DOCUMENT RESUME

ED 408 307 TM 026 516

AUTHOR	Ingebo, George S.
TITLE	Examining Item Response Theory: Consistency of Rasch Calibration in Basic Skills Item Banks.
PUB DATE	[93]
NOTE	24p.
PUB TYPE	Reports - Evaluative (142)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Achievement Tests; Difficulty Level; Item Banks; *Item Response Theory; *Measurement Techniques; Probability; *Reliability; *Test Items
IDENTIFIERS	*Calibration; Oregon; Portland Basic Skills Achievement Levels Testing; *Rasch Model

ABSTRACT

Item response theory (IRT) is based on the assumption that a direct relationship exists between an examinee's total performance on a set of items and the difficulty of each item on the test. The Rasch model represents this relationship mathematically on an equal interval scale. This paper argues that IRT, under the required conditions, provides measurement comparable to that used for experimentation in the natural sciences. IRT is a unification theory in that it brings two variables together in a mathematical relationship, and it is a probabilistic model because the nature of the data allows prediction of probability of success on any calibrated item, given the student's achievement in Rasch units. The Rasch item calibrating model is tested through demonstrations of experimental methods and results. The most important use of the Rasch model is to produce an extended equal interval scale on which basic skills items are calibrated for an entire elementary school continuum. The main reason for creating a Rasch calibrated item bank is to make possible the construction of different levels of tests, all yielding comparable results on the same scale. Studies of the Portland (Oregon) Levels Tests supported three hypotheses about Rasch model calibration: (1) that item bank calibrations do, with a few exceptions, remain stable across grade level groups performing on the same items; (2) that item calibrations remain consistent from year to year and from grade to grade given appropriate testing and calibrating procedures; and (3) that for the same items, Rasch-calibrated by the fixed parameter model, administered to different groups of students, all calibrations for each item will be the same. These results support the use of Rasch model calibration when proper precautions are taken. (Contains eight figures.) (SLD)

EXAMINING ITEM RESPONSE THEORY:
Consistency of Rasch Calibration in Basic Skills Item Banks

George S. Ingebo

EXAMINING ITEM RESPONSE THEORY:
Consistency of Rasch Calibrations in Basic Skills Item Banks

Item Response Theory is based on the assumption that a direct relationship exists between an examinee's total performance on a set of items and the difficulty of each item in that test. The Rasch model mathematically represents this relationship on an equal interval scale.

Item Response Theory (IRT) is a **unification** theory because it brings together these two variables in a mathematical relationship. The Rasch is a **probabilistic** model because the nature of the data allows prediction of probability of success on any calibrated item, given the student's achievement level in Rasch units.

The questions answered in this study concern the equal interval properties of the scale produced when the mathematical model includes only the two basic elements of the theoretical relationship.

This study shows that this measurement theory — under the required conditions — provides measurement comparable to that used for experimentation in the natural sciences.

Characteristics that physicists look for in a mathematical model designed to unify two phenomena such as test performance and success rate on individual items are:

(1) the mathematical formula must use all the data generated in the interaction of the two variables;

(2) no other parameters are added in examining interaction of the two basic variables; and

(3) values can be computed in either direction across the equation.

The Rasch model meets each of these criteria:

(1) it uses all information generated in producing the relationship stated by the Item Response Theory;

(2) it contains no parameters external to the theory; and

(3) it has symmetry because either student performance **or** item difficulty can be calculated from the equation.

Rasch calibrated items, linked to a single continuous scale in a basic skills subject, can provide reliable measurement of individual and group growth on a curriculum continuum. It is the purpose of this paper to test the Rasch item calibrating model through demonstrations of experimental method and results.

Linking field tests to develop a continuous scale

Investigating the equal interval aspect of IRT requires constructing an extended scale for measuring a learning continuum, and demonstrating its continuity.

In developing the scale, it is important to use data derived in a way that is consistent with the proposed use of the scale. With this in mind, the developers of the PPS/NWEA Rasch scale (RIT) organized the administration and linking of a large number of field tests (120 or more). After applying traditional item analysis procedures to identify unsatisfactory items, the remaining items were Rasch calibrated within each test. Several Northwest school districts and the San Jose, California School District participated in using these field tests. These field tests were then linked to a single scale. This was done through the known calibration values calculated for each of the items. Factors used to adjust calibrations from each test to all items in other trial tests ("linking constants") were calculated by reconciling the links among various tests with several confirming linkages.

The field test data came from thousands of students, with their varied personal reactions to the items, and from hundreds of field tests given under varied conditions of test administration. In forming the item bank, items showing errant performance within a test or contributing to linking irregularities were identified and excluded.

There are three distinct operations needed to build a continuous curriculum scale after traditional item analysis information has been used to screen out bad items:

(1) determine the usefulness of an item through examination of the mean square fit statistic, the item characteristic curve, and examination of the item's relationship to the curriculum outcome it purports to represent;

(2) examine plots of items in links between tests; and

(3) make adjustments to reconcile the links and multiple confirming link values in order to establish a calibrated bank of items.

Calibration consistency as a test of theory

The most important use of the Rasch model in education is to produce an extended equal interval scale on which basic skills items are calibrated for an entire elementary school continuum. The existence of the scale then makes it possible to compare results from different levels of tests in the same subject.

The main reason for creating a Rasch calibrated item bank is to make possible the construction of different levels of tests, all yielding comparable results on the same scale. With such tests a district can test each student at his/her level of functioning yearly or twice each year and provide all district personnel with both cross section and growth information. Such a program has existed in the Portland, Oregon Public Schools for over a dozen years, using tests created from item banks constructed with the procedures briefly described above.

All items in the Portland Levels Tests were calibrated and linked to a common scale covering the curriculum in mathematics, reading, and language usage in grades three through eight. All items used were subjected to continuous review over a period of years. Tests were administered to tens of thousands of student twice yearly in grades three through eight. The database created in this process provided the data for this study.

A demonstration of consistent item calibrations across the curriculum scale provides support both for Item Response Theory and the accuracy of measures secured through tests linked to the same scale. Achieving this consistency requires an equal interval scale for linking and a relationship with student achievement level that is independent of group or grade level.

As stated earlier, tests of increasing difficulty are administered at the functioning achievement level of each student in the Portland testing program. This helps assure all students receive tests that are neither too easy nor too difficult, which increases the accuracy of measures secured.

Uniformity in length, range of difficulty, and individual appropriateness make these tests comparable in terms of logit length, as discussed by Dr. Benjamin Wright in *Rasch Measurement*, Summer, 1990.

HYPOTHESIS 1.

Items calibrations will remain constant across groups of students in grades three through eight who were assigned the same test because previous test measures indicated a similar functioning achievement level.

Purpose: The purpose of this part of the study was to determine the extent of calibration consistency along a specific curriculum scale. The technology for examining calibration stability was developed by Dr. Fred Forster to monitor item calibrations in tests constructed in the Portland Public Schools from Rasch-calibrated item banks over a period of years.

Procedure: Students from grades three through eight were administered the same test from the levels test series. This test was selected on the basis of the students' most recent achievement test results. This insured that students were working with the test most appropriate for them to take. Thus, fast learning third or fourth grade students took the same test as lower achieving seventh or eighth grade students.

All students taking the same test were identified and grouped by grade levels. The Rasch model was used to analyze each group separately. Six new values were computed for each item as it was calibrated in the six different student groups.

Since the Rasch model sets the center of the calibrations at zero for each test as it is analyzed, the six calibrations for each item can be compared directly across grade groupings.

Findings: Theoretically, calibrations from each of the six groups should remain the same.

Figures 1 through 3 show the distributions of item calibrations for all six grade groups plotted for each of the three tests. The similarity of calibration patterns for the different grade groups for each year is shown in the figures.

Figure 4 shows percents of deviations for a large sample of 3,300 differences. Since the calibration of an item in one grade is compared with its calibrations in all other grades, a widely differing calibration in one grade adds several large differences to the distribution.

Conclusions: These findings support hypothesis (1) in that the item bank calibrations did, with few exceptions, remain stable across grade level groups performing on the same items.

HYPOTHESIS 2.

Item calibrations will be the same in the same tests when administered in consecutive years.

Purpose: One of the requirements of a levels testing system derived from a Rasch calibrated item bank is to provide longitudinal information for students over their elementary school experience.

The purpose of this analysis is to examine the calibration stability of items from year to year, thus demonstrating the integrity of the item bank over time and the consistency of Item Response Theory.

Procedure: Data from the Portland School district testing program were used to examine the degree of consistency in Rasch item bank calibrations as well as the equal interval nature of scales based on the Rasch model.

Test MO4RD fall administrations were used from the years 1985, 1986, and 1987. Calibrations for each item were compared across years.

Findings: Figure 5 shows the distributions of two calibration differences between items: 1985 to 1986 and 1986 to 1987.

Figure 6 is a distribution of differences in calibrations for three tests; LO3RD, LO4RD, and LO5RD.

The distribution of calibration differences aggregated for the three years shows 98% of the differences between calibrations for two different years are 3 RITs or less. The uncertainty inherent in the data, student responses to test items, and teachers individualized implementation of instruction will produce variation in calibrations, but within acceptable limits.

Calibration drift between the tests examined is well within practical limits for producing test measures that can be used for *reporting **both** growth and rate of growth* in basic skills for groups of students.

Conclusions: 1. According to these data, item calibrations do remain consistent from grade to grade and from year to year, given appropriate testing and Rasch calibrating procedures.

2. Confidence in year to year consistency of item calibrations in tests constructed from Rasch calibrated item banks make these tests suitable for testing Item Response Theory.

3. Item Response Theory is generally supported by the data generated with these tests developed from the Rasch calibrated item banks.

HYPOTHESIS 3.

When the same items are administered to different groups of students and Rasch calibrated by the fixed parameter model (see explanation below), all calibrations for each item will be the same.

Purpose: This study was designed to determine the consistency with which the fixed parameter model calibrates test items as a further test of the Item Response Theory.

The fixed parameter formula for linking field test items to an existing Rasch scale was first programmed by Dr. Fred Forster to increase the number of items in the banks used by the Portland Public School's Research and Evaluation department. It involves substituting Rasch scaled student measures in place of raw scores in the Rasch model. The values assigned to field test items calibrated in this way fall into place with those of existing bank items. They do not require special linking procedures because they are mathematically linked to a continuous scale already created by the more difficult and resource consuming methods required to establish the original scale.

The most important use of the Rasch model in education is to establish continuous measurement scales. Practical applications of linking have resulted in calibrated item banks such as the Northwest Evaluation Association item banks in reading, mathematics, and language usage. These banks have several thousand items related to each other on a continuous curriculum scale.

Securing fixed parameter model calibrations simplifies the process of expanding and tailoring such Rasch calibrated item banks. As indicated, statistically independent calibrations of field test items are computed by substituting achievement measures from an established Rasch scaled testing program for estimates of student's scores as required in the Rasch calibration formula.

Items with acceptable item characteristic curves can be taken directly into the item bank, providing the following conditions are met: (1) the item bank used to secure the fixed parameter achievement data was already Rasch calibrated on a single scale over its entire range; (2) the achievement measures come from tests drawn from this item bank and assigned at the students' levels of achievement; (3) the groups taking the tests have been exposed to the curriculum content of the items; and (4) teachers and students know that the test results will be reported to them.

Procedure:

In each basic skill area (reading, mathematics, and language usage), nine groups of 300 to 400 third grade students were given the same fifteen item field test.

Teachers administered the field tests in the spring. The regularly scheduled district wide levels test program followed a few days after field testing. These tests provided the achievement measures used in the fixed parameter equations.

Tests were assigned according to each student's own level of achievement. Student's performances in the district's regularly scheduled fall testing program indicated which level of test should be taken in the spring. The assignment was subject to adjustment by the teacher if indicated by the student's work over the school year.

The field tests consisted of fifteen new items. Their range was 30 RITs (3 logits). The new items had been considered by the curriculum committee, the sex and race bias committee, and the editing committee, but had not been tried out in any test.

The fifteen field test items for each of the three basic skills were calibrated independently for nine groups of students to examine the consistency with which the fixed parameter model calibrates items administered to different groups of students.

Third grade test data has more variation than other levels, making the test of item response theory most stringent in that respect. When the high and low extreme scores are identified in longitudinal consistency data, test LO4RD (Table 13 - 4) for example, grade three accounts for more of the extreme calibrations than any other grade.

Findings: Three replications of the study are reported separately for reading, mathematics, and language usage. Each is discussed following an initial aggregation of data. All of the 1620 differences in calibrations are included in frequency distributions in Figure 7. The same information is presented in actual differences, percentages, and cumulative percents in (table 8).

Conclusions: 1. The Item Response Theory is supported by the consistency of calibrations produced by the fixed parameter model across nine independent samples of students in each of three basic skills curricula. Different students in different schools were tested in each subject to secure the test results used in this study.

2. The fixed parameter model provides calibrations for field test items that permit adding such items to established item banks.

Research requirements

Rigorous attention to a variety of matters must be attended to if the quality of the data collected and analyzed is to be scientifically acceptable. Some of these matters include:

Instrument Validity

Item Response Theory states that achievement measures have a consistent relationship to individual item difficulty. To test IRT requires that test data are sufficiently accurate and appropriate to warrant interpretation of findings relevant to the theory's premises.

Testing instruments that meet these requirements must be made up of items that are neither too hard nor too easy for the student, and are distributed around the student's achievement level in that curriculum. Each test item must represent a specific outcome of curriculum and instruction. The items must include content to which the students have been exposed.

When such a test's items are balanced with respect to curriculum goals, are graduated in difficulty, and are assigned to students on the bases of their achievement level, *instrument validity* is satisfactory for experimental testing to affirm the premises of the theory.

Student Performance Reliability

Even though a test instrument is shown to have a high degree of instrument validity, erratic (unreliable) performance of a student who does not cooperate can destroy measurement accuracy. Individual reliability, unlike instrument validity, reflects personal manifestations of cognitive and emotional idiosyncracies impinging on student performance. Without satisfactory reliability of test performances, instrument validity is impossible.

Defining a Continuous Curriculum Variable

Curriculum content usually varies too much from school to school in subjects other than the basic skills to provide useful data for the present study. For this reason it is necessary to use only the basic skills curriculum scales that are consistent in elementary schooling.

The primary responsibility of the public school is to teach the basic conventions of communication and problem solving, *i.e.,* language and mathematics. Since these societal conventions are sets of man made rules, instruction in their use is implemented sequentially and consistently enough within a school district to make it possible to define a curriculum continuum. Each of the basic skill curriculums, mathematics, language usage, and reading are suitable variables for the development of scales. Research leading to acceptance of a Rasch based testing program is presented and documented in the National Association of Test Directors Symposium, 1987.

Origin of Data

Data are produced as students mark answer sheets. The marking is not judged —only counted— by the scoring machine. But the student's marking of the test sheet has to be viewed through a psychological filter. For simplicity, we refer to the "mental set" of the student before and while marking an answer sheet.

Unlike people, a seven foot long rod has no reactions, such as boredom, if the carpenter measures it with a six foot tape. Nor is it anxious if measured with a twenty foot tape. This is not true of a high achieving student who realizes a traditional standardized test is too easy. Nor is it the case for a low achieving student who knows that most items in a test are too puzzling to be a measure of what he or she knows.

When a student's raw score is less than a third of the number of items in a test, results are questionable indicators of what the student knows. Scores based on a large percentage of items easy for a student or scores approaching all items correct can hardly show excellence of performance.

A mental set exists prior to testing and is modified with every item encountered. Emotional carryover from past experiences with tests that have been personal defeats or that lacked challenge may produce unfavorable mental sets that influence performance. And as a low student gets into a test that is too difficult, the emotional impact can become serious.

The one thing we must do in administering a testing program to test a theory of measurement is to be sure that students are assigned tests appropriate for their levels of achievement and instructional experiences each time they are tested.

Scale development

Scale development with the Rasch is not a matter of finding a population central tendency for making comparisons. The population-independent nature of a Rasch scale comes from accounting for each individual's level of performance, not from ignoring it. Freedom from comparison with a population is the key to quality of information usable for experimentation. **Identifying and purging erratic data** is a necessary dimension of that quality. Erratic student records, faulty test administrations, and misleading links are unacceptable.

Following this principle requires extensive computer manipulation of information, but the computer can't do it all. Professional judgment must play a part in interpreting information produced by computer manipulation, especially in judging Item curves and linking patterns.

Statistical and Professional Estimates

It has been noted that only usable items are acceptable for processing. But not all field test items that appear acceptable for a Rasch analysis are useable in a test. Deciding to reject certain items and to keep others is helped by numbers such as student's scores, percent of students at different achievement levels answering the item correctly, and numbers of students responding to the item. Help is also provided by chi-square fit statistics and point-biserial correlations for the item.

These statistics are summary estimates and helpful to some extent, but the weight of evidence rests with observation of the plotted performance of students on each item (item characteristic curve). The plots show what happens when students meet the test.

There are three types of qualitative information needed to build a continuous curriculum scale after traditional item analysis has been used to identify undesirable items:

(1) determining the usefulness of an item through examination of the mean square statistic, the item characteristic curve, and each item's relationship of the curriculum outcome it purports to represent;

(2) examining plots of items in links between tests; and

(3) making adjustments to reconcile the links and multiple confirming link values in

order to establish a calibrated bank of items.

Recommendations for the Rasch model version of Item Response Theory

1. Customize Rasch calibrated basic skills item banks to District goals and construct tests with confidence that a continuous curriculum scale is consistent grades 3 through 8.

2. Utilize the year to year consistency of the curriculum scale to exploit the growth information from fall to spring and other desired time intervals.

3. Utilize the capability of the fixed parameter model to substitute an independent measure of ability in the item calibration formula. Given measures from calibrated tests, item calibrations can be secured much more efficiently than using traditional field tests and linking methods.

Figure 1. Calibrations for Items by Grade

Reading, Spring 1987

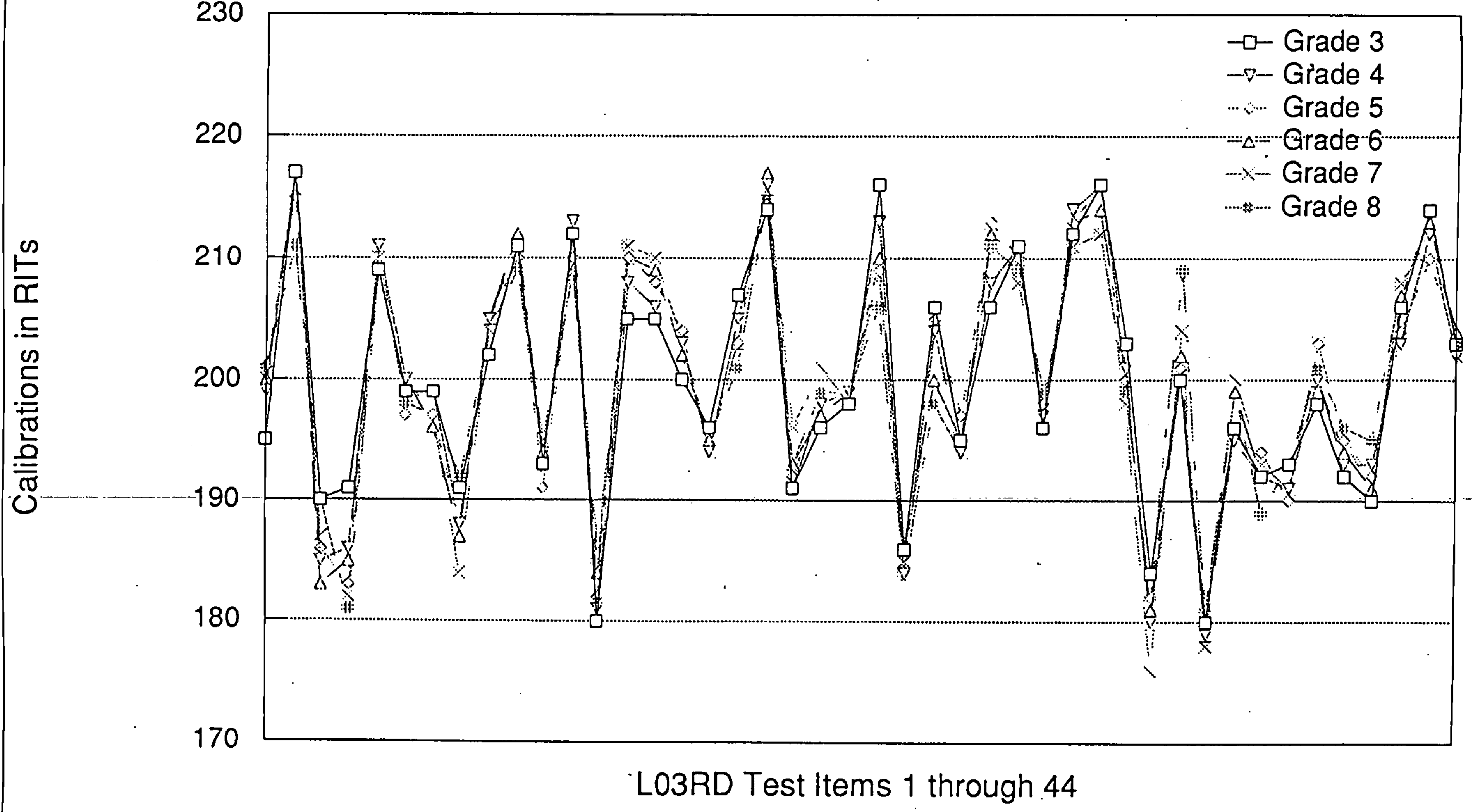


Figure 2. Calibrations for Items by Grade

Reading, Spring 1987

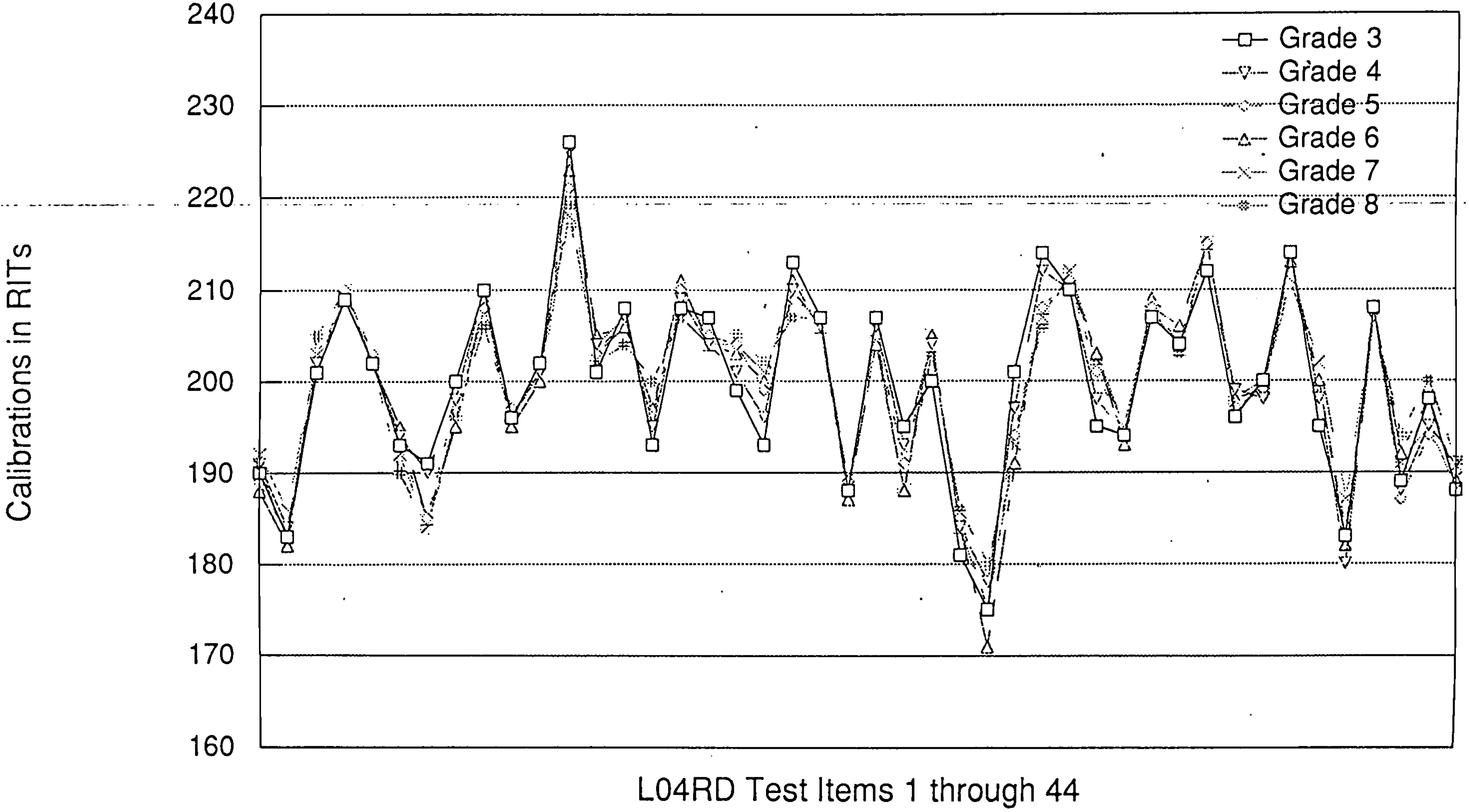


Figure 5. Differences in Year to Year Calibrations

For the Same Items in Reading Test M04RD

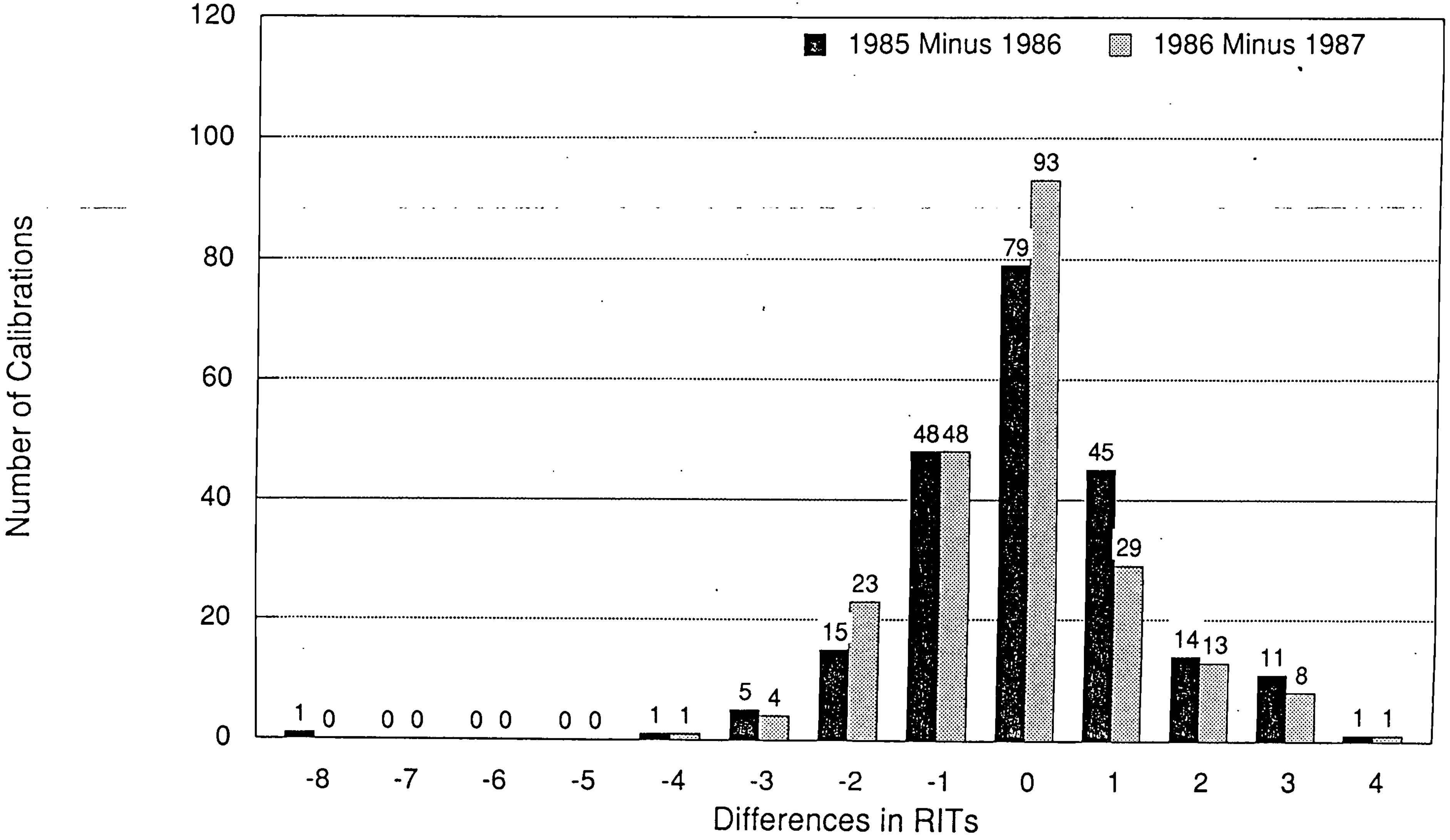


Figure 6. Distribution of Item Calibrations

Differences from 1986 to 1987 in the Same Tests

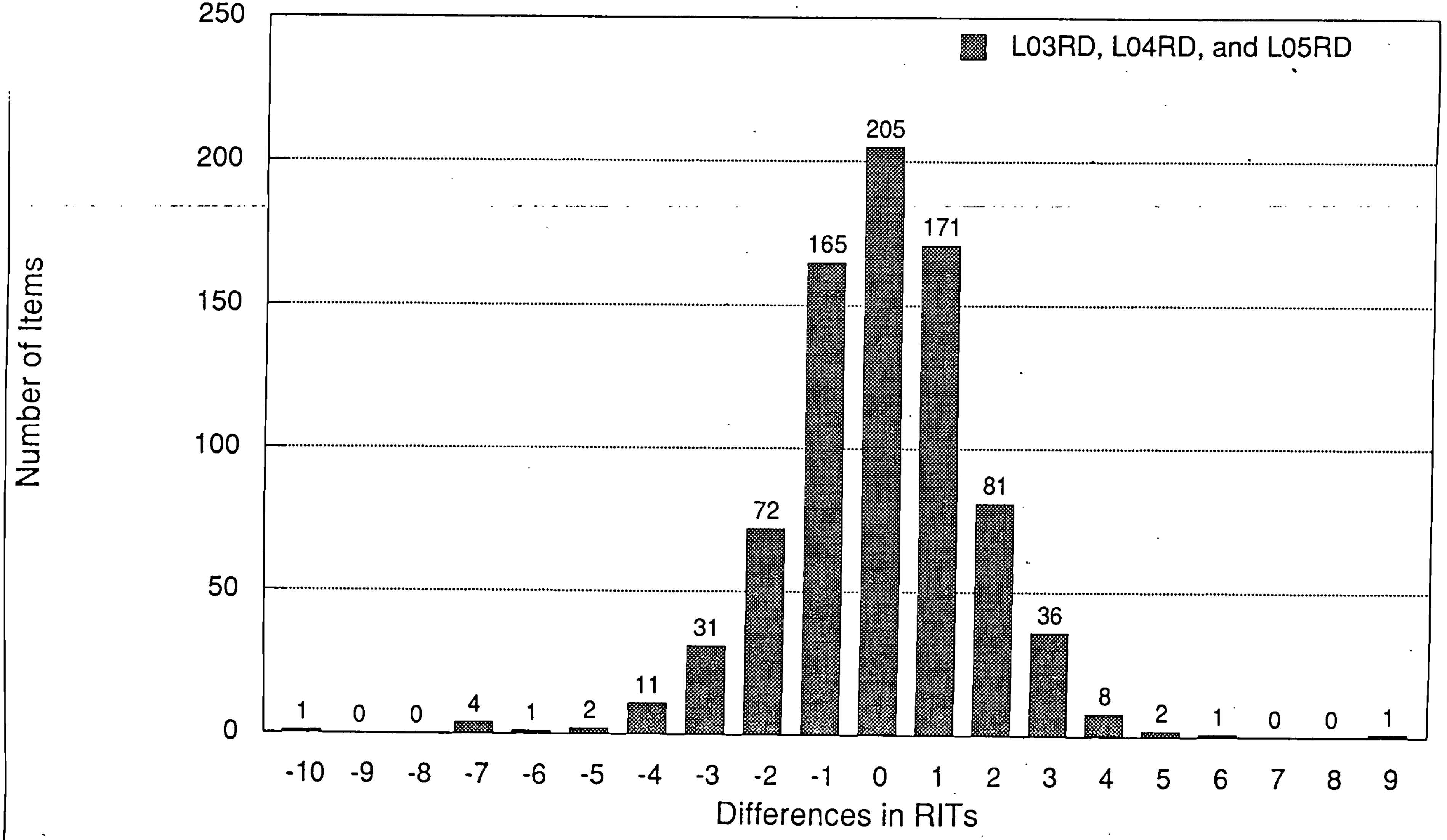


Figure 7. Calibration Differences Between the Same

Field Test Items in Reading, Math, and Language Usage

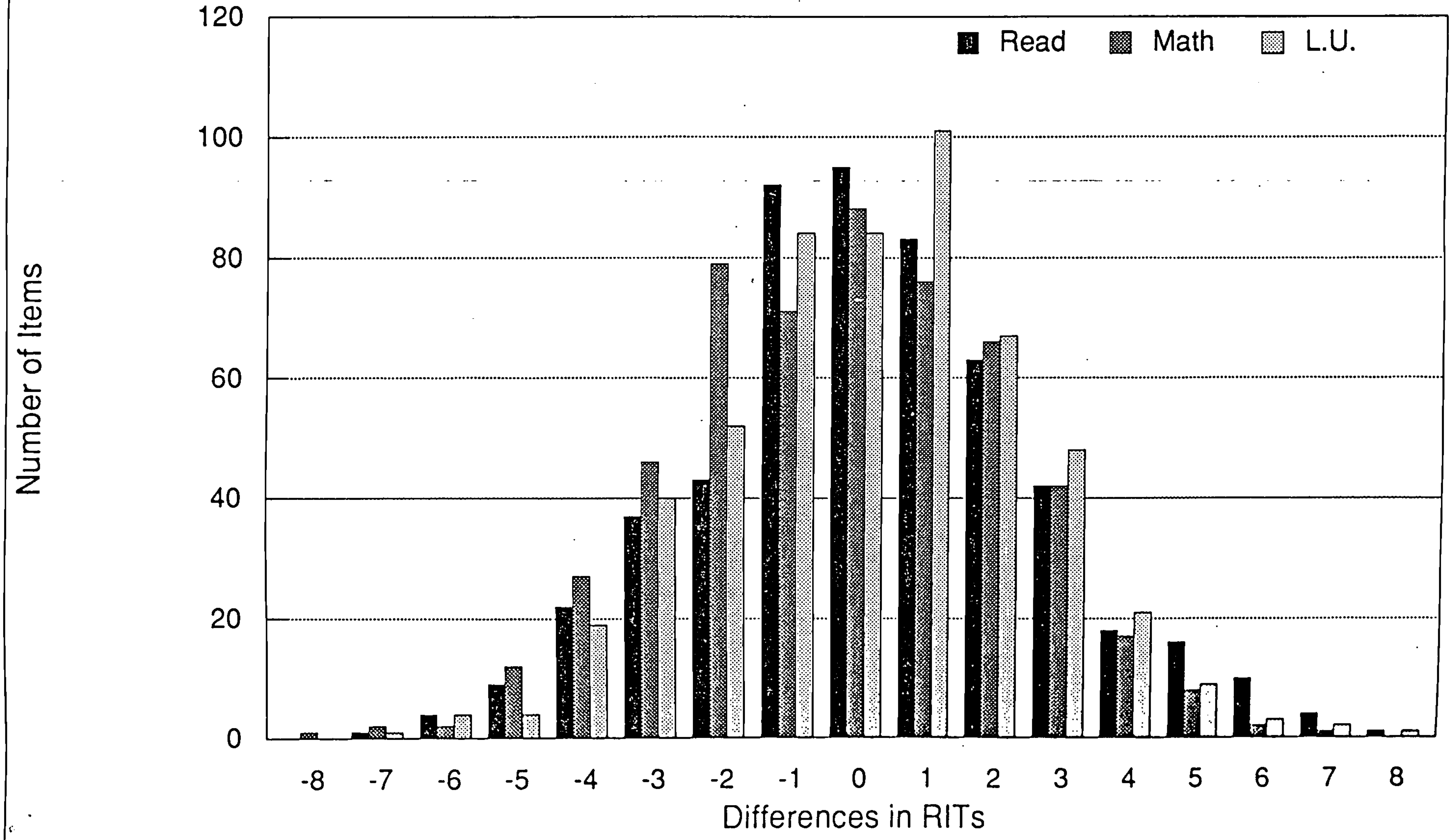
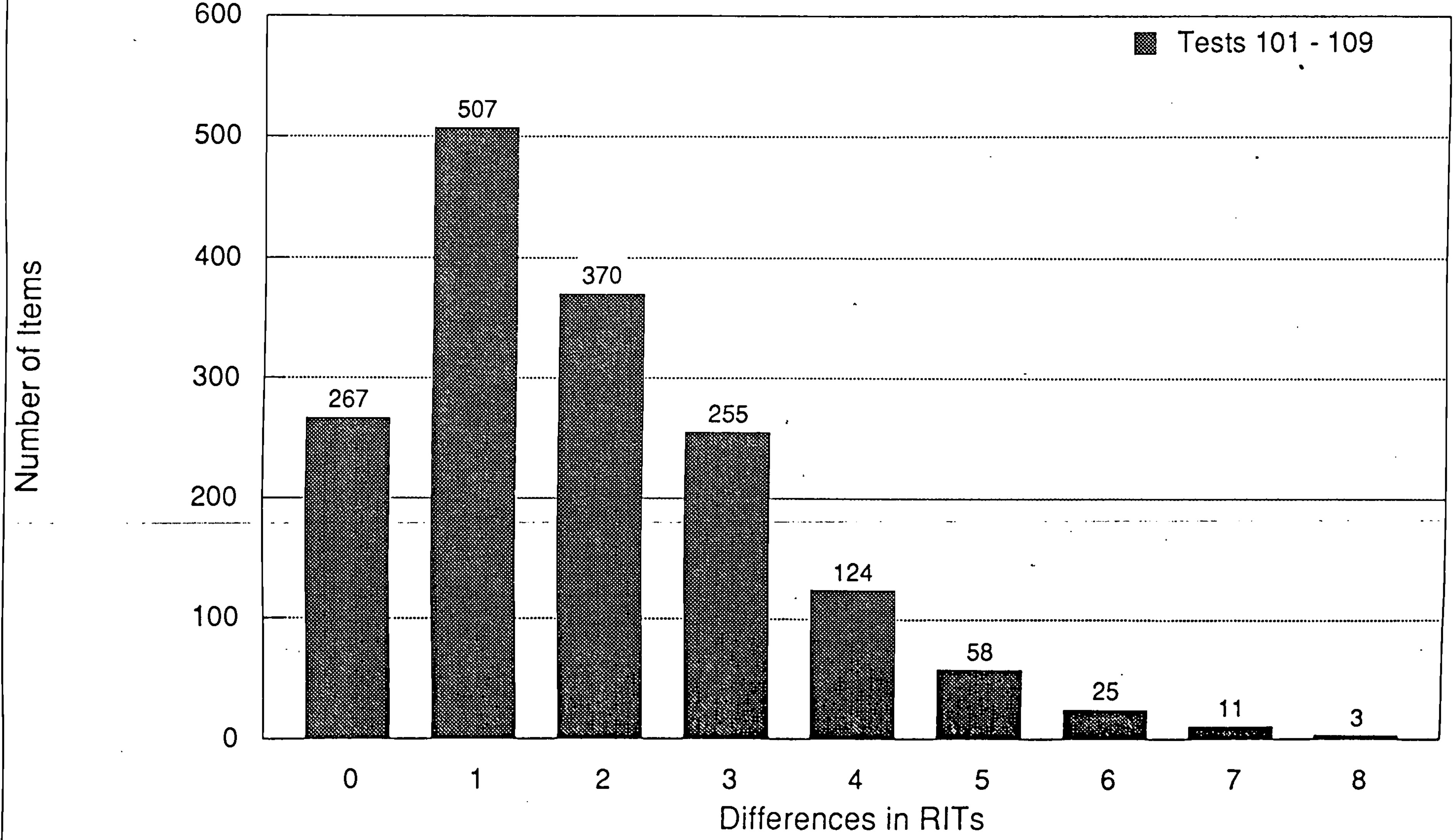


Figure 8. Calibration Differences Between Items

For All Field Tests in Reading, Mathematics, and Language Usage



TM026516

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title:
Consistency of Rasch Calibrations, Grades 3 to 9

Author(s): GEORGE INGEBO

Corporate Source:

Publication Date: 1/15/1993

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic/optical media, and sold through the ERIC Document Reproduction Service (EDRS) or other ERIC vendors. Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following two options and sign at the bottom of the page.

□ Check here
For Level 1 Release:
Permitting reproduction in microfiche (4" x 6" film) or other ERIC archival media (e.g., electronic or optical) *and* paper copy.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Level 1

The sample sticker shown below will be affixed to all Level 2 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN OTHER THAN PAPER COPY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Level 2

□ Check here
For Level 2 Release:
Permitting reproduction in microfiche (4" x 6" film) or other ERIC archival media (e.g., electronic or optical), but *not* in paper copy.

Documents will be processed as indicated provided reproduction quality permits. If permission to reproduce is granted, but neither box is checked, documents will be processed at Level 1.

"I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic/optical media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries."

Sign here→ please

Signature:	Printed Name/Position/Title: GEORGE S INGEBO, PhD. retired	
Organization/Address: Mr. George S. Ingebo 3708 NE 136th Pl. Portland, OR 97230	Telephone:	FAX:
	E-Mail Address:	Date:

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:

Address:

Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:

Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
1100 West Street, 2d Floor
Laurel, Maryland 20707-3598

Telephone: 301-497-4080
Toll Free: 800-799-3742
FAX: 301-953-0263
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com