DOCUMENT RESUME

ED 243 922 TM 840 220

AUTHOR Frechtling, Joy A.; Schenet, Margot A.
TITLE A Funny Thing Happened on the Way to the Printer: The Saga of Developing a Customized Test.
PUB DATE Apr 84
NOTE 6p.; Paper presented at the Annual Meeting of the American Educational Research Association (68th, New Orleans, LA, April 23-27, 1984).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Descriptive (141)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Achievement Tests; *Criterion Referenced Tests; Program Evaluation; Publications; Scoring; *Test Construction; *Testing Problems
IDENTIFIERS Montgomery County Public Schools MD; Test Publishers

ABSTRACT

A description of the difficulties encountered in constructing a criterion-referenced test to assess end-of-year skills as part of a program evaluation is presented. The problems encountered in preparing the customized test are described in humorous detail. The first problem involved the listening test obtained from a publisher without administration instruction. The test booklets were not prepared as ordered, and item formats had to be changed. Items were incorrectly scored when returned to the publisher, which complicated an item analysis. The norm data reports issued by the publisher were also inaccurate. The final revised instrument provided useful information about the listening comprehension skills of students who were nonreaders. Despite the difficulties encountered, useful data were elicited from the test. (DWH)

A FUNNY THING HAPPENED ON THE WAY TO THE PRINTER: THE SAGA OF DEVELOPING A CUSTOMIZED TEST

Dr. Joy A. Frechtling
Dr. Margot A. Schenet

Montgomery County Public Schools

Paper presented at the AERA Annual Meeting in New Orleans, Louisiana
April, 1984

A FUNNY THING HAPPENED ON THE WAY TO THE PRINTER: THE SAGA OF DEVELOPING A CUSTOMIZED TEST

Many school districts are moving away from a reliance on norm-referenced tests to a greater use of criterion-referenced tests, believing (probably rightly so) that the criterion-referenced tests can provide a better measure of learning. Research has fairly well established that norm-referenced tests do differ in content and that the overlap between curriculum and test content affects test scores. Given these somewhat simple-minded facts, it is clear that the closer the match between what is taught and what is tested, the more accurate and useful a test score will be.

Finding the right criterion-referenced test is not, however, a simple matter. Whether one goes with a test developed by "experts," the publishers, or prefers to adopt a home-grown variety, there are complications. What I'd like to do today is share with you some complications we encountered last year when we went the "expert" route. I assure you that everything I am going to be saying to you is the truth; I have not added anything to increase the dramatic effect.

First, a few words about the context of our effort. We were in the process of conducting an evaluation of a new reading/language arts program that had been developed in our county. The purpose of the evaluation was to determine whether the program was being implemented as planned and whether, once implemented, the program had any effects on achievement or attitudes toward reading. The study is a three-year, longitudinal effort in which two groups of students are being followed--a primary group (starting in Grade 1) and a later elementary group (starting in Grade 4). Because we did not at that time do any end-of-year testing in reading/language arts on a systemwide basis, we wanted a criterion-referenced test that we could administer at the end of the school year to assess the first and fourth graders' end-of-year skills.

Our first challenge was to find a set of items that the program developer would accept as being good measures of the objectives at these two grade levels. This was not an easy task as he had some very definite ideas about both what should be measured and how it should be measured. The fact that criterion-referenced tests had already been developed by the program developer and his staff (tests we did not choose to use both because they had been administered earlier in the year and because they had been highly criticized by some school staff) rendered selecting some outside measure even more difficult. We brought in several outside groups to pitch their wares, show us items, and negotiate dollars. Finally, after extensive review and time, we found a group that appeared to have a product we could live with.

The test bank, as it was described to us, had some uniquely appealing features. To start with, if one included certain items (regardless of their curriculum match) not only could the district receive criterion-referenced data but also norm-referenced data. Thus, national comparisons could be made as well as more local ones. Second, the available item bank was rather extensive, as the publisher had taken items from the many different tests produced by the company. Third, although items on the test had already been arranged into objective groupings, we were told that we could have our test

"customized." We could create our own item groupings and give them our own objective names. This was a new feature the company was offering and we became one of the first clients to take advantage of this option.

The next five months can only be described as a "comedy of errors." Having selected an "expert" group for its awareness of test construction pitfalls and developmental dilemmas, we wound up encountering one disaster after another. Whether these were caused by unworkable time constraints or incompetent staff, I won't venture to say. Let me lay out the problems we encountered and leave you to decide how you would allocate the responsibility.

PROBLEM ONE: THE LISTENING TEST

The normed items on the first-grade test were described to us by the local rep and her regional manager as orally administered. This allowed the possibility of obtaining measures of text comprehension, which were not necessarily tied to decoding skills. This had a good deal of appeal to us because our new curriculum stressed comprehension at all grade levels and some teachers felt that existing tests did not allow the student who was essentially a nonreader to show his/her listening comprehension skills. In building the first-grade test we decided, however, to include a second, nonoral section in order to be able to measure reading skills. In order to use the norm data, we included the norm items in a listening portion, added some oral items to more adequately cover the objectives we wanted to test, and then built a second, somewhat parallel "reading" portion. Accomplishing this without building a test that was too lengthy for the first grader wasn't easy, but with a good deal of work and negotiation between the evaluation and program development staff, we got it done.

Tests ordered, cover colors selected, we notified the principals whose schools were participating in the reading study that the test which had been tentatively promised was, in fact, about to become a reality. While the need for additional testing was not welcomed, the possibility of getting normed as well as criterion-referenced data, and the interest in the listening measure, seemed to result in a decision of the pluses outweighing the minuses. They were ready, if not eager.

One bit of information not included in the prepackaged material provided by the publisher was the instructions for administering the listening test. We needed to know how much time was allowed, how many times each passage and the alternative answer choices were read, etc. For some reason, the local rep was having problems getting such information from the main office and kept returning with the answer that there were no standard directions and that we should develop our own. We found this response extremely puzzling and wondered how one could have norm data without standard directions. We began to get very nervous.

Then one night around six I was home having my first scotch of the evening when I got a long distance call from the main office of the publisher. They were finalizing the test and had some questions about how items were to be grouped into subsections and objectives. In the course of this conversation, I naturally had occasion to refer to the listening section and the reading section. To make a long story short, and believe me, we had many conversations on this topic in a relatively short period of time, we found

out that the information we had been given about the items being normed as an oral test wasn't quite true. What was meant was that the item directions were to be given orally, but the test was a standard test of reading at the first grade, not a test of listening comprehension as we had been told. We had created a truly customized product and it was too late to regroup. All we could do was go ahead and create directions for test administration, and hope.

PROBLEM TWO: THE TEST INSTRUMENTS

Shortly after the conversations described above the tests arrived, nearly two thousand of them, ready for administration. We had assumed we'd be sent a galley for proofing before the actual printing, but either through oversight or perceived lack of time, this step was skipped.

The first thing we noticed as we unpacked the test booklets was that the covers were white, not the colors which we had painstakingly chosen. We were disappointed, but we figured we could live with white covers, as long as the charge for colored covers was taken off the bill.

Next we opened the booklets and found ...Pandora's box. To start with, despite our previous conversations, some of the items were wrong. We hadn't selected them, they didn't match, and they couldn't be used. If that weren't enough, the items, because they had been taken from a variety of tests, were in varied type styles. Finally, directions for the items were inconsistently worded and in some cases omitted.

It took us many conversations and displays of temper to get changes made in the format of the items so that they were reasonably consistent. While the publisher's staff did not necessarily agree that inconsistencies in format and directions might throw off first and fourth graders, they finally gave in and agreed to make as many changes as possible. They also agreed to give us tests with the items we had selected instead of ones which apparently had been developed by gremlins. We threw out the initial cartons of test booklets and waited. This time, facsimiles were telecopied to us before printing, so that we could see the instruments. The next batch was what we had ordered and looked useable.

PROBLEM THREE: THE ITEM TAPE

We made it through test administration with no new disasters, collected the booklets, and sent them off to be scored. Since the tests were being administered for the first time and we were not totally sure of whether or not all the items would be acceptable, we asked for a tape of item responses so we could conduct some preliminary analyses before score reports were produced for individual schools. We had a suspicion that some items might not "work" and that they would have to be thrown out.

Items of special concern were ones where many students had selected an incorrect answer. We wanted to be sure that the students hadn't been in some way thrown off by the wording or structure of the question and that finding a high failure rate would provide instructionally useful information. What we found instead was that in a number of instances the item had been incorrectly scored. The gremlins were at it again. When we called the publisher about this problem, we found out that they had found

and corrected that error on their tape, but had somehow failed to do so on ours. Rather than wait for them to send us a new tape (and all the possibilities that creating the new tape might open up), we went ahead, made our own corrections, and proceeded with our item analyses.

PROBLEM FOUR: ADDING TO 100

Reports of school performance included the average percentage of students mastering each objective, as well as average subtest and total test performance. When we looked at the first grade data, we found some totals that struck us as rather odd. In a number of cases, more than 100 percent of the students had passed each objective.

Solving this problem was not really very hard. What had happened was the following. At the first-grade level in grouping objectives into subtests, an objective was sometimes included on two different subtests. (The item wasn't actually on the test twice, it was used twice for scoring purposes.) The mean percentages had been calculated by a program that did not take into account this possibility. Thus, when the average was calculated, it was divided by the unique number of objectives, not the number of objectives that had actually been mastered. Apparently, no one had looked at the reports we were sent before they were shipped, as one would expect percentages of 120, 150, etc. to catch the eye. However, since apparently no one reviewed the test booklets or the item tape before these were sent, we shouldn't have been surprised.

PROBLEM FIVE: REPORTS ON NORM DATA

The publishing company offered a variety of ways of reporting norm data. The options varied in unit of analysis and type of score presented (stanines, grade equivalents, etc.). A certain number of scores came as part of the package; others required additional funds.

When we received the scores comparing our fourth grade students to the national sample, the gremlins got in the way once more. Instead of receiving the four scores that we had requested, we received one that we had requested and three that we had not. By this time, we pretty much expected that something would go wrong. Calmly, we called and explained. The correct reports were duly dispatched.

CONCLUSION

Believe it or not, after all that, the data we got from the tests were extremely useful. And, it turned out that the listening test for first graders provided us with some very useful information about the listening comprehension skills of students who were nonreaders. In fact, we were surprised at how well some students did on what had been judged to be fairly complicated passages.

How about our relationship with the publisher. It's a bit sensitive, but we're still on speaking terms (or at least were until this AERA presentation). We haven't crossed them off our list, but we're not buying another test this year. For a variety of reasons, we are turning to the home-grown alternative. From what's been happening so far, I think I'll be back next year with a whole new set of stories.