The Irish Journal of Education, 1985, *xix*, 1, pp. 39-50.

THE REACTIONS OF TEACHERS TO SURPRISING STANDARDIZED TEST SCORES

Peter Archer*
Educational Research Centre
St Patrick's College, Dublin

Research findings indicate that teachers rarely alter their opinions of students in the light of information they receive from standardized tests. To explore why this might be the case and to get an overview of teachers' responses to standardized-test results, 30 primary-school teachers (in second and fifth standards) were interviewed about their reactions to the scores obtained by their students on a previously administered ability test. It was found that teachers often did not regard as surprising test scores that appeared, on the basis of statistical criteria, to differ from the teachers' previously given assessments of students. Thus, in a large proportion of the cases where there was room for test information to have an effect on teachers' assessments, the teachers did not revise their assessments since they had not perceived a discrepancy between the test results and their own assessments. In the cases in which teachers did perceive discrepancies, they were able to find legitimate reasons for them; in the majority of cases, the test score was simply regarded as being 'wrong'.

Knowledge of students' scores on standardized tests rarely causes teachers to revise their assessments of students. In the majority of cases, there is no room for revision since test scores will closely agree with teachers' assessments (1, 5, 10). In those cases where a discrepancy does exist between a test score and a teacher's assessment, teachers are more likely to adhere to their own assessments than to accept the results of the test (6, 9, 10, 12). If teachers rarely alter their opinions of pupils in the light of test information, then an important question for research is 'why not?'.

A major limitation of much of the work on the impact of test information on teachers is that it contains an implicit assumption that teachers' reactions to discrepant test information can be encompassed within a change/no change dichotomy. The belief appears to be that the response

* Requests for off-prints should be sent to Peter Archer, Educational Research Centre, St Patrick's College, Dublin 9.

of teachers to test scores will be one of two types Either they will stick to their guns and not change their opinions or their opinions will shift in the direction suggested by the test There is, however, good reason to believe that this is an unduly simplistic view of the situation and that a variety of other reactions might be anticipated (cf 11 for discussion of responses to persuasive communications) For example, teachers might react negatively to the intrusion of tests into their domain and respond by adjustmg their opinions away from the test score Or teachers might over compensate and adjust their opmions too far in the direction of the test Or the test might generate a generalized uncertainty which could result in changes of opinion in all directions Indeed, it can be argued that the effect of providing test information to teachers might be to make teachers' opmions more stable Since in the majority of cases the results of a standardized test will probably confirm a teacher's opinion, it might be that exposure to such confirmatory evidence would result in the teacher having more confidence in his or her own opinions and so be reluctant to change The variety of ways in which a teacher might react to standard ized test information indicates that this is a complex issue and that simply attempting to establish whether test information causes teachers to alter their assessments in line with test results will present, at best, a limited picture

The objective of the study reported in this paper was to examine teachers' reactions to standardized tests in the broader context discussed above This objective was pursued by interviewing a small sample of teachers who had been provided with test information about their reaction to that information, especially their reaction to cases where the information differed from the teacher's previously expressed assessment

METHOD

Sample

The sample consisted of 13 second standard teachers and 17 fifth standard teachers who, as part of a larger study of the consequences of educational testing (10), had received the results of a standardized group ability test which had previously been administered to their students

Procedure

Early in the school year (between November 1st and December 15th) teachers rated the intelligence and other characteristics of their students At about the same time tests of ability were administered to the students

of all the participating teachers. Achievement tests were also administered in the classes of five second standard teachers and ten fifth standard teachers. The tests were returned to the headquarters of the research project for scoring and teachers were later provided with information on the test performance of their students. Student performance was described in terms of raw score, standard score, and percentile rank based on norms derived from standardizations of the test.

A few months after the test information had been provided, the interviews which form the basis of the present study were carried out. The interviews were usually conducted in the classroom of the teacher. Although all teachers had received the results of their students' test performance some time before the interview, a copy of these results was brought by the interviewer to all interviews.

Instruments

(i) *Ability tests.* The Otis-Lennon Mental Ability Test, Form J (8) was administered to second-standard students. This test is an adaptation of an American test and contains items designed to measure classification ability, the ability to follow directions, comprehension of verbal concepts, quantitative reasoning, and reasoning by analogy. Norms for standard and time of year are available for the Irish version. The Drumcondra Verbal Reasoning Test (DVRT) (4) was administered to fifth-standard students. This test was standardized on a national sample of Irish school children aged from ten years to twelve years eleven months. It contains items designed to measure the ability to use and reason with verbal symbols and includes sections on analogies, the identification of word opposites, and problems in classification and in inductive and deductive reasoning. The DVRT is age-normed.

(ii) *Teachers' ratings of intelligence.* A Pupil Evaluation Form was completed by teachers. On the form, teachers were asked to provide on five-point rating scales judgments of a variety of academic and non-academic student characteristics. The intelligence rating, which is used in the present study was obtained in response to the question: 'How do you rate this pupil's general intelligence, well below average, below average, average, above average, or well above average?'

(iii) *Teachers' reactions to test results.* In an interview, teachers were asked to select from a copy of their students' test scores any ability-test result which surprised them. For each student mentioned, the teachers

were asked to indicate their present estimate of the student's intelligence, what their opinion of the student had been prior to receiving the results, and how well they thought the student had done on the test All these estimates were requested in terms of five categories well below average, below average, average, above average, and well above average For each student mentioned, teachers were also asked why they thought the discrepancy had arisen and if they had altered their opinion of the student's intelligence and, if so, whether the test score had influenced the change of opinion

It may be noted that there is a possible difficulty with terminology here insofar as teachers were asked both during the interview and on the Pupil Evaluation Form to assess their students' 'intelligence' while the two tests used purported to measure 'ability' Attention was not drawn to this discrepancy by the interviewer and no interviewee raised it as a problem It seems reasonable, therefore, to assume that in this context, at least, teachers in the present study had no difficulty using the terms intelligence and ability interchangeably

Analysis

As well as obtaining teachers' perceptions of discrepancies, statistical discrepancies were also identified on the basis of a comparison of each student's test score with the teacher's initial rating (Pupil Evaluation Form) This was achieved by means of a linear regression of the ability test score on the intelligence rating with the data from each teacher in the interview sample The procedure was as follows For each teacher, regression was used to predict the ability test score from the intelligence rating and the residual (predicted test score minus the actual test score) was computed for each pupil A discrepancy was considered to exist whenever the residual was greater than both the standard error of estimate and the beta weight derived from the regression equation For a two-variable regression, the beta weight is equal to the difference between two adjacent predicted scores, in the present case, therefore, the beta weight may be seen as being the width of the rating categories

Teachers' responses to the open-ended questions about discrepancies were coded in four stages

Stage 1 Each response was examined to determine whether the teacher regarded the discrepancy as having arisen from a misperception on his or her part or from an error in the test Initially categories called 'teacher error', 'test error', and 'don't know' were established Later it was

necessary to set up two additional categories to cater for the situation where the teacher was not certain about why the discrepancy occurred but 'tended towards' either the teacher-error or test-error categories. It was also necessary to include a category for uncodable responses.

Stage 2. Teachers' explanations were examined to see if the teachers could specify the source of their own or the test's error. If, for example, a teacher regarded the test score as being wrong, he or she could attribute this error to some pupil attribute or behaviour at the time of the test – nervousness, illness, guessing, or cheating. Alternatively, a teacher could attribute the perceived error to some aspect of the testing process, for example, by claiming that the pupil had entered his or her age incorrectly or that an error in scoring had occurred. Teachers could also state that they regarded the test as an invalid measure of intelligence. Where the teachers regard the discrepancy as arising from their own error, their responses could be further analysed on the basis of whether or not an 'excuse' was presented for the error. For example, teachers might claim that they were misled by some pupil behaviour or by the report of a previous teacher.

Stage 3. The teacher's perception of the outcome of a discrepancy between the test score and the teacher rating was classified. The categories in this stage were- 'change', 'no change', and an intermediate category where the teacher, although not altering his or her opinion, reported that he or she questioned the initial assessment. The 'change' category can be divided into those cases where the teacher attributed the change of opinion to the influence of the test score and those cases where the teacher did not.

Stage 4. The teacher's first rating of intelligence (i.e., that given on the Pupil Evaluation Form) was compared with his or her more recent rating of the students' intelligence (i.e., that given in response to the interview question about the teacher's present opinion of the student). Three outcomes were possible: no change, change in the direction suggested by the test score, and change in the direction opposite to that suggested by the test score.

RESULTS

Test scores found surprising by teachers

In response to the question, 'which intelligence- test scores did you find surprising?', the 30 teachers in the interview sample mentioned 167 pupils. This represents 17.89% of the participating teachers' students for whom test scores were available. Of the 167 'surprises', 112 were cases where

the student had performed worse on the test than the teacher had expected and 55 were cases where the student had performed better than expected

Using the procedure described above for identifying statistical discrepancies for each of the 30 teachers in the sample, 220 discrepancies between ratings and test scores were identified Thus the number of statistical discrepancies is somewhat larger than the number of discrepancies identified by teachers

Table 1 represents a first stage in the examination of overlap between statistical and perceived discrepancies It contains the four possible combinations of perceived and statistical discrepancies Using a modification of the procedure suggested by Snedecor & Cochran (13) for setting the confidence limits of a sample proportion, the level of agreement between the two kinds of discrepancy was found to be statistically significant ($p < .01$) However, it is worth noting that there are a large number of statistical discrepancies which were not perceived as such by teachers and that a slightly smaller number of the perceived discrepancies were not discrepant using statistical criteria

TABLE 1

CROSSTABULATION OF PERCEIVED AND STATISTICAL DISCREPANCIES

	Perceived Discrepancy		
	Discrepant	Not Discrepant	Total
Statistical Discrepancy			
Discrepant	60	160	220
Not Discrepant	107	606	713
Total	167	766	933

Initially, the explanations offered by teachers for surprising test results were placed in one of the six categories described in Stage 1 of the analysis of open-ended interview questions. In the case of half of the discrepancies, the teacher regarded the test as being wrong. In about 15% of the cases, the teacher felt that he or she had misperceived the student. Teachers were unable to offer any explanation for a similar percentage of discrepancies. In less than 10% of cases, the teachers were unsure of the reason but 'hypothesized' that the test score was wrong, while, in a smaller number of cases (4%), the teachers were unsure but hypothesized that they themselves were wrong. The remaining 7% of responses could not be coded.

TABLE 2

NUMBERS (AND PERCENTAGES) OF EXPLANATIONS OFFERED BY TEACHERS FOR SURPRISING TEST RESULTS, BY TYPE OF EXPLANATION

Type of Explanation	No. of explanations	% of explanations
Test Error	79	50.97
Teacher Error	23	14.84
Don't Know	23	14.84
Hypothesized Test Error	13	8.39
Hypothesized Teacher Error	6	3.87
Uncodable	11	7.10
Total	155	100.01

Stage 2 of the coding procedure involved exploring teachers' explanations of discrepancies further in an attempt to specify the source of errors either on the part of the teacher or the test. The analysis of the teachers' attempts to specify the source of what they regarded as test errors did not yield a single instance where the discrepancy was explained by reference to any general fault in the test, such as lack of validity. In fact, when the 79 test-error responses are further analyzed, we find that in

two cases the teacher was unable to specify the source of the test's error Three discrepancies were attributed to factors relating to the testing process in one case, the teacher believed that the pupil's age had been computed incorrectly and in the other two cases the teacher mentioned 'random error' The source of the remaining 74 test errors was seen as relating to the physical and mental state of the student at the time of testing or to his or her behaviour during the test The most frequently occurring responses when the test score was lower than the teacher expected were nervousness, giddiness, and overconscientiousness and, when the test score was higher than expected, cheating and guessing

An attempt to specify the source of the teacher's error reveals a pattern rather similar to that found for the test-error cases Of the 23 cases in this category, there was one response where it was not possible to establish what the teacher saw as the source of the error In the remaining 22 cases, the source of the error was attributed to aspects of the pupil's classroom behaviour which had misled the teacher in forming his or her initial opinion Typically, the behaviour in question was the degree of the students' participation in the classroom (asking and answering questions, etc), where the level of participation was low, teachers said they had been led to underestimate students' ability and, where the level of participation was high, teachers had been led to the opposite conclusion

The final set of results concern the outcome, if any, which the discrepant test information was seen by teachers as bringing about Both reported and actual outcomes will be considered (see the descriptions of Stages 3 and 4 of the coding procedure) In relation to reported outcome, there were only 28 cases where the teacher reported a change of opinion and four cases where teachers said they were in the process of questioning their opinions In the remaining 123 cases, the teachers felt that the test had had no impact It is somewhat surprising that there were so few reports of teachers questioning their opinions in the light of the test information Many of the explanations offered by teachers for discrepancies between their assessments and test scores seemed to imply that the tests, at least initially, did create in the teachers some doubts about the validity of their assessments of students One must assume that any such doubts had been dispelled by the time the interviews took place

Actual outcome (i e , change of opinion) was assessed by comparing teachers' current rating of the student's intelligence given during the interview with the rating given on the Pupil Evaluation Form Where

these values were different, it was counted as a change. There were 60 such cases. Of these, 37 were in the direction of the test score and 23 were not. Of the 37 in the direction of the test score, 16 were acknowledged by the teacher as changing to conform with the test score and of these 16, 12 were attributed to the influence of the test score. Of the 23 changes which did not move in the direction implied by the test score, only four were acknowledged by teachers. Most of the changes (10 out of 12) which were attributed by teachers to the test's influence were changes in an upward direction. It would appear that teachers are more likely to alter their opinions of pupils and to attribute these changes to the test when it involves seeing the pupil in a better light. This finding contrasts with our earlier finding that the majority of perceived discrepancies were cases where the student was said to have done worse on the test than the teacher had expected.

Finally, it should be noted that there were eight cases where teachers reported that a change of opinion had taken place but where a comparison of the current and previous ratings revealed that no such change had taken place.

DISCUSSION

The first finding reported in the present paper is the low level of correspondence between perceived and statistical discrepancies (see Table 1). One partial explanation for this is the fact that a majority of the perceived discrepancies (67%) were cases where the test score was reported by the teacher as being less than expected, whereas the actual discrepancies were more or less equally divided between cases where the test score was better than one would have expected on the basis of the rating and cases where the test score was worse than expected. It should be noted that the linear regression method will tend usually to produce this kind of situation since the sum of the residuals will always be equal to zero.

Whatever the reason for the absence of a marked overlap between perceived and statistical discrepancies, this finding has two important implications for some of the issues raised elsewhere in the paper. Firstly, the fact that so few of the statistical discrepancies were perceived as discrepancies by teachers goes a long way towards explaining why teachers who received test information did not often align their judgments with the test information. Secondly, it is important in relation to our analysis of teachers' explanations of discrepancies (presented below) to keep in

mind that the discrepancies in question bore only a slight resemblance to our set of statistical discrepancies

A marked tendency among teachers to regard discrepancies as resulting primarily from errors in the test scores was noted This would appear to lend some support to Jackson's view that when contradictions between test scores and teacher judgment occur, the teacher seems more likely to deny the accuracy of the test information than to alter her previous assess ment of the student (9, p 124) However, further analysis of teachers' explanations of discrepancies indicate that the situation is more complex than Jackson's statement might suggest It seems that rather than looking for general weaknesses in a test, teachers are prompted by the existence of a discrepancy to look for more specific excuses for what they perceive as incorrect test scores The fact that the teacher succeeds in finding such an 'excuse' in the vast majority of attempts (93 67%) may be taken as evidence of teachers' reluctance to denigrate the source of discrepant information (i e , the test) In this respect, the findings of the present study may be said to be in line with those of other studies which shows that, in general, teachers see standardized tests as yielding accurate information on the ability and achievement of pupils (e g , 3, 10) It would seem from the present study that exposure to discrepant test information does not prompt teachers to question the validity of the test but to offer reasons or excuses for each individual discrepancy

It should be pointed out that in talking about 'excuses' we do not mean to imply that the explanations offered by teachers are not legitimate It is quite likely, in fact, that many of the discrepant cases selected by teachers do in fact represent instances where the test score does not reflect the pupil's 'true' ability and that the explanations given by the teachers are perfectly valid

In addition to the obvious preference of teachers to see discrepancies as resulting from errors in the test rather than from errors in their own assessments, another striking feature of the analysis of teachers' explana tions is the large extent to which errors, both on the part of the teacher and on the part of the test, are attributed to student variables The physical and mental condition of the pupil at the time of testing was mentioned most often in cases where the test score was seen as wrong, while classroom behaviour, which might have misled the teacher, was the most frequently mentioned source of an error on the part of the teacher It would seem that teachers find themselves when confronted with

discrepant test information in a situation of conflict between two highly regarded sources of information – the two indices of intelligence. They cannot denigrate the tests without by implication denigrating their own assessments, since those assessments agree in general with the test information. Thus, they may try to resolve the conflict by shifting the responsibility for the discrepancy to a 'third party', the student.

Two features of the analyses of outcome of discrepancies presented here deserve further comment. Firstly, although the number of opinion changes attributed to the influence of test information is small, almost all represent upward revisions. A similar directionality was noted in other studies (10, 12). However, as was noted, this pattern is surprising in the present study given that approximately two-thirds of the discrepancies nominated by teachers were cases where the pupil's test score was lower than the teacher had expected. Secondly, the low level of agreement between reported and actual changes of opinion raises the possibility that, like other forms of persuasive communication, test scores may sometimes induce changes of opinion in teachers, of which the teachers themselves have no subjective awareness (2, 7).

In the introduction to this paper it was suggested that the available evidence indicates that teachers rarely align their opinions of pupils with the results of standardized tests. The present study may go some way towards explaining the relatively small impact of standardized-test information. Firstly, we saw that test scores which appeared to be discrepant with the teacher's opinion on the basis of statistical criteria were only perceived as discrepant slightly more often than chance alone would predict. This means that in a large proportion of the cases where one could have expected the test information to have had an effect, the teachers did not have any reason to revise their assessments, since they did not perceive the test results as differing from those assessments. Secondly, the study revealed that where discrepancies were perceived, the teachers were able to find legitimate reasons for them. Furthermore, since in the majority of these cases the test result was regarded as wrong, a change of opinion was not seen to be necessary.

REFERENCES

1. ARCHER, P. *A comparison of teacher judgements of pupils and the results of standardized tests.* Unpublished doctoral thesis. University College Cork, 1980

2. BEM, D.J., & McCONNELL, H.K. Testing the self-perception explanation of dissonance phenomena: On the salience of premanipulation attitudes. *Journal of*

Personality and Social Psychology 1970 *14*, 23 31

3 BRIM O G , GOSLIN, D A , GLASS, D C , & GOLDBREG, I The use of standardized ability tests in American secondary schools and their impact on students teachers and administrators Technical Report No 3 New York Russell Sage Foundation, 1965

4 EDUCATIONAL RESEARCH CENTRE Drumcondra Verbal Reasoning Test Dublin Author, 1968

5 EGAN O , & ARCHER P The accuracy of teachers ratings of ability A regression model *American Educational Research Journal*, 1985 *22*, 25 34

6 FLEMING, E S & ANTTONEN R G Teacher expectancy or My Fair Lady *American Educational Research Journal*, 1971 *8*, 241 252

7 GOETHALS, G R , & RECKMAN, R F The perception of consistency in attitudes *Journal of Experimental Social Psychology*, 1973 *9*, 491 501

8 GREANEY, V , & KELLAGHAN, T Otis-Lennon Mental Ability Test (Irish version) Elementary 1 Level Form J Dublin Educational Research Centre 1972

9 JACKSON, P *Life in classrooms* New York Holt, Rinehart & Winston, 1968

10 KELLAGHAN T MADAUS, G F , & AIRASIAN, P W *The effects of standardized testing* Boston Kluwer Nijhoff 1982

11 McGUIRE, W J Attitudes and attitude change In G Lindzey & E Aronson (Eds) *The handbook of social psychology (3rd ed) Vol 3* Reading MA Addison Wesley 1985

12 SALMON-COX L Teachers and standardized achievement tests What's really happening? *Phi Delta Kappan* 1981, *62* 631-634

13 SNEDECOR, C W , & COCHRAN, W G *Statistical methods* Amos, Iowa Iowa State University Press 1967