Mid-Western Educational Researcher

Volume 10 | Issue 1 Article 2

1997

Effectiveness Testing Practices: Educators' Perceptions of the Effectiveness of Their Schools' Standardized Testing Practices

Ronald N. Marso
Bowling Green State University

Fred L. Pigge
Bowling Green State University

Effectiveness Testing Practices

Educators' Perceptions of the Effectiveness of Their Schools' Standardized Testing Practices

Ronald N. Marso and Fred L. Pigge
Bowling Green State University

Abstract

This study was designed to collect and then to compare teachers', principals', supervisors', and testing directors' (N=484) ratings of the effectiveness of selected standardized testing program management practices in their schools. It was found that these educators, who were selected for being knowledgeable about their testing programs, rated their schools' performance in standardized testing higher than in meeting other district responsibilities. The highest rated testing practices were use of quality tests and materials, maintenance of pupil records, and use of understandable scores and reports. The lowest rated testing practices were the use of test results to evaluate instruction, availability of written policies, and use of publisher instructional guides accompanying achievement batteries. Comparatively, educators assigned to secondary schools tended to rate the testing practices lower than did their elementary school cohorts; just the ratings of the teachers differed significantly among the various job assignment groups; and the job assignment groups provided similar relative ratings of the testing practices with most Spearman Rho coefficients being +.73 or higher.

Educators generally do not have a high regard for standardized testing despite the increased use of these tests in recent school reform efforts (Haney & Madaus, 1989). For example, many classroom teachers appear to have an unfavorable to indifferent attitude toward standardized testing (Borg, Worthen, & Valcarce, 1986), and school administrators tend to view standardized testing as being a relatively unimportant administrative function in their schools (Sproull & Zubrow, 1981). Additionally, assessments of the research literature reveal that testing and evaluation practices receive less attention from educational researchers than many other aspects of education (Crooks, 1988).

This less than positive regard for standardized testing is also revealed in what many educators believe about testing. Classroom teachers commonly believe that standardized testing skills are less needed than are other testing skills (Marso & Pigge, 1988); many teachers perceive the primary benefits of their school districts' standardized testing programs accrue not to themselves but to the school administration (Salmon-Cox, 1981); building principals typically do not perceive the need for testing specialists to be involved in the selection of standardized tests (Kinney, Brickell, & Lynn, 1988); and school counselors frequently feel testing services dominate too much of their time (Miller, 1977).

Furthermore, this less than positive attitude of educators toward standardized testing may be having an undesirable impact upon standardized testing practices in the K-12 schools. For example, many teachers report very limited use of the results from standardized testing in their classroom instruction (Linn, 1990), and educational administrators frequently do not convey the results from standardized testing to their teachers (Wood, 1982). Further curtailing the effective use of the results from standardized testing, the results of this testing, if made available, typically are not available to educational staff until six or eight or more weeks after test administration (Hall, Carroll, & Comer, 1988).

Additionally, some researchers have attributed the rather recent movements toward alternate pupil achievement assessments to the belief that existing standardized measures are too narrow in scope and may even have a negative impact upon classroom instruction (Miller & Legg, 1993). Other research findings have suggested that recent pressures in schools to show improved achievement scores have lead to questionable, if not unethical, methods of raising test scores (Nolen, Haladyna, & Haas, 1992). For example, observations of classroom instruction have revealed that external testing programs may substantially reduce time available for instruction and reduce teachers' use of the variety of instructional materials and methods available to them (Smith, 1991). Surveys of teachers reveal the existence of perceived pressures, particularly in lower socio-economic schools, to improve test scores by planning instruction around tests, by increasing time spent on reviewing previously presented content, and by teaching various test-taking strategies (Herman & Golan, 1993). Relatedly, surveys of adolescent pupils indicate that they have become suspicious and cynical about standardized tests and commonly do not respond with positive test-taking strategies when being tested (Paris, Lawton, Turner, & Roth, 1991).

In brief, the existing research literature does not specifically address the effectiveness of K-12 schools' standardized testing practices. This existing research literature has indicated, however, that educators do not hold standardized testing in high regard, that limited management attention is

veys (85%) completed by the testing directors. A check of school district size indicated that size in itself did not influence whether or not a testing director participated in the study (Marso & Pigge, 1990). Also, just those teacher supervisors employed by selected school districts were included in the supervisors group. Several school superintendents reported either that no formal teacher supervisor positions existed in their district or that teacher supervisory services were provided through their county offices of education.

The respondents were employed in schools organized by city district (42%), local county district (44%), and exempted village district (14%), in schools located in geographic settings described as rural (37%), suburban (57%), and urban (6%), and in small schools (11% with fewer than 1,000 pupils), moderately sized schools (34% with 1,000 to 2,000 pupils), moderately large schools (34% with 2,001 to 4,000 pupils), and large schools (21% with more than 4,000 pupils). These proportions of respondents representing different types of school settings were judged to be approximately similar to the composition of all such schools as reported in the Ohio Education Directory.

The focus of the present report is upon these educators' responses to 10 survey items related to their school district's practices associated with the management of standardized testing. They responded to each of the 10 testing practices by rating the "relative effectiveness" of their school district's testing practices or procedures during the past year or two. The reference to this time period was provided to create a common time period for the ratings and to avoid consideration of proposed, but yet to be implemented, state-mandated high school proficiency testing in the schools. The data collection for this study was completed during spring term of 1989 prior to the initiation of state-mandated standardized testing programs; therefore, the directions to the respondents as to which standardized tests to consider in their ratings were not necessary. Previous surveys of the public schools in Ohio had indicated that group standardized testing primarily consisted of the scheduling of reading achievement, achievement batteries, and scholastic aptitude tests in the elementary schools and of interest inventories, multiaptitude tests, and very limited use of subject area achievement tests in the secondary schools.

In addition to the time reference, the educators also were provided with a second common rating reference. They were directed to rate their schools' effectiveness in performing the 10 testing practices compared to their schools' overall performance in meeting responsibilities as educational institutions. It was assumed that most respondents would lack a common comparative performance reference across school districts but that they would possess knowledge of the overall performance of their own schools. It was determined, therefore, that the overall district performance reference point would provide much more meaningful ratings than would allowing the respondents to bring to the rating task whatever unspecified reference point that occurred to them at that moment.

A five-point scale with narrative descriptions at each scale point and with an accompanying "DK" response option, defined as "I really do not know," was provided with each of the 10 testing practices items. The "I really do not know" response option was added to discourage ratings of testing practices about which the respondents might feel uninformed. This was deemed to be consistent with the researchers' goal of seeking ratings just from educators knowledgeable of their schools' testing practices. This scale ranged from "we perform well below our average" (1) to "we excel" (5).

Three sets of statistical analyses of the collected data were completed. One and two-way ANOVA procedures were used to identify significant rating mean differences among the teacher, principal, supervisor, and test director respondent groups and among these groups when classified by secondary or elementary school assignments. The job assignment and grade level interactions were also tested and discussed. An alpha level of .05 was selected for the ANOVA's while a .10 level was selected for the pair-wise post-hoc Scheffe tests. This pair-wise procedure readily handles unequal n's and is the most conservative of these procedures to the extent that Scheffe recommends use of the .10 level (Hinkes, Wiersma, & Jurs, 1994). These ANOVA procedures were completed on the data derived from respondent ratings of each of the 10 testing practices. In addition, Spearman Rho correlations were completed between the various groups of educators' ranked rating means for the selected testing practices to ascertain the extent of agreement among the educators as to which of their schools' testing practices were rated to be more or less effective.

Results

Each of the four groups of educators, testing directors, classroom teachers, teacher supervisors, and principals rated their school's performance of the selected 10 testing practices about average or somewhat higher (3 or higher on the five-point scale) compared to the performance of their schools in meeting their overall responsibilities as educational institutions. Only when the teachers, principals, and supervisors were classified by elementary and secondary school assignments were any rating means found below the "about average performance for us" or '3' level. Just two of the rating means of the secondary teachers and one of the rating means of the secondary supervisors were below this average, whereas none of the mean ratings of the secondary principals, the testing directors, and the elementary level educators were below the "about average" or '3' level.

The testing practices rated more effective by the educators were management of pupil records, use of quality tests and materials, selection and administration of tests, and use of understandable scores and reports (items 8, 3, 1, and 5, respectively). Practices rated less effective were use of the results of achievement battery testing to evaluate district classroom instruction, provision of instructional guides ac-

The one-way ANOVA procedures indicated that elementary and secondary teachers as a collective group rated test selection and administration (item 1) significantly lower than did the combined groups of elementary and secondary supervisors, principals, and directors. These teachers also rated test scheduling at times to aid decision-making and prompt return of testing results (items 2 and 4) lower than did the supervisors and directors. In contrast, the teachers rated the provision of criterion-referenced data from achievement batteries (item 9) higher than did the testing directors. When these means were rank ordered, the directors' ratings were found to be highly related to those of the principals (Rho = +.93) and the supervisors (Rho = +.93), but somewhat less so with the teachers (Rho = +.73).

The one-way ANOVA and Scheffe procedures just for the directors and the elementary educators indicated that the elementary teachers' ratings were lower than the directors' ratings of practices related to test selection-administration (M's = 3.57 & 4.01), test scheduling (M's = 3.40 & 3.90), and prompt return of test results (M's = 3.13 & 3.70), items 1, 2, and 4, respectively. In contrast, the elementary teachers' ratings were higher than the directors' ratings for the provision of criterion-referenced data (M's = 4.05 & 3.29) and the handling of pupil permanent records (M's = 4.41 & 4.03), items 9 and 8, and the elementary teachers' ratings were higher than the directors' and elementary principals' ratings of the provision of instructional guides (M's = 3.79, 3.20, & 3.23, respectively) and the availability of written school policies regarding pupil records (M's = 3.65, 3.10, & 3.00, respectively), items 6 and 7. The Spearman Rhos between the rank ordered rating means of the testing directors and the three groups of elementary educators indicate that the elementary teachers perceived their schools' relative performance of the various testing practices somewhat differently than the other educators but that considerable agreement existed among the other groups of educators. Positive Rhos of +.49, +.55, and +.60 were obtained between the elementary teachers and directors, elementary principals, and elementary supervisors, respectively; whereas Rhos between the elementary principals and supervisors, directors and elementary supervisors, and directors and principals were +.80, +.85, and +.93, respectively.

The one-way ANOVA procedures just for the directors and the secondary educators indicated that the secondary teachers' ratings were lower than the secondary principals' ratings of the use of understandable scores and reports and of the use of achievement batteries to evaluate district instruction (items 5 and 10). The secondary teachers' ratings were lower than both the directors' and secondary principals' rating of the practices of test selection-administration (M's = 3.51, 4.01, & 3.97, respectively), test scheduling (M's = 3.36, 3.90, & 3.87, respectively), test and materials quality (M's = 3.81, 4.17, & 4.25, respectively), and promptness of test results (M's = 3.14, 3.70, & 3.67, respectively) items 1, 2, 3, and 4, respectively. Additionally, the ratings of the secondary teachers (M = 2.42) were lower than both the directors' (M = 3.20) and supervisors' ratings (M = 2.92) for the provision of instructional guides to aid instruction (item 6). Unlike the elementary teachers' ratings, all of these ratings of the secondary teachers were lower than those of the other noted groups. The secondary teachers, however, perceived the relative effectiveness levels of their schools' performance of the selected testing practices more similar to the other secondary education groups than did their elementary teacher cohorts. The Spearman Rhos between the rating means of the secondary teachers and directors, secondary teachers and principals, and secondary teachers and supervisors were +.87, +.94, and +.92, respectively. The related Rhos among the secondary pairs of directors and principals, directors and supervisors, and principals and supervisors were +.95, +.79, and +.84, respectively.

The two-way ANOVA procedures, completed without the directors but with the elementary-secondary assignment classification of the remaining groups of educators, revealed that the elementary school educators (combined principals, supervisors and teachers) rated higher the provision of instructional guides for instruction and use of scores for evaluation of district instruction (items 6 and 10) than did their secondary cohorts (see Table 1). The job assignment main effect comparisons identified significant differences in the ratings of the teachers, principals, and supervisors for test selection and administration (item #1), test scheduling (item #2), and making test results available promptly (item #4). In each case the rating means of the teachers were the lowest of the three groups; however, the Scheffe pair-comparisons identified a difference among the rating means just for the test selection and administration practice.

These two-way ANOVA procedures also revealed significant job-group and grade-level interactions among the rating means for four items. For each of these four testing practices, understandable scores and reports, availability of instructional guides, presence of school policies, and provision of criterion-referenced test data, the secondary teachers' ratings (items 5, 6, 7, and 9, respectively) were sharply lower than those of the elementary teachers. Additionally, the ratings of the elementary supervisors and secondary supervisors differed sharply on the effectiveness of the provision of criterion-referenced analysis from achievement batteries (item #9). Figure 1, the graph of the rating means for the provision of criterion-referenced data, illustrates the elementary and secondary teachers' differences common to

elementary educators. The ratings of the testing directors, supervisors, and the principals did not differ significantly one from the other for any of the 10 testing practices, and the Rhos between the ranked rating means of these groups all exceeded +.90. Also, few differences were identified between the respondents when grouped as secondary and elementary educators, and when these differences were identified they resulted from differences between the ratings of the elementary and secondary teachers with but one exception.

The differences found between the ratings of the elementary and secondary teachers may simply reflect the differences in the focus of standardized testing in the elementary as compared to the secondary schools. In the elementary schools, the focus of standardized testing is upon the guidance of pupil instruction with reading tests, achievement batteries, and scholastic aptitude tests being most frequently administered. In the secondary schools, achievement batteries and general aptitude tests are less frequently scheduled as typically the focus of standardized testing has changed from instruction to career selection with the administration of multiaptitude batteries, vocational interest inventories, and college admission tests (Mehrens & Lehmann, 1987). Consequently then, one might expect secondary teachers to perceive standardized testing programs to be of less use to them than do their elementary school cohorts as was the case in the present study.

Similarly, the statistical interactions identified between the job assignment and the job grade level classification in the present study might also be explained by differences in the focus of the standardized testing programs in the secondary and elementary schools. For example, the nature of score reports, the practices related to the storage of cumulative pupil records, the availability of instructional remediation guides, and the provision of criterion-referenced data after achievement battery testing are all practices likely to vary considerably between elementary and secondary schools. The elementary grade aptitude and achievement test reports tend to be less complex than the secondary school vocational aptitude and interest test reports; remedial instructional guides accompanying achievement batteries are less commonly used in secondary schools than in elementary schools; cumulative pupil records typically are stored within self-contained elementary classrooms but typically are stored in central locations in secondary schools; and typically criterion-referenced data are available just for achievement batteries which are more frequently administered in elementary schools than in secondary schools.

The pattern of high and low rating means for the 10 testing practices noted in the present study suggests possible implications for the management of standardized testing programs. Certainly, first and foremost, the ratings of these educators suggest that standardized testing programs are perceived to be functioning effectively as compared to the overall performance of the schools in meeting their overall goals as educational institutions. Each of the groups of educators in the present study appeared to be satisfied with the quality of the tests, testing materials, report forms, and the management of pupil records. On the other hand, these educators appeared to be less positive about the effectiveness of the use of achievement battery scores in part to evaluate classroom instruction. The teachers appeared to be less satisfied with test selection, test administration and scheduling, and the prompt availability of the results from testing than were the other three groups of educators. Conversely, the elementary school teachers appeared to be more satisfied with the effectiveness of the guides for remedial instruction and of criterion-referenced data accompanying achievement batteries than were the other three groups of educators.

Practicing testing directors might prudently build upon the present satisfactions of their administrative cohorts but strive to enhance interactions with classroom teachers related to the operation of their testing programs. In particular, it appears that these testing directors along with the other educational administrators ought to work more closely with teachers in the selection, administration, and scheduling of tests; in the prompt dissemination of test results; in preparing written policies for school testing programs; and in making available to teachers remedial instructional guides accompanying achievement batteries to better enhance classroom instruction. The differences in typical standardized testing in the elementary and secondary schools and the present findings suggest that these collaborative efforts might be more essential in the elementary as compared to the secondary schools. Lastly, it would seem that testing directors should investigate the major discrepancy that appears to exist between elementary teachers' and elementary teacher supervisors' perceptions of the effectiveness of criterion-referenced data in linking testing results with classroom instructional activities. Measurement specialists typically expect those educators and administrators most directly responsible for classroom instruction, such as elementary teacher supervisors, to be the strongest advocates of the provision of criterion-referenced data to support classroom instruction (Mehrens & Lehmann, 1987), but it appeared that this may not have been true of the elementary supervisors in the present study.

Appendix: Rating Form

SECTION IV. School Standardized Group Testing Program Practices or Procedures.

Please rate each of the following group testing practices or procedures in terms of the relative effectiveness of what happens in your school(s). Please respond to each item the best you can although you may be more or less informed about some of these practices. Please circle your rating of effectiveness using the code below.

	Relative Effectiveness* Response Codes
'1'	We perform well below our average* here
'2'	We perform below our average here
'3'	About average performance for us
'4'	We perform somewhat above average here
'5'	We excel here
'DK'	I really do not know

* Your perception of your school's performance on this practice relative to its overall performance as an educational institution.

	Practice or Procedure	Relative Effectiveness					
		LOW				HIGH	(?)
1.	Effective test selection/administration/scheduling for standardized testing program (overall)	1	2	3	4	5	DK
2.	Tests are scheduled at times to aid decision-making	1	2	3	4	5	DK
3.	Quality tests, materials, and reports are used	1	2	3	4	5	DK
4.	Results of tests are available promptly to aid use of results	1	2	3	4	5	DK
5.	Understandable scores, narrative reports and pupil profiles are used to report performance	1	2	3	4	5	DK
6.	Teachers' instructional guides are made available to all teachers to aid instructional use of achievement battery results	1	2	3	4	5	DK
7.	Written school policies are available for access/dissemination/ storage of test results	1	2	3	4	5	DK
8.	Student permanent records are updated periodically (dated information removed, new added, etc.)	1	2	3	4	5	DK
9.	Criterion-referenced achievement battery results are provided as well as norm-referenced scores	1	2	3	4	5	DK
10.	Achievement battery scores are used in part to evaluate district classroom instruction	1	2	3	4	5	DK