).
PUB TYPE	Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Computer Assisted Testing; Cutting Scores; Data Collection; Elementary Secondary Education; Evaluators; *Licensing Examinations (Professions); *Scoring; *State Programs; Statistical Analysis; *Teacher Certification; *Testing Problems; Testing Programs
IDENTIFIERS	Angoff Methods; *Georgia Teacher Certification Testing Program; *Standard Setting

ABSTRACT

A standard-setting procedure was developed for the Georgia Teacher Certification Testing Program as tests in 30 teaching fields were revised. A list of important characteristics of a standard-setting procedure was derived, drawing on the work of R. A. Berk (1986). The best method was found to be a highly formalized judgmental, empirical Angoff procedure with iterations, in which content area specialists in each field would be provided, after each round of ratings, with field test results and an estimate of the effects of ratings. The data collection process used a computer-based data entry form, and each judge entered ratings onto an electronic form. At the conclusion of a round, the forms (spreadsheets) were merged with a master form that automatically calculated means and ranges of ratings and a passing score, and the computer created an output dataset to estimate passing rates and ratings for the panel. The method has been used to collect over 50,000 ratings from about 100 judges with no loss of data. In operation, the method was practical and could be administered to 20 judges by 2 staff members in a 2-day period. The method was sensitive to examinee performance data and appeared statistically sound. The method was approved by the State Board of Education, the Technical Advisory Committee, and the State Teacher Certification Testing Advisory Committee. Two tables and three figures supplement the text. (SLD)

SOME PRACTICAL SOLUTIONS TO STANDARD-SETTING PROBLEMS:

THE GEORGIA TEACHER CERTIFICATION TEST EXPERIENCE

Stephen E. Cramer, PhD
Division of Assessment
Georgia Department of Education

Presented at the annual meeting of the
National Council on Measurement in Education
Boston, MA
April 16, 1990

INTRODUCTION

As more and more states implement minimum competency certification and accountability programs for students and teachers, the issue of how to set the passing score for a test becomes more and more relevant. In spite of much research in the area, we are still far from agreement on the "right" way to perform this increasingly important task.

Indeed, some authorities would contend that there is no right method, but only a variety of methods that, since they yield different standards, must therefore all be wrong (e.g., Glass, 1978; Poggio, Glassnapp, & Eros, 1981; Linn, Madaus, & Pedulla, 1982). Others (e.g., Jaeger, 1989), recognizing that all standard-setting is judgmental, have sought to develop methods that, while perhaps not yielding "the right answer" (p. 492), provide "the best obtainable answer", and one that provides additional useful information to answer legislatively-mandated questions such as "Should this student receive a diploma?" or "Should this applicant receive a certificate to teach children?" Granted, this is "an artificial dichotomy imposed on a continuous test score distribution" (Berk, 1986). Nonetheless, it is possible to envision a person who does not know enough to teach a given subject, and also to envision someone who does. The passing score must lie between these two scores. The difficulty, of course, is finding it.

OBJECTIVES

Our task was to develop a standard setting procedure for use in the Georgia Teacher Certification Testing Program. We were in the process of doing complete revisions of tests in 30 teaching fields, and needed to have new passing scores. As the process was being developed, we were also in litigation with the Georgia Association of Educators, an NEA affiliate, over test validity and bias issues (This has since been settled, to the satisfaction of both parties). We needed, therefore, to develop a procedure that had bulletproof statistical merit and also a high level of validity. Face validity was an essential consideration, as well.

Berk (1986) catalogued 38 standard-setting methods, and provided a rating system usable by a "consumer" such as the director of a state testing program. The rating system ("aka Frequency of Despair Record") contains 10 criteria, grouped into the areas of Technical Adequacy and Practicability. While each of the methods reviewed had advantages and disadvantages, Berk's ratings appeared to indicate that procedures that incorporated both expert judgments and empirical information met the largest number of criteria.

Moreover, many of the criteria did not appear to be intrinsic to the methods. In other words, it appeared to be possible to treat

Berk's list as a buffet of techniques. This, then, was the goal, to create an eclectic method that incorporated the advantages of several methods while overcoming their disadvantages, to produce a standard setting method that was of technical quality, practical to administer, credible, and legally defensible.

A review of the recent literature and the advice of the Department of Education's Technical Advisory Committee indicated several characteristics that seemed to be important in creating a technically adequate, credible, and defensible standard-setting procedure. A non-exhaustive list (borrowing heavily from Berk) of these includes:

1. Sensitive to examinee performance (e.g., field test data)
2. Provides outcome information (i.e., pass rates)
3. Sensitive to training of raters
4. Statistically sound
5. Yields decision validity evidence
6. Provides opportunities for judges to adjust ratings
7. Includes relevant expert judges
8. Easy to implement
9. Easy to compute
10. Easy to interpret to laypeople (and Boards of Education)
11. Credible to all audiences

Based on these desiderata, we determined that the best method for our purposes would be a highly formalized judgmental-empirical Angoff procedure with iteraticns, in which content specialist judges in each test field would be provided with field test results and with an estimate of the effect of their ratings after each round of ratings.

Such a procedure obtains Angoff ratings made by relevant expert judges (7, above), and is relatively easy to implement (9, above), although it is challenging to convene a panel of raters that includes 25% teacher's union representatives, 25% minority representatives, and otherwise constitutes a valid statewide sample of educators in a given field. By using an iterative process with three rounds of ratings, judges are given the opportunity to adjust their ratings (6, above).

By including information on item p-values (discussed more fully below) and projected pass rates (also discussed below), the procedure is very sensitive to examinee performance and provides outcome information to the panel of judges (1 and 2, above). The judges are clearly sensitive to this information (3, above), and to feedback on the mean, high, and low ratings of the panel. This is indicated by the changes in passing score across the three ratings (see Table 1). Changes in the variability across the raters (see Table 2) demonstrate the convergence of ratings within the group, due in part to empirical data being added to the process, and in part to discussions in which outlying judges are encouraged to defend their ratings. Finally, although the method itself does not supply decision validity evidence (5, above), this can be obtained from the Subkoviac (Subkoviac, 1994) statistic calculated from the operational administration.

Insert Table 1 about here

Insert Table 2 about here

In the actual rating process, judges respond to the traditional Angoff question: "What percentage of minimally competent examinees would be expected to get this item correct?" They are permitted to give an item any integer rating between 0 and 100 percent. By not restricting judges to 20%, 25%, 33%, 50%, and 100%, which a Nedelsky (Nedelsky, 1954) rating task essentially does, or 10, 20,...90, 100, which some Angoff procedures do, the process allows the judges to make fairly fine distinctions among items and to make small adjustments from one round to the next.

These procedures appeared to deal with most of the disadvantages noted in the reported methods. However, they create problems of their own, the solutions of which will be the focus of the remainder of this discussion.

INNOVATIVE METHODOLOGY

1. Data Collection.

In iterative procedures, data must be collected repeatedly. In our process, we were dealing with approximately 150 items times 20 raters times 3 iterations, or about 9,000 ratings. We also needed to be able to provide rapid feedback to the raters. It seemed clear that ratings needed to be collected in machine readable form. Our prior Angoff methods had restricted raters to "multiple-choice" item ratings such as 10, 20,..., 90, 100 or even coarser divisions, since we were using scannable sheets.

Our new data collection process overcame this problem with a PC-based data entry form (Figure 1) using a spreadsheet program (AsEasy, Trius Software). Each judge used a computer to directly enter item ratings onto an electronic form. The forms are structured so that the rater can enter values into only one column at a time. At the conclusion of each round, these spreadsheets are merged with a master form (Figure 2), which automatically calculates means and ranges of ratings and a passing score, and creates both an output dataset to be used to estimate passing rates and the mean, high, and low ratings for the panel. The latter information is merged back into the judges' rating forms. Through the use of macros, this information can be provided to judges within 30 minutes of the completion of a round of ratings. Although several judges were complete computer novices, the method has been used to collect over 50,000 item ratings from about 100 different judges so far with zero loss of data.

Insert Figure 1 about here

Insert Figure 2 about here

2. Projection of pass rates.

The literature contains several references to the importance of providing feedback to judges on the effect of their ratings. In the present instance, such feedback would take the form of "Based on your ratings so far, XX% of the prospective teacher applicants taking this test will pass and receive a certificate; YY% will fail ana be denied."

For tests that have already been administered, this is easy to calculate. However, since we were setting standards for new test forms that had never been administered as a whole, the actual score distribution for the set of items being rated (the new form) was unknown. Only field test p-values for each item and the response vectors from the field test forms (four or five for each new form) were available.

The difficulty was overcome by an analysis program that used field test data from four administrations to create a unique test form and a passing score for each field test examinee. These forms differed in numbers of items and in passing scores, but this was not a problem, since the essential information for this purpose was the proportion of examinees above and below their defined passing score and, to a lesser extent, the distribution of examinees immediately around this score. We were also initially concerned that the field test items were not randomly distributed across forms, but since each unique form had a passing score determined by the ratings for items appearing on that form, this did not seem crucial.

The general logic of this program was as follows:

1. Begin with an incomplete data matrix of rows (examinees) and columns (items), based on results from all of the field test administrations. Values are 1 (right), 0 (wrong), and blank (did not take the item.

2. For each examinee, sum the item ratings for the items that appeared on the test form he took. Call this value CUT.
3. For each examinee, count the number of items that he got right that appear on the proposed operational form. In effect, this means to sum across each row. Call this value RAW.
4. For each examinee, subtract CUT from RAW. If this value is zero or greater (i.e., at or above his cut score), the examinee is a projected pass; set the value of a variable called PASS to 1. If the value is negative (i.e., RAW < CUT), set PASS = 0.
5. Calculate the frequency distribution of PASS; the percentage of 1's is the projected percentage pass rate.

The actual information presented to the judges was the projected percentage pass figure and a cumulative frequency distribution graph (Figure 3) of examinees at each point above and below the passing score, without regard to actual total score. It is clear that this information was very important to the judges. Not only did it force them to look at their item ratings in a new light, but it invariably occasioned a serious, sometimes passionate discussion of what their expectation of minimal competence really consisted of, and how essential it was to insist on that standard in their own field.

These estimates were found to be close to the actual obtained percentage of examinees passing the operational form (See Table 1), although they appear to generally underestimate the operational pass rate. This may be explained by the fact that the program does not account for examinees who failed the test, perhaps repeatedly, on the administrations in which field testing was being carried out. Plans are underway to evaluate the use of only first-time test takers for this analysis in the future.

3. Conditional p-values of minimally competent examinees.
One of the most difficult things to explain to Angoff judges when providing empirical data is the interpretation of empirical p-values. In essence, we say to the judge: "Here is a measure of how well all examinees perform on this item, but this is not what you are to estimate. Your task is to tell us how well a select group, the minimally competent, will perform. Their average score on this item is (probably) below the total group p-value, but we don't know how far below."

As an aside, there was one group of raters who challenged the assumption made in the previous sentence. "How do you know," they asked, "that we don't consider the minimally competent educator to be above the average level of competence of the goup of people who take this test?" As it turns out, their operational (by our method) definition of minimally competent corresponds roughly to the 25th percentile, well below the average.

In an effort to simplify this task, we tried using the first and second round ratings to identify examinees who scored at or near their unique passing score that was calculated as described above. We labeled these people "hypothetically minimally competent." We then computed p-values for this group and presented this information to the judges with instructions like "These figures show how well the people that your first ratings identify as 'minimally competent' actually performed. Taking these data into consideration, re-rate each item."

Although this technique may have potential in certain situations, and had definite impact on judges' ratings, in our initial attempt to implement it, we ran into severe problems. In the first round of ratings for the field of Administration and Supervision, the judges set a standard that would have passed only about 50% of the examinees. Since the distribution of raw scores for these tests is strongly negatively skewed (i.e., most examinees score high), the process identified examinees whose scores were considerably above the mean as hypothetically minimally competent. Thus, the p-values repoıced for the "hypothetically minimally competent examinees" were actually far above the p-values for all examinees. This had the effect of causing the judges to re-rate the items higher in their second round, resulting in a still lower pass rate.

At this point, we discovered what was going on, and explained the situation to the judges before they completed their third and final ratings. Based on these final ratings, we projected the pass rate based on the third round of ratings to be about 96%. This massive fluctuation indicated to us that the judges had overshot in their attempt to recover from Rounds 1 and 2. We therefore invalidated the results. A subsequent standard setting procedure for this field, omitting the "hypothetically minimally competent" feedback, resulted in an 80% pass rate.

In reconsidering this process, it now appears that there is a basic flaw in its logic. Although the goal of the "hypothetical minimally compe-ent p-value" information is to bring the raters more in line with real data, it will tend to have the opposite effect. If the initial ratings are extremely high, the minimally competent p-values calculated based on this standard will tend to indicate that minimally competent examinees perform even better than the raters estimated, which leads the raters to increase their standard in round two. Conversely, if the initial ratings are low, the minimally competent p-values will tend to look lower still, and depress the standard further.

CONCLUSION

The methods discussed above, and others, have been used successfully to set new passing scores for five Georgia Teacher Certification Test fields so far. In operation, the methodology appears to meet most of the desiderata listed above. It is highly practical, capable of being administered to 20 judges by two staff members in a two day period, although it requires the availability of several computers and mainframe access. It is highly sensitive to examinee performance information, yields appropriate classification information, and appears to be statistically sound. The methodology is credible enough to have been approved by our State Board of Education, our Technical Advisory Committee, and our State Teacher Certification Testing Advisory Committee, and to be well accepted by representatives of the major teacher union.

References

Berk, R.A. (1986) A consumer's guide to setting performance standards on criterion-referenced tests. RER, 56(1), 137-172.

Glass, G V (1978) Standards and criteria. JEM, 15, 237-261.

Jaeger, R.M. (1989) Certification of student competence. In Robert L. Linn, Ed., Educational Measurement (3rd edition). Washington, D.C.: American Council on Education.

Linn, R.L., Madaus, G., & Pedulla, J. (1982) Minimum competency testing: Cautions on the state of the art. AERJ, 91, 1-35.

Nedelsky, L. (1954) Absolute grading standards for objective tests. EPM, 14, 3-19.

Poggio, J.P., Glasnapp, D.R., & Eros, D.S. (1981, April). An empirical investigation of the Angoff, Ebel, and Nedelsky standard-setting methods. Paper presented at the annual meeting of the American Educational Research Association, Los Angeles.

SAS Institute Inc. SAS User's Guide: Basics, Version 5 Edition. Cary, NC: SAS Institute Inc., 1985.

Subkoviak, M.J. (1976) Estimating reliability from a single administration of a mastery test. JEM, 13, 265-276.

Trius Software (1989) As-Easy-As (Ver.3.00) 231 Sutton Street, Suite 2D-3, North Andover, MA, 01845.

Table 1: Changes in projected passing rates over three rounds of item judgments

Field	Round 1	Round 2	Round 3	Actual
Early Childhood	58.2	76.0	83.9	87.0
Middle Childhood	91.3	90.5	86.9	89.3
Mental Handicaps	56.6	60.0	80.3	83.6
Interrelated Special Education	67.5	69.1	69.7	***
Spanish	88.2	80.9	79.4	76.5*
Counseling	72.4	71.1	70.1	77.0
Administration & Supervision	50.8**	49.5**	95.8	79.7

*--includes production items not considered in this process

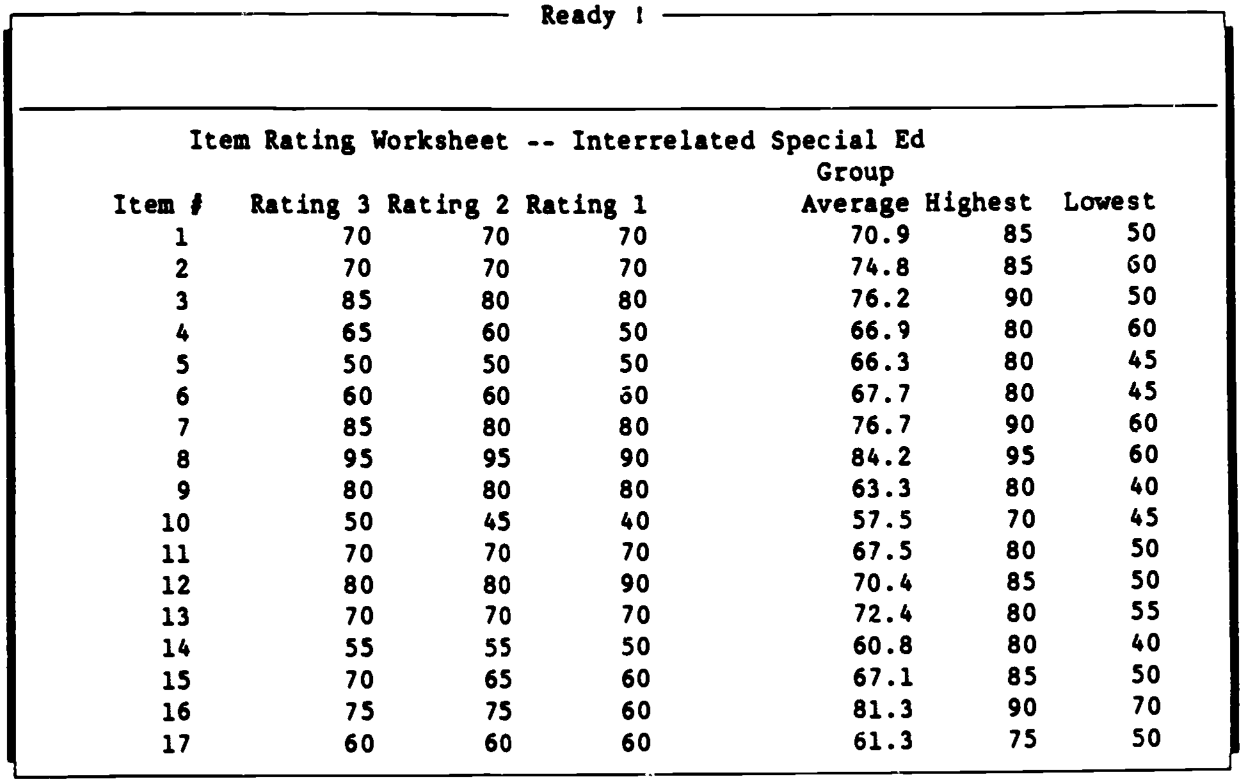
**--using "hypothetical minimally competent" p-values

***--in process; available 3/30/90

Table 2: Changes in a /erage variability of ratings across all raters over three rounds of item judgments

Field	Round 1	Round 2	Round 3
Early Childhood	13.42	10.11	9.25
Middle Childhood	17.14	12.89	12.03
Industrial Arts	18.02	13.71	12.53
Interrelated Special Education	13.39	8.47	7.35
Spanish	11.01	6.14	5.92
Counseling	13.85	9.40	9.26
Administration & Supervision	13.69		5.13

Figure 1: Screen dump of electronic item rating form used by judges. Seen after third round of ratings.

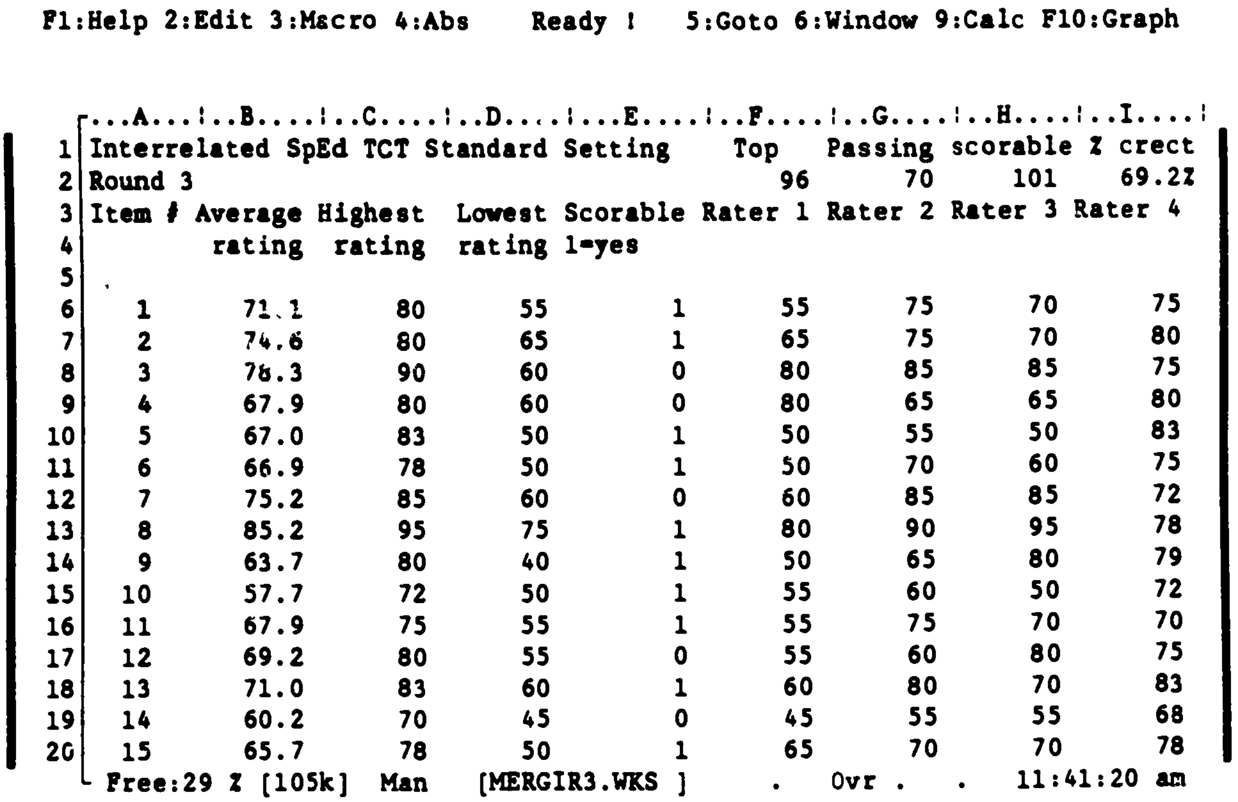


```
─────────────────── Ready ! ───────────────────



          Item Rating Worksheet -- Interrelated Special Ed
                                                   Group
  Item #   Rating 3 Rating 2 Rating 1            Average Highest  Lowest
     1          70       70       70               70.9      85      50
     2          70       70       70               74.8      85      60
     3          85       80       80               76.2      90      50
     4          65       60       50               66.9      80      60
     5          50       50       50               66.3      80      45
     6          60       60       60               67.7      80      45
     7          85       80       80               76.7      90      60
     8          95       95       90               84.2      95      60
     9          80       80       80               63.3      80      40
    10          50       45       40               57.5      70      45
    11          70       70       70               67.5      80      50
    12          80       80       90               70.4      85      50
    13          70       70       70               72.4      80      55
    14          55       55       50               60.8      80      40
    15          70       65       60               67.1      85      50
    16          75       75       60               81.3      90      70
    17          60       60       60               61.3      75      50
```

Figure 2: Screen dump of electronic item rating form used to merge judges' ratings and calculate passing scores. Seen after third round of ratings.



```
F1:Help 2:Edit 3:Macro 4:Abs    Ready !   5:Goto 6:Window 9:Calc F10:Graph

  r...A...!..B....!..C....!..D....!...E....!..F....!..G....!..H....!..I....!
 1|Interrelated SpEd TCT Standard Setting    Top   Passing scorable % crect
 2|Round 3                                    96      70      101    69.2%
 3|Item # Average Highest  Lowest Scorable Rater 1 Rater 2 Rater 3 Rater 4
 4|        rating  rating  rating 1=yes
 5| .
 6|   1     71.1      80      55        1      55      75      70      75
 7|   2     74.6      80      65        1      65      75      70      80
 8|   3     76.3      90      60        0      80      85      85      75
 9|   4     67.9      80      60        0      80      65      65      80
10|   5     67.0      83      50        1      50      55      50      83
11|   6     66.9      78      50        1      50      70      60      75
12|   7     75.2      85      60        0      60      85      85      72
13|   8     85.2      95      75        1      80      90      95      78
14|   9     63.7      80      40        1      50      65      80      79
15|  10     57.7      72      50        1      55      60      50      72
16|  11     67.9      75      55        1      55      75      70      70
17|  12     69.2      80      55        0      55      60      80      75
18|  13     71.0      83      60        1      60      80      70      83
19|  14     60.2      70      45        0      45      55      55      68
20|  15     65.7      78      50        1      65      70      70      78
  L Free:29 % [105k]  Man   [MERGIR3.WKS ]     .  Ovr .   .   11:41:20 am
```

Figure 3: Cumulative pass-fail percentages--Third round ratings
Based on field test statistics for total examinee population

Interrelated Special Education TCT

```
                                SAS
CUMULATIVE PERCENTAGE BAR CHART

DIFF   PTS FROM PASSING      FREQ    CUM.   PERCENT    CUM.
                                     FREQ            PERCENT

        ]
  -7    ]*                     13      34     1.23      3.21
        ]
  -6    ]*                     22      56     2.08      5.28
        ]
  -5    ]**                    37      93     3.49      8.77
        ]
  -4    ]**                    32     125     3.02     11.79
        ]
  -3    ]***                   56     181     5.28     17.08
        ]
  -2    ]****                  51     232     4.81     21.89
        ]
FAIL    ]******                89     321     8.40     30.28
        ]
PASS    ]********             105     426     9.91     40.19
        ]
   1    ]**********           127     553    11.98     52.17
        ]
   2    ]*************        143     696    13.49     65.66
        ]
   3    ]***************      120     816    11.32     76.98
        ]
   4    ]*****************    105     921     9.91     86.89
        ]
   5    ]*******************   85    1006     8.02     94.91
        ]
   6    ]********************  47    1053     4.43     99.34
        ]
   7    ]********************   7    1060     0.66    100.00
        ]
        ----+---+---+---+---+
           20  40  60  80  100

         CUMULATIVE PERCENTAGE
```