DOCUMENT RESUME

ED 398 284 TM 025 628

AUTHOR O'Neill, Thomas R.; Lunz, Mary E.

TITLE Examining the Invariance of Rater and Project Calibrations Using a Multi-facet Rasch Model.

PUB DATE Apr 96

NOTE 14p.; Paper presented at the Annual Meeting of the American Educational Research Association (New York, NY, April 8-12, 1996).

PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.

DESCRIPTORS Ability; Comparative Analysis; *Difficulty Level; Equated Scores; Estimation (Mathematics); *Interrater Reliability; *Item Response Theory; Judges; Probability; *Scoring; *Test Items; Test Results

IDENTIFIERS Benchmarking; *Calibration; Invariance; Polytomous Variables; *Rasch Model

ABSTRACT

To generalize test results beyond the particular test administration, an examinee's ability estimate must be independent of the particular items attempted, and the item difficulty calibrations must be independent of the particular sample of people attempting the items. This stability is a key concept of the Rasch model, a latent trait model of probabilities that permits items and persons to be analyzed independently, yet still be compared using a common frame of reference. An extension of the Rasch model, the multi-facet Rasch model, can estimate examinee ability, item difficulty, and other facets for polytomous data. It was hypothesized that the multi-facet Rasch model would yield invariant, sample-free, slide and judge calibrations for a certification test for histology completed by 364 candidates. Eighteen qualified judges graded the test, which required examinees to prepare laboratory slides. Results of the study confirm that the slide and judge calibrations were essentially stable across diverse samples of examinees. This indicates that slide and judge calibrations can be used to anchor test administrations to a benchmark scale, making the equating of two administrations of the examination possible and supporting the hypothesis. (Contains 2 figures, 2 tables, and 15 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND
DISSEMINATE THIS MATERIAL
HAS BEEN GRANTED BY
THOMAS R. O'NEILL

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)

ED 398 284

Examining the Invariance of Rater and Project Calibrations Using a Multi-Facet Rasch Model

Thomas R. O'Neill and Mary E. Lunz

American Society of Clinical Pathologists

Paper presented at the Annual Meeting of the American Educational Research Association, New York, NY. April 1996.

NO25628

BEST COPY AVAILABLE

Version: March 21, 1996

Examining the Invariance of Rater and Project Calibrations Using a Multi-Faceted Rasch Model

Introduction

Some things cannot be adequately measured using a multiple choice test. For example, it would be absurd to have a multiple choice test for a diving competition, a karate test, or a needlepoint contest. In order to measure these abilities, examinees must perform the task, rather than answer questions about the task. While the actual performance of the task in question enhances the face validity of the test, judges must be included to assess the quality of the performance. The introduction of a third facet affects the reproducibility of the results. If a judge's decision is treated as absolute, there is some probability that different judges make different decisions about the same candidate performance. This means the judges and the examinees are not really independent of each other, ergo interpretation of a score is impossible without relating it to the specific judge.

Rosenthal and Rosnow (1984) contend that in practice "usually there is no special interest in individual differences among observers or judges when we consider issues of interobserver or interjudge reliability. We simply decide on the type of judges or observers we want (college students, clinical psychologists, linguists, mothers, etc) and then regard each observer or judge within that sample as to some degree equivalent to, or interchangeable with, any other observer or judge within the sample" (p. 162). This approach is feasible if judges' ratings of the same performance are comparable. However, even with extensive training, interjudge reliability correlations have often been low (Michael, Cooper, Shaffer & Wallis, 1980), causing some serious concerns about the actual interchangeability of judges based upon those scores (Lunz & Stahl, 1990; Lunz, Wright, & Linacre, 1990; Lunz, Stahl, & Wright, 1994). If all examinees are scored by the same judges, the influence of judge severity is equal across all examinees. However, it is often impractical for all judges to grade all tests. For instance, few judges would have the time to grade an essay for every fourth grader in Chicago, much less Illinois. This also assumes the judge can remain self-consistent when grading a large number of essays.

In order to generalize test results beyond the particular test administration, an examinee's ability estimate must be independent of the particular items attempted, and the item difficulty calibrations must be independent of the particular sample of people attempting the items. This stability is a key concept of the Rasch model (Rasch, 1980). The Rasch model is a logistic latent trait model of probabilities which permits items and persons to be analyzed independently, yet still be compared using a common frame of reference. The Rasch model calibrates all items onto a single scale, regardless of item format or classification system, provided that they represent a single conceptual domain. Wright and Masters (1982) assert that "the method by which observations are turned into calibrations and measures must contain the possibility of *invariance* over a useful range of time and place. It must also provide the means for verifying that a useful approximation to this invariance is maintained in practice"

BEST COPY AVAILABLE

(p. 5). The degree to which examinee measures and item, judge, project, or task calibrations are invariant will limit the extent to which those associated values will have meaning over time and across contexts.

It has been demonstrated that the two facet Rasch model yields invariant or stable item calibrations and person measures. Wright (1967) demonstrated that comparable item calibrations are derived from two mutually exclusive groups of examinees with different mean abilities. He also demonstrated that equivalent person ability measures are derived regardless of which particular items are used to compute ability. Rudner (1983) found similar results.

An extension of the Rasch model, the multi-facet Rasch model (MFRM, Linacre, 1989), can estimate, examinee ability, item difficulty, and other facets, such as judge severity, grading session severity and/or task difficulty, etc. for polytomous data. As in the basic Rasch model, all elements within a facet are expected to maintain the same relative position on the scale. Thus, judge severity estimates should be comparable (within the error of measurement) across different forms of the test and across administrations. Using the multi-facet Rasch model to analyze three types of examinations (essay, oral, and clinical), Lunz and Stahl (1990) found that most judges are reasonably consistent in their individual level of severity. The invariance of judge severity calibrations was tested when differences in examinee attributes are extreme, and found to be generally stable (Lunz, Stahl, & Wright, in press).

It is hypothesized that the multi-facet Rasch model yields invariant, sample-free, slide and judge calibrations. Thus, comparable judge severity and slide difficulty calibrations are expected even when the calibrations are based upon examinee groups of significantly different ability levels.

Methods

Instruments

A certification examination for histology, the branch of anatomy concerned with "the minute structure of cells, tissues, and organs" (Stedman, 1982, p. 652) was used to collect data. Examinees were required to submit an examination that included 15 types of histology slides that met defined specifications regarding tissue type, stain, processing and microtomy.

Subjects

There were 364 candidates for certification in histology. All examinees had to meet specific educational and experience requirements to be eligible to submit a test. All examinees also had to sign a pledge of independent workmanship that was co-signed by their employer or supervising pathologist.

There were 18 qualified judges, who graded the test. Before grading the slides, the judges spent three hours in an orientation session to review the grading criteria and sample slides. Sixteen of the judges had prior experience grading this examination.

Scoring Plan

The examination required examinees to prepare 15 laboratory slides to detailed specifications. Each slide was graded on three tasks. Two tasks were rated 0/1 (pass or fail) and the third was rated 0, 1, 2, 3, where 0 is unacceptable, 1 is marginal, 2 is satisfactory, and 3 is excellent. Since it would be impractical for all judges to rate all slides for all tasks for all candidates (15 slides x 3 ratings per slide x 364 examinees =16,380 ratings per judge) candidate performances were allocated to judges using a rotation plan.

Each examinee's slides were grouped into three batches of five slides (1-5, 6-10, 11-15). The rotation plan enabled each judge to grade each of the 15 slides on all three tasks at sometime during the grading session. Each examinee's performance was rated by a random combination of three judges who each graded one subset of five slides. Although each slide was rated only once, three judges provided ratings for the candidate's overall performance. Using this rotation plan, slides were rated by only one judge so that the examination structure was 15 slides x 3 tasks x 1 judge= 45 ratings per examinee. Common tasks and slides provided the necessary linkage to connect all facets. Examinees had all slides, all tasks, and some judges in common.

Multifacet Rasch Analysis

The ratings for the 364 examinees were scored using FACETS, a computer program for the analysis of exams with 3 or more facets (Linacre, 1994). The units of measurement are expressed in log-odds units, or logits. In this analysis, the probability of person n with ability B_n achieving a rating of x on slide s with difficulty D_s on task t with complexity C_t from judge j with severity S_j was modeled as:

$$\log\left(\frac{P_{nstjx}}{P_{nstjx-1}}\right) = B_n - D_s - C_t - S_j - F_x$$

where:

P_{nstjx} = Probability of examinee receiving score x by judge j on slide s on task t.
$P_{nstjx-1}$ = Probability of examinee receiving score $x-1$ by judge j on slide s on task t.
and
Bn = Ability of examinee n (Facet 1)
Ds = Difficulty of slide s (Facet 2)
Ct = Complexity of task t (Facet 3)
Sj = Severity of judge j (Facet 4)
Fx = Difficulty of achieving score x relative to score $x-1$ on the rating scale.

Three analyses were completed. The initial FACETS analysis generated calibrations for slide difficulty, task complexity, and judge severity, as well as measures for examinees using the total (N= 364) population. A median split of examinee ability measures was used to divide the examinees into two groups, low ability and high ability. A t-test was used to ascertain the difference in ability between the groups. The low and high ability groups were then used in two subsequent analyses to recalibrate judge severities, slide difficulties, and task complexities separately. For both slides and judges, the high group calibrations were plotted against the low group calibrations, and Z-score analysis was used to identify significant differences in severity or difficulty estimates.

Fit statistics were reviewed to assure that the data was appropriate to be used with the Rasch model. It is important to ensure that 1) the data, as a whole, adequately fit the model and 2) no element (an individual examinee, judge, or slide) deviates so far from the model that the measure or calibration no longer adequately summarizes the responses that comprise that element vector. Fit statistics are chi-square based "unexpectedness" indicators for the responses within a particular facet.

Results

The mean ability estimate for the low ability group was 1.15 logits and the mean ability estimate for the high ability group was 2.31 logits. The two groups had statistically different variances, F (181,181) = 4.181, p = .042, thus the unequal variances t-test formula was used. The difference in ability levels between the two groups was statistically significant (t = 23.62, df=328, $p < .001$).

Judges

The judge severity calibrations for the high and low groups are listed in Table 1. Individual judges differed in their level of severity; however, 94% of the judges had statistically comparable severity estimates, across ability groups. Figure 1 shows the plot of judge severity estimates. The diagonal line is an identity line, which represents where points would fall if the calibrations were identical. Around the identity line, confidence bans (2 SEs) were plotted (Luppescu, 1995). The Z-score analysis shows that for the distribution of judge calibration differences, only judge 17's difference was greater than expected. Judge 17 was significantly more lenient than expected (more than two standard deviations) when grading the low ability group. This judge was a new grader and apparently inclined to give less able performances the "benefit of the doubt". The mean of the calibration differences, .001 logits, suggests that the calibration differences as a whole were not biased toward either group. The Infit and Outfit statistics confirm that judges were generally consistent in how they graded (criteria MS = .5 to 1.5).

Insert Table 1 and Figure 1 about here

Slides

The slide difficulty calibrations are listed in Table 2. Although, the slides differed in difficulty, 93% of the slides maintained statistically comparable difficulty estimates across ability groups. Figure 2 shows the plot of slide difficulty estimates. The diagonal line is an identity line, which represents where points would fall if the calibrations were identical. Around the identity line, confidence bans (2 SEs) were plotted (Luppescu, 1995). The Z-score analysis shows that only the calibration for slide 14 varied. Slide 14 was significantly easier (more than two standard deviations) for the high ability group. The mean of the calibration differences, -.004 logits, suggests that the calibration differences as a whole were not biased toward either group. Using the fit criteria of MS= 0.5 to 1.5, all slides were within range.

Insert Table 2 and Figure 2 about here

Discussion

The goal of measurement is to determine how much ability an individual has. The practice of measurement involves making specific comparisons, such as comparing the side of a board to the marks on a ruler or comparing a person's performance on a math test to the difficulty of the items. If the measurements or calibrations are to be useful or generalizable, they must be stable. The idea of length would not be useful, if the markings on a tape measure changed depending on where or when the measurement was taken or what was being measured. The markings on the yardstick must not change as different objects are being measured, nor should the object change size depending on whose yardstick is used.

The purpose of this paper was to demonstrate that multi-facet Rasch model analysis yields comparable judge severity and slide difficulty calibrations across diverse samples of examinees. The calibrations computed for slides and judges are not infinitely precise, they are estimates. No real items or judges will be perfectly consistent in difficulty or severity over all time and across all occasions. While some variation in a probabilistic model is expected, excessive variation can limit the usefulness and generalizability of the results.

The results of this study confirm that slide and judge calibrations are essentially stable across diverse samples of examinees. This indicates that slide and judge calibrations can be used to anchor test administrations to a benchmark scale, thereby equating two administrations of a performance examination. Thus, the same criterion standard can be used for subsequent administrations of a performance examination, so that candidates have a comparable opportunity to pass regardless of the test administration in which they participate.

Sometimes item calibrations change over time and across geography because the meaning behind the question has changed, although the text of the question has not. For example,

BEST COPY AVAILABLE

"What is HIV?" would have been a difficult biological sciences question in 1980, but by 1987, it became a relatively easy current events question. A geographic example is, "What is the capital of Maine?" Obviously, this question will be easier for people living in Augusta, Maine than those living in California. When the meaning of an item changes over time or across regions, its use should be reconsidered because its meaning will be idiosyncratic. Items on performance examinations should be monitored for idiosyncratic behavior, just as items on multiple examinations.

In this study, slide difficulty was found to be comparable across ability groups. This was expected because slides are similar to multiple choice items in that the prompt (the question or the project specifications) is written and does not change across administrations. However, it is possible that a new technology could suddenly become widely available that would make some items easier. Under these circumstances, the prompt would be the same, but the meaning behind the successful completion of the project would be different. For example, 'Bone' is a difficult slide to prepare because bone is brittle. It is difficult to get a cross-section 5 micrometers thick without the fragment shattering. Should a new blade or preparatory process suddenly become available that makes cutting bone easy, then the successful completion of this slide will become significantly easier and the calibrated difficulty of the slide might change. If this happens, the slide should be re-calibrated.

Unexpected judge ratings should also be considered. In this paper, judge severity was found to be comparable even when examinee ability was significantly different. It is possible that in other tests, where the judges receive less training, or perhaps don't agree on the meaning of the grading criteria, the judges may be less self-consistent. For example, judges may be less severe on poor performers because they wanted to "give them a break", or conversely judges may be more severe on poor performers because "this examinee obviously doesn't deserve full credit". Similar checks could be run for gender, ethnicity, or other extraneous variables.

While this paper supports the stability of judge and slide calibrations, it is also important to check the requirement that judges are self-consistent. The internal consistency of the judges is verified by examining the fit statistics (deviation from expected). No difference between observed and expected is anticipated. If judges rate a slide or a candidate differently from the way he/she grades the others, a misfit will occur. If the judge is not internally consistent, that judge may need some additional training. All judges met the expectations for self-consistency in this study further demonstrating that while judges are different from each other, they are consistent within themselves.

Summary

In this study, the judge and slide calibrations were found to be stable across examinee groups of significantly different ability. This supports the premise that multi-facet Rasch analysis calibrations are independent of the particular sample from which they were derived. This is consistent with other studies (Wright, 1967; Rudner, 1983). Lunz, Stahl, and Wright (in

BEST COPY AVAILABLE

press) tested the invariance of judge severity and item difficulty across ability strata and found that judges generally maintain the same level of severity across exam administrations, although new judges may vary more due to lack of experience or unfamiliarity with the grading practices.

References

Linacre, J. M. (1989). *Many-Facet Rasch Measurement*. MESA Press: Chicago.

Linacre, J. M. (1994). *A User's Guide to FACETS: Rasch Measurement Computer Program*. MESA Press: Chicago.

Lunz, M.E., and Stahl, J.A. (1990). Judge consistency and severity across grading periods. *Evaluation & the Health Professions, 13* (4) p. 425-444.

Lunz, M.E., Wright, B.D., and Linacre, J. M. (1990). Measuring the impact of judge severity on examination scores. *Applied Measurement in Education, 3* p. 331-345.

Lunz, M.E., Stahl, J.A. and Wright, B.D. (In press). Invariance of judge severity calibrations. In G. Engelhardt (Ed.), *Objective Measurement: Theory into Practice, volume 3*. Ablex Publishing: Norwood, NJ.

Lunz, M.E., Stahl, J.A. and Wright, B.D. (1994). Interjudge reliability and decision reproducibility. *Educational and Psychological Measurement, 54* (4) p. 913-925.

Luppescu, S. (1995). Comparing measures. *Rasch Measurement Transactions, 9*, p.410-411.

Michael, W. B., Cooper, T., Shaffer, P., and Wallis, E. (1980). A comparison of the reliability and validity of ratings of student performance on essay examinations by professors of English and by professors in other disciplines. *Educational and Psychological Measurement, 40* p.183-195.

Rasch, G. (1980). *Probabilistic Models for Some Intelligence and Attainment Tests*. MESA Press: Chicago.

Rosenthal, R. and Rosnow, R. L. (1984). *Essentials of Behavioral Research: Methods and Data Analysis*. McGraw-Hill: New York.

Rudner, L. M. (1983). A closer look at latent trait parameter invariance. *Educational and Psychological Measurement, 43*, p.951-955.

Stahl, J.A. and Lunz, M.E. (1991). *Judge performance reports: Media and message*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Stedman, T. L. (1982). *Illustrated Stedman's Medical Dictionary 24th Edition*. Williams & Wilkins: Baltimore.

Wright, B. D. (1967). *Sample free test calibration and person measurement*. Paper presented at the Invitational Conference on Testing Problems. Educational Testing Service, NJ.

Wright, B. D. & Masters, G. N. (1982). *Rating Scale Analysis: Rasch Measurement*. MESA Press: Chicago.

Table 1. Comparison of Judge Severity Calibrations for High and Low Ability Groups

	High Ability Group				Low Ability Group					
Judge #	Severity	SE	Infit MS	Outfit MS	Severity	SE	Infit MS	Outfit MS	Calibration Difference	Difference Z-score
1	.11	.10	0.8	0.8	.11	.08	0.8	0.8	.00	-.01
2	-.09	.09	0.8	0.8	.01	.07	0.8	0.8	-.10	-.57
3	-.27	.10	1.3	1.1	-.38	.09	1.1	0.9	.11	.61
4	-.09	.11	0.9	1.1	.09	.08	0.9	1.0	-.18	-1.02
5	-.34	.13	1.0	0.7	-.11	.07	1.1	0.9	-.23	-1.30
6	-.03	.12	1.0	1.1	.14	.09	0.9	1.0	-.17	-.97
7	.17	.10	0.8	0.8	.29	.08	0.8	0.8	-.12	-.68
8	-.07	.13	0.9	0.6	-.08	.09	0.9	0.8	.01	.05
9	-.09	.11	0.9	0.9	-.05	.08	0.9	0.9	-.04	-.23
10	-.46	.13	1.0	0.7	-.27	.10	0.9	0.9	-.19	-1.08
11[1]	-.07	.11	0.8	0.8	-.21	.07	0.9	0.9	.14	.78
12	.30	.10	0.7	0.9	.17	.08	0.6	0.9	.13	.73
13	-.26	.12	0.8	0.7	-.23	.07	0.9	0.9	-.03	-.18
14	.74	.07	0.7	1.1	.60	.07	0.6	0.8	.14	.78
15	-.22	.11	0.8	0.8	-.06	.08	0.9	1.0	-.16	-.91
16	.36	.12	0.7	1.0	.37	.09	0.9	1.1	-.01	-.06
17	.39	.09	0.7	0.9	-.01	.10	1.8	0.8	.40	2.25[1]
18	-.08	.11	0.9	0.9	-.40	.09	1.0	0.9	.32	1.80
								Mean =	.001	

[1] While judges calibrated slightly different across examinee groups, the difference in Judge 17's calibrations was significantly greater than that of the other judges.

Table 2. Comparison of Project Difficulty Calibrations for High and Low Ability Groups

	High Ability Group				Low Ability Group					
Slide	Difficulty	SE	Infit MS	Outfit MS	Difficulty	SE	Infit MS	Outfit MS	Calibration Difference	Difference Z-score
1	-.07	.10	0.7	0.7	-.15	.08	0.6	0.7	.08	.60
2	.73	.08	0.8	0.9	.56	.07	0.7	0.9	.17	1.29
3	.03	.10	0.7	0.9	.19	.07	0.8	0.9	-.16	-1.23
4	-.50	.12	1.2	1.0	-.42	.08	1.0	0.8	-.08	-.62
5	-.15	.10	0.9	0.8	-.17	.08	1.0	0.8	.02	.14
6	-.13	.10	0.6	0.5	-.21	.08	0.6	0.7	.08	.60
7	-.08	.10	0.8	1.2	-.08	.07	0.9	1.3	.00	-.01
8	-.26	.11	0.7	0.7	-.24	.08	0.6	0.6	-.02	-.16
9	.02	.09	0.9	0.8	.17	.07	0.9	0.9	-.15	-1.16
10	.39	.08	0.9	0.7	.37	.07	1.0	0.8	.02	.14
11	-.08	.10	0.7	0.9	-.20	.08	0.7	0.9	.12	.91
12	.05	.09	0.8	0.9	.00	.07	0.8	0.8	.05	.37
13	-.10	.10	1.0	0.7	-.23	.08	0.9	0.7	.13	.98
14	-.04	.10	0.9	1.2	.28	.07	1.1	1.3	-.32	-2.45[1]
15	.21	.09	1.1	1.4	.13	.07	1.2	1.4	.08	.60
								Mean =	-.004	

[1] While slides calibrated slightly different across examinee groups, the difference in Slide 14's calibrations was significantly greater than that of the other slides.

Figure 1. Plot of Judge Severity Calibrations For High and Low Ability Examinees

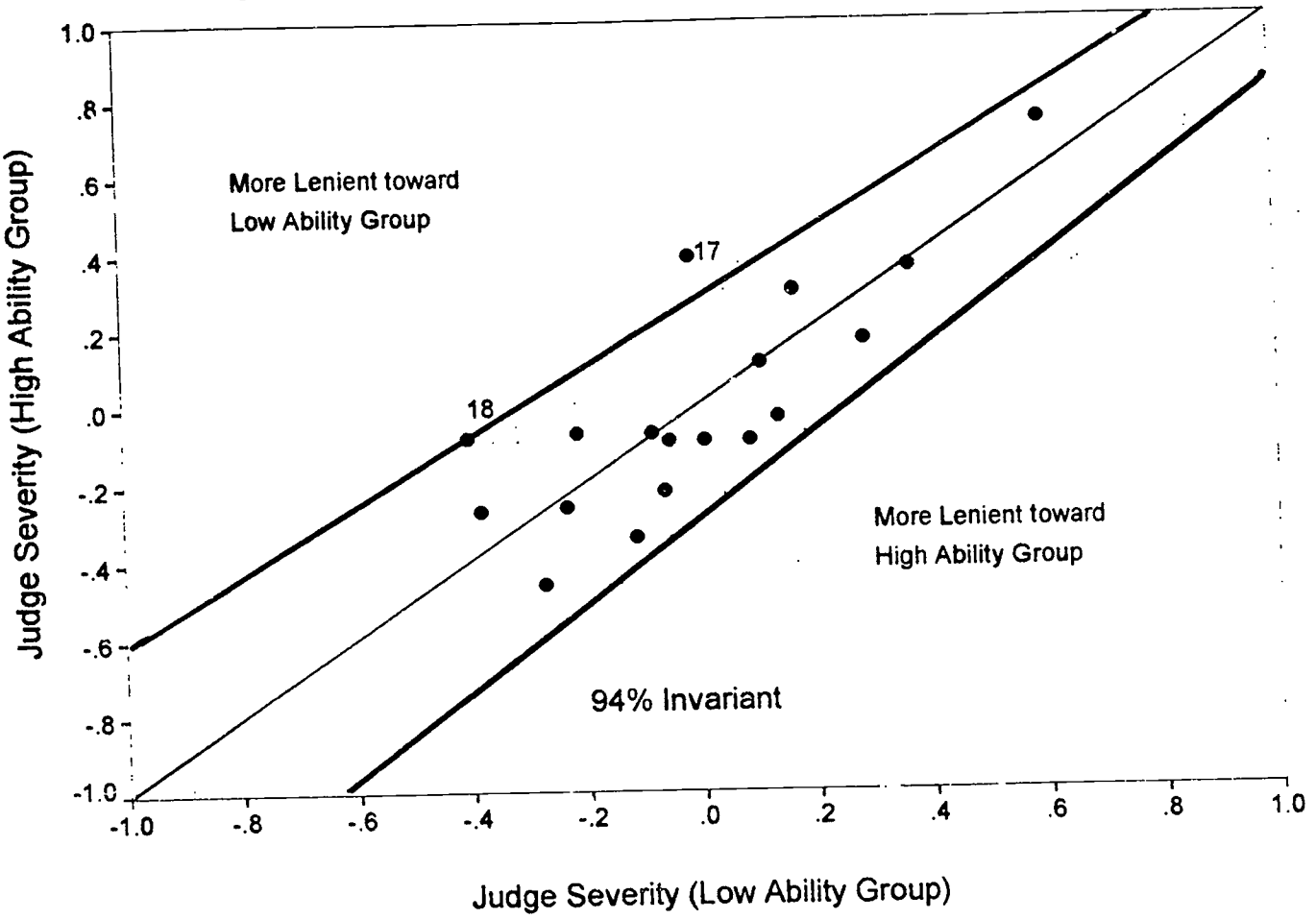

Figure 2. Plot of Slide Difficulty Calibrations For High and Low Ability Examinees