PAUL G. LEMAHIEU AND RICHARD C. WALLACE, JR.

Establishing the Credibility of Test Results

The Pittsburgh Public Schools validated the results of their regular testing program by administering a second, similar test under secure conditions.

The public that supports American education has every right to know how schools are doing and that the information it receives on student achievement is accurate. Therefore, charges that call into question the integrity of the testing process are extremely serious. Nevertheless, every school board member, superintendent, and director of testing will admit in confidence to having heard of "nonstandard practice," "testing irregularities," or outright cheating within their districts.

Monitoring Achievement Test Data in Pittsburgh

In the fall of 1980, under the guidance of a new superintendent, the Pittsburgh Public Schools initiated a host of reforms designed to improve the quality of education and to increase student achievement in basic skills throughout the district. Over the next several years, Pittsburgh (like so many other districts engaged in concerted self-improvement efforts) observed dramatic gains in its achievement test results. In 1980, 53 percent of students in grades 1 through 8 scored at or above the national norm in mathematics; by 1984 that figure had increased to 74 percent. Gains in other subjects over the same period were nearly as dramatic: from 49 percent to 60 percent in reading, and from 48 percent to 71 percent in language. Almost as soon as these gains were announced, questions arose about the credibility of the district's achievement test data.

In 1984, we decided to investigate the validity of our achievement test results. The public's faith in the district demanded it, and the district's leadership, with its need for high quality data to inform educational decisions, demanded it as well. Although it was important to seek answers to these questions, we did not want to engage in a punitive exercise in investigation and detection for several reasons.

1. *The detection approach to this problem is inherently self-fulfilling.* Where there are many teachers (perhaps thousands) involved in test administration, somebody may be up to something. This is especially true in that teachers may have varied reactions to testing circumstances and approach tests with different motives and different degrees of perceived stress. Frank recognition of the existence of some nonstandard practice in the administration of a testing program cannot and should not be equated with condoning it. A district must act swiftly and decisively in addressing irregular practices whenever they become known.

Approaching the issue as one of detection, however, will only validate the objections without answering the important questions. If some are indeed engaged in irregular test administration, it will undoubtedly be detected and thereby substantiate every anecdote and rumor that calls the testing results into question. This will be true even if only one person in the district cheats and is not doing anything on a scale that would invalidate the district's results. Recognizing this only serves to point out that the important question to ask at the district level is not, "Is anything irregular happening?" but rather the more salient, "Is anything irregular happening to a degree that causes us to distrust what we are saying about ourselves to the public?"

2. *The investigative/detective mode of thought is impossible to support.* No district has enough resources to maintain a fully secure testing program. The mechanics of administering a program as well as the ready availability of test forms militate against it. Indeed, the very notion of a fully secured program is probably quixotic in the

context of increased (and in many places unreasonable) emphasis on test scores.

3. *The investigative/detective approach would critically compromise the testing office's ability to be of genuine service to the district.* It would quickly become apparent that the office's role is one of enforcement, and if this sort of policing becomes a primary function, it would soon become unable to perform the many services that are of greater educational value to the district. It would be a serious loss to proscribe the participation of the testing office in the educational life of the district by placing it in this role.

For these reasons, we decided to approach the credibility issue as one of validation rather than investigation. In the spring of 1984, we conducted a study to determine whether districtwide achievement as indicated by the results of our regular testing program presented an accurate assessment of the students in Pittsburgh. In short, we administered a second achievement test on an extremely secure basis to provide independent estimates of student performance throughout the district and used it to monitor our regular testing program.

Design of the Study

A plan was formulated to provide a representative sample of students contributing to achievement estimates in reading, language, and mathematics. All of the elementary schools in the district were stratified by size and racial composition as well as prior student achievement as indicated by the past years' testing. The schools were organized according to this scheme, and then classes were randomly selected at appropriate grade levels within each building. Thus, students from all buildings at all levels of achievement, racial composition, and size contributed to each separate achievement estimate. All schools participated in the validation process. The estimates were developed from samples ranging in size from 225 to 313, or approximately 9 to 13 percent of the population of interest.

All students in the district were tested using the appropriate level of the California Achievement Test (CAT), as has been a longstanding district policy. The students in the validation sample were required to take an additional test, the Comprehensive Test of Basic Skills (CTBS), to which the district's CAT performance was compared. This test was chosen because it offered certain technical advantages:

1. It was the most recently normed of the available tests; thus, performance could be estimated with reference to students' achievement in 1982, the most up-to-date comparison possible.

2. The developers of the test (CTB/McGraw-Hill, Inc.) had carried out an equating study relating the scores on the CTBS to those on the CAT. The CTBS scores could then be translated into estimated CAT scores to permit ready comparisons of the Pittsburgh students' performance on the two tests.

3. The two tests are quite similar in physical structure; thus, the use of the CTBS minimized error or bias in estimating CAT performance that might have been introduced by differently structured tests.

The students who participated in the validation study were tested using a CTBS test in one subject area in addition to the regular CAT battery. The order of test administration was counterbalanced to control any possible practice or fatigue effects: half of the participants received the CTBS test first, while the other half received the CAT test first. The time lag between the administration of the two tests was one day. Classroom teachers administered the tests according to the directions for standard administration, and the CTBS administration was also monitored in each classroom by a support professional (supervisor, program specialist, and so on) to certify standard practice. These observers obtained the testing materials just two days prior to the actual test date and

transported them to the assigned schools on the actual day of testing. After monitoring the tests, they immediately returned all testing materials and student answer sheets to the Division of Testing and Evaluation. In short, the validation instrument was maintained and administered as securely as is possible.

Analyses and Results

Separate frequency distributions of national percentile scores were created and inspected for each grade level and aggregated across grades 1 through 8. Next, test equating tables prepared by the California Test Bureau were used to identify the performance levels on the CTBS that corresponded to the norm of the CAT. This equated CTBS score was used as a cutoff point in calculating the proportion of students whose performance was at or above the point on the CTBS corresponding to the national median of the CAT. Thus, an estimate of the proportion of Pittsburgh students scoring at or above the national norm on the CAT was derived from the closely monitored administration of the independent CTBS. This estimate is directly comparable to the "Percent of Students Scoring At or Above the National Norm" on the district's regular CAT administration.

Figure 1 summarizes the results of these analyses. Column 4 presents the percentage of district students actually observed in the regular CAT testing program to be at or above the national norm in each of the three subject areas. Column 6 lists the estimates based on the secured testing event. To evaluate the question central to this study, these estimates should be compared to the observed CAT results. When we inspected those differences for grade levels 1 through 8, we found the CTBS estimate to be only one percentage point lower than the CAT performance in mathematics, four percentage points lower in reading, and three percentage points higher in language. When these broad estimates are further examined by organizational units or grade levels, the differences vary somewhat by grade level and subject. For the most part, however, the CTBS estimates appear to be well within tolerable limits of the performance level indicated by the CAT.

Summary

The results of our validation study and analyses indicate that there are no consistent statistical differences between the equated scores on the CTBS and the district's CAT results. There are specific grades and subjects in which differences may be observed, but these differences have no consistent patterns. Thus, the results of the regular testing are credible, and the public can have confidence in them. In general, secure items that test the same content domains as those on the CAT produce comparable test results. Further, when examining the percentage of correct items on the CTBS, the ones similar in content to the CAT resulted in equally high performance levels (CAT, 65 percent correct; CTBS, 75 percent correct). This indicates that teachers did not simply teach the CAT *items*. Even with intensely focused test pressure, the national norm interpretations given to rising CAT scores seem believable overall.

While these results support the integrity of the district's regular testing program as a whole, our validation study was never meant to—and therefore does not—address issues of individual behavior with respect to achievement testing. The results only confirm that compromising behavior is not prevalent to a degree that would call into question the statements of achievement for the district as a whole.

"[T]he important question to ask at the district level is . . . 'Is anything irregular happening to a degree that causes us to distrust what we are saying about ourselves to the public?' "

We have described how districts might address one of the toughest questions they face concerning the credibility of their test performance. The profession demands a high standard of integrity of our teachers and principals even in the face of extreme pressure. If that same measure of integrity extends to the superintendent's office, then these questions must and will be addressed directly, as they have been here. The risks may be great, but they must be taken. Like the questions, not all the answers may be pleasant ones; nonetheless, confidence in a growing perception of improvement in public schools is essential. Indeed, the very fact of willingness to ask such tough questions can only redound to a district's benefit in the form of increased support from a properly informed public.□

Figure 1. Validation Study Results (Estimates of CAT Performance Based on CTBS)

(1) Subject	(2) Grades Subsumed by Analysis	Observed CAT Performance (3) Number Tested	Observed CAT Performance (4) Percent At or Above National Norm	Estimates Based on CTBS (5) Number Tested	Estimates Based on CTBS (6) Percent At or Above National Norm
Reading	2–8	17,898	57	1,752	53
Language	2–8	17,944	71	1,797	74
Mathematics	1–8	20,834	74	2,098	73

Paul G. LeMahieu is Director, Division of Testing and Evaluation, and **Richard C. Wallace, Jr.** is Superintendent of Schools, both with the Pittsburgh Board of Public Education, Administration Building, Bellefield and Forbes Avenues, Pittsburgh, Pennsylvania 15213.