DOCUMENT RESUME

ED 263 144 TM 850 618

AUTHOR	Doss, David; Ligon, Glynn
TITLE	Empty Bubbles: What Test Form Did They Take?
INSTITUTION	Austin Independent School District, Tex. Office of Research and Evaluation.
REPORT NO	AISD-ORE-84-39
PUB DATE	Apr 85
NOTE	9p.; Paper presented at the Annual Meeting of the American Educational Research Association (69th, Chicago, IL, March 31-April 4, 1985).
PUB TYPE	Speeches/Conference Papers (150) -- Reports - Research/Technical (143)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Educational Testing; *Error of Measurement; Latent Trait Theory; *Measurement Techniques; Secondary Education; *Test Format; *Testing Problems; Test Results
IDENTIFIERS	*Austin Independent School District TX; Person Fit Measures; Rasch Model; *Rasch Scaled Scores; Sequential Tests of Educational Progress

ABSTRACT

Upon learning that a form of the Sequential Tests of Educational Progress was incorrectly distributed to an unidentified number of high school students along with an answer sheet pregridded with an alternate test form, the Austin Independent School District performed the following research analyses: (1) scored the tests using the key for each form, calculating a daily total raw score; (2) Rasch-calibrated the tests using the key for Form B, the form the students were supposed to have taken; (3) obtained an average person-fit statistic for each day; (4) created a file which included, for both days of testing, the average person-fit score and the total raw score for both Forms A and B, and also the person-fit scores for each subset; (5) sorted the file by Day 1 average person-fit score and print; (6) sorted the file by Day 2 average person-fit score and print; and (7) examined the output. Six samples illustrate the case types found. Apparently 63 of 65 booklets were used with 78 students taking Form A at least one day. However, there is no empirical evidence to suggest what kind of error rate there is with this procedure. (PN)

-EMPTY BUBBLES: WHAT TEST FORM DID THEY TAKE?

David Doss and Glynn Ligon
Austin Independent School District
Austin, Texas

Paper Presented at the Annual Meeting of
the American Educational Research Association

Chicago, Illinois April, 1985

Publication #84.39

84.39

EMPTY BUBBLES: WHAT TEST FORM DID THEY TAKE?

David Doss and Glynn Ligon
Austin Independent School District

The "crisis" developed out of an almost casual remark by a high school counselor about a teacher's reporting that she "thought" some of her students had taken Form A of the STEP. That may appear to be an innocuous enough statement; however, all students at the school were supposed to have been given Form B, and we had no way of knowing which ones had taken Form A.

First a little background. When our district gave the Sequential Tests of Educational Progress (STEP) to our high school students, we gave half of our schools Form A and half Form B. The following year they would get the alternate form. We did this because there was only one level on the STEP, and we did not want students to take the same form all four years of high school. We also pregridded or preslugged our answer sheets with the required demographic information for each student enrolled. Part of the preslugging was the test form, so only those students who did not receive a preslugged answer sheet actually coded the test form on their answer sheet.

We confirmed that we had a problem when an examination of the test booklets returned by the school showed that 65 Form A booklets had been sent with the 1300 Form B booklets the school was supposed to receive. The situation was complicated by the fact that the testing was done on two days. Five tests were given on day 1 and three on day 2. The test booklets and answer sheets were collected separately at the end of the first day and redistributed the next. Therefore, a student could have taken the incorrect form either or both days.

Our immediate thought was to score the tests using both keys and give the students the higher score for each test. However, the students could use the STEP results in reading and mathematics to meet the district's minimum competency requirement. We thought it best to try a less sweeping approach. We wanted to identify those students who had taken the wrong form of the test and change their scores only. It was time to bring out the big guns, the Rasch person-fit statistic (Wright and Stone, 1979).

The Rasch Person Fit Statistic

We were both familiar with the Rasch person-fit statistic from attending an AERA presession on the Rasch latent trait model presented by Benjamin Wright and his colleagues, but we had had no opportunity to use it in our work. This situation seemed to be the perfect opportunity.

The reader is referred to Best Test Design, Wright and Stone's very clear and readable how-to book on measurement with the Rasch model for a full explanation of person fit. However, the logic of the statistic is described below.

The Rasch model holds that the probability of a student's answering an item correctly is a function of two parameters, the student's ability and the item's difficulty. When the student's ability matches the item difficulty, the probability of a correct response is one half. With increasing student ability the probability of a correct response to the item approaches one. With decreasing ability, the probability approaches zero. An incorrect response to an easy item by a student of high ability is inconsistent with the model as is a correct response to a difficult item by a low-ability student. The person-fit statistic is a measurement of the degree to which a person's pattern of correct and incorrect responses is consistent with the individual's raw score. The higher the fit score, the greater the discrepancy.

In our case, when students' tests were scored with the wrong key, they should have appeared to be answering at random, giving them low raw scores. Furthermore, they should have been just as likely to answer a difficult item correctly as an easy one. Therefore, they should have had high fit scores indicating a misfit with the Rasch model.

It should have been easy to detect which form high-scoring students took from a comparison of raw scores alone. The hope presented by using the Rasch fit statistic was that we could identify those low-scoring students who took Form A. Their raw scores might be similar on each form, but if they were not guessing to a great extent, their fit score should be noticeably high when computed using the Form B key.

The Procedure

The analyses were carried out using the SPSS package of statistical programs and program RASCH in the PRIME system of statistical programs (Veldman, 1978). Most analyses were conducted on the University of Texas Dual Cyber computer system.

The steps in the analyses were as follows:

1. Score the tests using the key for each form, calculating a daily total raw score.

2. Rasch calibrate the tests using the key for Form B, the form the students were supposed to have taken. Our initial notion was to Rasch calibrate the tests using both scoring keys. However, with further thought it did not seem to make sense to calibrate the tests with the Form A key because perhaps 95% of the students would be scoring randomly. The item difficulties would have had essentially random values. Had we not been working in a "crisis" situation, trying to get the test scores back to the campus as quickly as possible, we could have calibrated Form A using the schools which

gave it and then have written a special program to calculate the person-fit statistic using reasonable item calibrations and the students' responses. As it was we had to settle for calibrations using the Form B keys only.

3. Obtain an average person-fit statistic for each day.

4. Create a file with the following information:

   Student Name
   Identification Number
   Classroom in Which Tested
   Day 1 Average Person-Fit Score
   Day 2 Average Person-Fit Score
   Day 1 Total Raw Score--Form A
   Day 1 Total Raw Score--Form B
   Day 2 Total Raw Score--Form A
   Day 2 Total Raw Score--Form B
   Person-Fit Scores for Each Subtest

5. Sort the file by Day 1 average person-fit score and print.

6. Sort the file by Day 2 average person-fit score and print.

7. Examine the output. Students likely to have taken the wrong form should have person-fit scores in the top percentile ranges and Form A raw scores that are greater than Form B raw scores. Sorting by the person-fit scores should bring most of the students of interest to the top of the printout.

   The classroom variable was included because the way the Form A booklets had been found in the boxes from the school suggested that they would be concentrated in a few classrooms. Once likely students had been identified for the first day, the classroom information was expected to be helpful in identifying those getting the wrong form the second day and in verifying that the first day's conclusions were reasonable.

Results

Figure 1 provides the results for six sample students to illustrate the types of cases found. They are the scores of actual students and are presented as if sorted by day 1 average fit scores. The daily average fit scores had mean values of about 1.00 (1.007 and 1.022 for day 1 and day 2 respectively) and standard deviations of about .3 (.297 and .299).

The first student had the highest day 1 fit score. It really makes no difference which form this student was given. He obviously guessed at most of the items. He answered correctly only two to four items on each subtest attempted regardless of the scoring key used. For all practical purposes, this student should be considered to be not tested.

Figure 1: SAMPLE OUTPUT FROM ANALYSES

STUDENT	AVERAGE FIT		DAY 1 RAW SCORE		DAY 2 RAW SCORE		FORM B FIT SCORES BY SUBTEST							
	DAY1	DAY2	FORM A	FORM B	FORM A	FORM B	1	2	3	4	5	6	7	8
1	4.11	1.83	15	9	4	4	2.77	8.50	0.00	2.44	2.76	1.83	0.00	0.00
2	3.10	1.11	46	130	38	77	.72	12.42	.99	.76	.70	.84	1.22	1.27
3	1.80	1.04	94	50	36	54	1.51	2.02	1.31	2.32	1.86	1.16	.99	.96
4	1.79	1.32	156	68	104	54	2.10	2.31	1.26	1.75	1.51	1.41	1.25	1.29
5	1.09	1.04	54	79	30	45	1.39	1.06	.69	.85	1.44	1.05	1.22	.84
6	1.09	1.10	40	68	37	35	1.07	1.46	1.13	.90	.90	.98	1.08	1.25

Subtests and Days Given:

1 = Math Concepts, Day 1	5 = Capitalization and Punctuation, Day 1
2 = Reading 1, Day 1	6 = Science 1, Day 2
3 = Reading 2, Day 1	7 = Science 2, Day 2
4 = Spelling, Day 1	8 = Math Computation, Day 2

The second student might appear from the fit score to have taken Form A on day 1 and Form B on day 2. However, an examination of her daily total raw scores indicates that she scored much higher when the Form B key was used. The chances of obtaining a score of 130 on the wrong form by chance are very small. Her high fit score comes from test 2, one of the reading tests. Her performance on that test was extremely inconsistent with the Rasch model. Our conclusion was that she took Form B, the correct form both days.

The third student would appear from his average fit scores to have taken Form A on the first day and Form B on the second. Examination of the total raw scores and the fit scores by subtest supports that conclusion.

The fourth student is the only example student to take Form A both days.

The fifth student appeared to take Form B both days, although two of the subtests given the first day had high fit scores indicating some problem with his performance.

The sixth student is an example of a low-scoring student who apparently took Form B on both days, yet has a higher score on Form A for the second day. The results seem to indicate that she attempted the questions and did not guess at random. Apparently she did not know many answers.

Figure 2 summarizes the results. Apparently 63 of the 65 booklets were used. Altogether, 78 students took Form A at least one day.

Second Period Teacher	Class Size	Form Taken Day 1/Day 2			Number of Form A Booklets Used
		A/A	A/B	B/A	
006	17	1	3	1	4
507	30	26	1	2	28
508	24	0	2	3	3
605	24	2	4	4	6
606	24	0	7	7	7
710	15	15	0	0	15
Total	--	44	17	17	63

Figure 2: NUMBER OF STUDENTS TAKING FORM A BY CLASSROOM

In retrospect, what might we do differently if we faced the same problem again? Probably not much. We try to provide test scores to campuses with weekend turnaround. They get the answer sheets to us on Friday, we get results to them on Monday. In this case we were not able to get the corrected scores for these 78 students to the school in that short period of time; however, our commitment to a quick response and other pressing work still caused us to omit activities that might have been interesting and perhaps helpful. We probably would not act differently today. The close match between the results and the distribution of booklets in the boxes from the school gives us confidence in our results. Furthermore, the school did not question any of the changes made.

In examining the printouts in preparing this paper, one additional idea did suggest itself. For the most part, the major factor in the decision was the difference between the daily total raw scores. The primary value of the fit statistic was the confidence it gave us about the correctness of our decisions. Therefore, it would probably have been better to have sorted the file by the difference between the Form A and Form B daily raw scores rather than by the average fit scores. When that is done, it appears that a greater percentage of those with the wrong form are listed at the very top of the printout.

As mentioned above we are reasonably confident about the reasonableness of our decisions. However, we have no empirical evidence to suggest what kind of error rate there is with this procedure. It would be a simple task to randomly select a sample of students tested with one form and add them to the students in another school tested with the alternate form. Naive subjects could identify the misfits following our procedures. However, time does not permit that at this time.

If such a procedure did show that there was a high level of accuracy in the judgments made, then a program could be written incorporating known item calibrations and the decision rules so that misfitting records could be automatically identified.

References

Veldman, D. The PRIME System: Computer Programs for Statistical Analyses. Austin: Research and Development Center for Teacher Education, The University of Texas, 1978.

Wright, B. D. and Stone, M. H. Best Test Design, Chicago: Mesa Press, 1979.