DOCUMENT RESUME

ED 274 721	TM 860 614
AUTHOR	Eads, Gerald M., II; And Others
TITLE	Identification of Improper Preparation of Students in a State-Mandated Norm-Referenced Testing Program.
PUB DATE	Apr 86
NOTE	8p.; Paper presented at the Annual Meeting of the American Educational Research Association (67th, San Francisco, CA, April 16-20, 1986).
PUB TYPE	Speeches/Conference Papers (150) -- Reports - Research/Technical (143)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Elementary Secondary Education; Goodness of Fit; Grade 4; Grade 8; Grade 11; *Mathematical Models; *Norm Referenced Tests; School Districts; *State Programs; Statistical Distributions; *Test Coaching; *Testing Problems; *Testing Programs; Test Norms; Test Wiseness
IDENTIFIERS	Chi Square; *Virginia

ABSTRACT

The purpose of this research was to develop and test a method of identifying groups of students participating in a mandated statewide norm-referenced testing program who have undergone highly specific training in preparation for, or coaching during, test administration. The implementation of the analysis procedure was intended not only to identify such groups but also to discourage school staff from engaging in student test preparation. A Chi-square based analysis was developed which compared classroom, school, and district test score distributions against empirically derived "expected" distributions anchored to the district median score for a given subtest. The statistical model for generating the expected distributions was built using the data from the 1983 administration. The fit between the expected distributions created by the model and a sample of empirical observed distributions was exceptionally close. Significance levels were set, so that only groups with highly discrepant score distributions would be identified. The analysis performed in 1984 was replicated without modification in 1985. Results showed that since the analysis and related notification procedures were implemented, there has been some reduction in the number of groups exhibiting score distributions that may be associated with improper test preparation. (Author/JAZ)

Identification of Improper Preparation of Students in a State-Mandated Norm-Referenced Testing Program

GERALD M. EADS II
Virginia Department of Education

DAVID D. S. POOR
Educational Computer Software, Inc.

LOIS S. RUBIN
Virginia Department of Education

CLARICE P. GRESSARD
University of Virginia

Running head: Identification of Improper Test Preparation

A paper presented at the 1986 American Educational Research Association Meeting
San Francisco, California

Abstract

The purpose of this research was to develop and test a method of identifying groups of students participating in a mandated statewide norm-referenced testing program who have undergone highly specific training in preparation for, or coaching during, test administration. The implementation of the analysis procedure was intended not only to identify such groups but also to discourage school staff from engaging in such specific student preparation.

A Chi-square based analysis was developed which compared classroom, school, and district test score distributions against empirically derived "expected" distributions anchored to the district median score for a given subtest. The statistical model for generating the expected distributions was built using the data from the 1983 administration the year prior to initiation of the study. The fit between the expected distributions created by the model and a sample of empirical observed distributions was exceptionally close, with no comparison exceeding $p<.20$. Significance levels were set, using a sample of the 1983 data, so that only groups with highly discrepant score distributions would be identified; additional reports (item analyses and frequency distributions) were generated for classrooms, schools and/or districts only when the significance level for that level of analysis was exceeded. These additional reports were reviewed by state department of education testing personnel to make final determination as to whether those results bore enough evidence to justify direct inquiries to the district.

The analysis performed in 1984 was replicated without modification in 1985 to determine whether any decrease in the incidence of discrepant score distributions occurred as a function of the public attention the analysis received in the intervening year. The analysis identified discrepant score distributions in 35 districts in 1984 and 25 districts in 1985. For both years nine districts were selected for further investigation.

Significant Chi-square tests do not in any sense absolutely implicate schools of improper preparation of students. Nevertheless, since the analysis and related notification procedures were implemented, there has been some reduction in the number of groups exhibiting score distributions that may be associated with improper test preparation, and a <u>notable</u> reduction in the number of anecdotal reports of improper practices.

Introduction

In 1971 the Virginia State Board of Education adopted a policy which required the State Department of Education (SDOE) to release results of a mandatory norm-referenced testing program by individual school district. Since then, district test results have been compiled and made available to the public. After statewide testing, the SDOE receives system-aggregate distribution summary reports for each of three grades in each of the 134 districts in the Commonwealth. Approximately 220,000 students are tested each year. Initial concerns about statewide results, in comparison to the nation, seem to have diminished over the years. This is presumably due to consistent improvements in average scores (overall, state average scores have increased each year since 1973-74 and all average scores are now at the national norm or higher). However, there continues to be much concern over the results of the tests in the individual districts.

Despite much of the usual caution to the contrary, the public and many school administrators continue to see the state program as a primary indicator of student achievement and even quality of schooling. The higher the scores, the better the job being done instructionally. Most appear not to be interested in other factors which mediate achievement and test scores (*e.g.*, socio-economic factors, curriculum-test match). Because school district personnel are aware of such perceptions by their clientage, and in many cases even hold these perceptions themselves, there is considerable pressure, especially in those districts with relatively low scores, to improve their measured performance in order to demonstrate the effectiveness of their instructional programs. There is substantial pressure on teachers and administrators, if not students, to generate high test scores. While most school staff respond by insuring that test content is fairly represented in the curriculum, others apparently choose to be somewhat more specific in their instruction.

The SDOE predictably takes the position that it must administer its norm- referenced testing program so that valid results are assured to the extent possible. It cannot be in the position of releasing test results known or believed to be invalid while publicly taking the position that the results bear some relationship to the "true" achievements of students. Efforts to improve test scores may in some cases involve strategies which give students a distinct advantage over students receiving normal instruction (even that influenced by test content), possibly raising test scores without a concomitant rise in overall achievement. Such practices most assuredly invalidate norm-referenced test results.

The problem of overzealous student preparation reached crisis proportions following the release of the spring 1983 results. As a result of direct evidence of improper preparation of students for testing, the SDOE worked directly with several school districts to ascertain the validity of their results. Some of these districts had produced both average scores and percentile-based frequency distributions at substantial variance with data that had been collected in previous years. One such district's students were retested with the alternate form of the commercial norm-referenced test used in the state program; the average scores from that testing were comparable to the previous years' data. In some cases districts were accused by the public of improper testing practices by individuals associated with those school systems.

For these reasons, the SDOE became interested in identifying or developing a means to identify, and hence reduce the incidence of, such practices as providing (1) specific instruction concerning test items and (2) hints or other help to students while they are in the testing situation. Any strategy to identify improper testing preparation needed to incorporate the following characteristics:

1. be capable of identifying "suspicious" data at the class and school as well as the district level.
2. use data from the current year's test administration. Comparison of year-to-year changes was considered less than ideal due to the size and expense of the analysis.
3. produce a data set that could be readily and efficiently analyzed.

After an extensive search, it appeared that the literature on detection of improper test-taking practices was almost non-existent, and that which did exist (*e.g.*, Angoff, 1974; Frary & Tideman, 1977) was limited to the detection of individual cheating during administration. However, SDOE interest was in identifying potential instances of what might be referred to as *group* cheating. Several alternate strategies were delineated and considered. Some of the methods comprised some form of statistical analysis of data from the form administered in the program, and others made a comparison of the data from one form of the test with data from alternate forms.

Development of a Model

A statistical procedure using a χ^2 distribution model was chosen as most likely to be successful, within the given constraints, to detect test score distributions that failed to conform to an expected distribution. The procedure had been delineated but not yet developed or tested. Development was organized into three phases:

a. Discrete subsets of data were selected from the initial year's data using the newly adopted test to obtain empirical distributions of test scores of students from high-, middle-, and low-achieving districts representing all geographic areas of the state. High-achieving districts were defined as those having grade-level average subtest scale scores corresponding to at least the 60th percentile; middle-achieving districts were defined as those having aggregate percentiles between the 45th and 55th percentiles; and low-achieving districts were defined as those having aggregate percentiles at or below the 40th percentile. Some exceptions were made to these criteria in order to equalize representation by geographic region. Using these empirical distributions as baseline data, a statistical model was generated to predict distributions, given the overall characteristics of the subgroup, and tested for closeness of fit between the empirical distributions and the model.

b. The statistical model for determining the "expected" distributions and testing the fit between the observed and expected distributions was examined on a selected sample of data so that the automated identification of non-conforming distributions could be manually checked against historical data.

c. The automated process was utilized to analyze the 1984 data as part of the normal scoring and reporting procedures.

Determination of empirical distributions

School districts were selected as having high, middle, and low achievement in Reading, Mathematics, and Language, for each of the three grades tested (4, 8, and 11). Composite scores were also analyzed at grade 4. Thirty-two separate computer runs were made to provide state distribution summary reports with the selected subsets of districts.

From each of the distribution summary reports, the percentage of observed students in each of the ten deciles defined by the national percentiles was recorded. These observed decile percentages defined the empirical data which were used as a baseline for preliminary evaluation of the statistical model.

Specification of the statistical model

The preliminary evaluation of the empirical distributions indicated that the mean test scores were less stable predictors of the distributions than the median test scores from the same distributions. Cursory examinations of the distributions indicated relatively stable distributions for specific median test scores.

The expectation distribution was defined in terms of ten deciles: *i.e.*, the percentage of students falling in each of the ten deciles, as defined by the national percentile. It was also noted that even when the median percentile was 50, the scores did not indicate a flat distribution, but were more leptokurtic than the normal distribution would predict. Because of this determination, it was decided that the model should utilize only the median as the measure of central tendency, and explicitly not include any measure of dispersion. This determination was made so that the model would not be influenced by any reduction in variation in scores due to administrative irregularities.

It was then decided to build the statistical model on the Normal Curve Equivalent (NCE) metric. The inital model was tested with the following parameters:

a. The "expected distribution" is anchored at the observed median. 50% of the cases in the statistical model are placed above the median, and 50% of the cases are placed below. The shape of the curve from the median to the distribution endpoint was assumed to conform to the shape of the normal curve, only compressed or expanded to cover the range from the median to the endpoint.

b. The endpoints were set at NCE 1 and NCE 99 (*i.e.*, -2.326 and +2.326 *z*-score units from the normal curve median).

The initial model "expected distributions" and the empirical distributions indicated that the model tended to place too few students in the lowest and highest deciles relative to the empirical distributions. In order to compensate for this deviation and the truncated range, the model was adjusted to extend the endpoints to -3.0 and +3.0 *z*-score units in a normal distribution. This resulted in NCE endpoints of -13.18 and +113.18.

This revised model provided a better fit at the end of the distribution closest to the median, but placed too many cases in the extreme deciles most distant from the median. Based on this observation, it was decided to implement a "floating endpoint" model in which the endpoint was adjusted from 2.326 to 3.0 *z*-score units away from the norm distribution median as a function of the location of the median. (For a median NCE of 30 or less, the endpoints were set to -3.0 and 2.326. For a median NCE of 70 or higher, the endpoints were set to -2.326 and 3.0. Endpoints for medium values between 30 and 70 were set as a linear function between the two so that there was a uniform set of endpoints only with a median NCE of 50.)

This revision of the model essentially acknowledged that, for groups of students with observed median performance higher than the 50th percentile, the lower limit of the lst percentile is reasonable, but that the 99th percentile is an artificial ceiling that prevents students from indicating their true ability. A similar effect occurs for groups with median scores below the NCE of 50, in which the 1st percentile becomes an artificial floor.

An additional adjustment was made to the model in order to reduce the expected dispersion of the scores. This adjustment was implemented because the scores from any subgroup of like-performing schools or districts would probably be more homogeneous than the total set of scores in the national norm group. With this adjustment, each "expected distribution" of scores was compressed relative to the prior model. This compression was implemented by having the 50 percent of the cases distributed from the median to the appropriate endpoint less some percentage of the range between the midpoint and the endpoint. Three different factors of compression were tried, representing compression of 20%, 15%, and 10%. Because the 15% and the 20% compensation placed too many cases in the deciles near the median, floating endpoints with a 10% compression cf the distribution to recognize the relatively homogeneous nature of a subgroup's data were finally chosen.

The final parameters of the statistical model are as follows.

a. The "expected distribution" is "anchored" at the observed median. 50% of the cases in the statistical model are placed above the median, and 50% of the cases are placed below. The shape of the curve from the medium to the distribution endpoint is assumed to conform to the shape of the normal curve, only compressed or expanded to cover the range from the median to the endpoint.

b. The endpoints are set as a function of the median, with the lower endpoint ranging from a NCE of -13.18 to 1 (z-score of -3.0 to -2.326), and with the upper endpoint ranging from a NCE of 99 to 113.18 (z-score of 2.236 to 3.0).

c. The expected distribution is further compressed by ten percent (relative to the normal distribution).

Fit between the model and the empirical distributions

The model, as described above, was compared to the 32 separate empirical distributions that had earlier been generated. The distribution that demonstrated the greatest discrepancy with the model (grade 11, low achieving districts in Mathematics) was (a) noticeably more leptokurtic than other distributions with the same median percentile score, and (b) not statistically significant (p<.20).

The comparison between the expected and observed models showed an exceptionally good fit with 22 of 32 cases producing χ^2 test results of less than 2.53. This is in contrast to the expectation of having values greater than 2.53 for 98% of non-conforming distributions.

Significant statistical analyses automatically produced a set of score reports (including distributions summaries and item analysis reports) of suspected groups at the classroom and school level for review. However, during the first year of using the analysis, it was discovered that "traditional" probability levels produced a large amount of data that proved to be of little value in producing observable evidence of improper preparation. It was therefore decided to select for report generation groups producing χ^2 values only above certain (high) levels. Ultimately, reports were generated only for distributions generating χ^2 values in excess of 100. This change resulted in reports that in some cases revealed extremely unlikely events (*e.g.*, all students in the same classroom choosing the same incorrect foil, or students in a classroom performing far above expectations on one subtest, but performing as expected on all others).

The net effect of the use of the procedure during the scoring process, and its timely feedback, has been to substantially decrease the incidence of at least the *reports* of improper testing preparation and coaching. It has, however, proven extremely difficult to assess the impact on actual practice.

Discussion

The analysis using the statistical model appears to perform reasonably well in detecting aberrations at the classroom level. Because data are usually aggregated at the classroom level in the lowest grade tested (grade 4), the analysis is more "fruitful" at that level. However, since data are typically aggregated only at the school and district level in the middle and high school grade levels (8 and 11), the analysis is sensitive only to relatively rampant school-wide preparation of students for individual subtests. Thus, for a meaningful analysis at these upper levels, student data would need to be grouped into subject area classrooms.

References

Angoff, W. H. (1974) The development of statistical indices for detecting cheaters. *Journal of the American Statistical Association. 69*, 44-49.

Bunch, M. B., and Kahn, P. L. (1982). The validity of local program evaluation results: An empirical investigation of Murphy's Law. Paper presented at the Annual Meeting of the Eastern Educational Research Association.

Frary, R. B. (1976). Evaluation of statistics for detection of cheating on multiple-choice tests. Paper presented at the Annual Meeting of the American Educational Research Association.

Frary, R. B., Tideman, T. N., and Watts, T. M. (1977). Indices of cheating on multiple-choice tests. *Journal of Educational Statistics. 2* (4).

Seretsky, G. D. (1984). Treatment of scores of questionable validity: The origins and development of the ETS Board of Review. ETS Archives Occasional Paper, Princeton, NJ: Educational Testing Service.