The Irish Journal of Education, 1989, xxiii 2, pp 40-50

RELATIVE BIAS IN TEACHER JUDGMENTS AND STANDARDIZED TESTS IN THE IDENTIFICATION OF LITERACY PROBLEMS

Thomas Kellaghan and Patricia Fontes
Educational Research Centre
St Patrick's College, Dublin

Sixth grade teachers in Irish schools were asked to identify pupils in their classes that they perceived as having literacy problems 11% of pupils were nominated Pupils performance on a standardized test of reading was also assessed and the lowest scoring 11% of pupils were identified One third of pupils were identified by both teacher and test (n 108) one third by the teacher only (n 102) and one third by the test only (n 107) Pupils in the three groups were compared in terms of age gender social class classroom behaviour and social behaviour In a MANOVA and canonical discriminant function analysis the main differences were found between the group of pupils identified by both test and teacher and the groups identified only by the test or only by the teacher pupils in the groups identified by both teacher and test were found to have relatively low achievement oriented behaviour Pupils identified only by the test tended to score relatively low in sociability Pupils identified only by teachers showed the least relative bias on the characteristics that were assessed

Teachers' assessments of pupils' scholastic ability and achievement show considerable agreement with the assessments which one would obtain if one used standardized tests In correlational terms, the relationship between the two forms of assessment is something of the order of 6 (Airasian, Kellaghan, Madaus, & Pedulla, 1977) When corrections are made for the unreliability of measures, it has been estimated that just under 20% of variance can be identified as not being common to teachers' estimates and test scores (Kellaghan, Madaus, & Airasian, 1982) While the strength of the relationship suggests that teachers and tests focus on the same phenomenon, there is sufficient disagreement between the results of the two types of assessment to indicate that teachers advert to factors which are disregarded by tests, while tests focus on factors which are not included in teachers' appraisals

Following the development of 'objective' measures of ability and achievement in this century, it seemed reasonable to assume that where tests and teachers disagreed, the teacher, since he or she had to rely on 'subjective'

evidence, was the one that was in error (see Binet, 1911, Terman, 1919) On this assumption, a number of investigators set out to examine teachers' bias in the assessment of their pupils, often using an objective test of intelligence or achievement as their criterion measure against which to evaluate teachers' judgments The evidence from these studies, while not always consistent, suggests that teachers' assessments are influenced by a range of factors including pupils' gender (Carter, 1952, Doyle, Hancock, & Kifer, 1972, McCandless, Roberts, & Starnes, 1974), age (Terman, 1919, Thompson, 1936, Varner, 1923), physical appearance (Clifford & Walster, 1973, Seligman, Tucker, & Lambert, 1972, Varner, 1923), social behaviour (leadership qualities, participation in classroom activities, conformity) (Gordon & Thomas, 1967, Morrison, 1970, Rist, 1970), personality characteristics (originality, attention span, persistence) (Morrison, McIntyre, & Sutherland, 1965, Pedulla, Airasian, & Madaus, 1980, Varner, 1923), and social class (Rist, 1970, Frender, Brown, & Lambert, 1970) It was in response to evidence of this kind that commentators from Binet (1911) to Cronbach (1963) concluded that 'teachers have various biases, which enter into their impressionistic evaluations' (Cronbach, 1963, p 550) Objective measurement was seen as providing a corrective for these biases

There is also a tradition — one that has been put with increasing vigour in recent years — that objective tests may be misleading in the information they provide about pupils (see Jensen, 1980, Kellaghan, Madaus, & Airasian, 1980) This tradition is less firmly rooted in research than is the tradition on teacher bias While the most frequently voiced position is that tests are biased in terms of social class, race, and ethnic and cultural background (see Quinto, 1977, Samuda, 1977, Tyler & White 1979), some of the other factors which have been considered in the context of teacher bias could also play a role in test performance For example, although efforts may be made to remove sex-biased items in test construction (see Ironson & Subovlak 1979, Rudner, Getson, & Knight, 1980, Scheuneman, 1979) performance may be related to gender on some kinds of test (Stockard, 1980)

If we talk of bias, we are assuming a criterion against which it may be measured A criterion raises the question of validity in one or other of its forms However, without taking up the questions of absolute criteria and validity, or without making assumptions about the relative 'validity' of tests or of teachers' assessments, we can ask in what way tests and teachers agree and differ in their assessment of pupils' scholastic achievement For example, to what extent do the assessments of tests and of teachers covary with particular characteristics of pupils?

In an attempt to answer such questions we selected literacy as an aspect of achievement which is undoubtedly important and might also appear relatively concrete If one agrees on a definition of literacy, then it is possible to see to what extent teachers and test performance coincide in their designation of pupils as literate or as lacking literacy Further, where there is disagreement, we can examine the characteristics of pupils who are designated as lacking literacy by teachers, but not by tests, and the characteristics of pupils designated by tests, but not by teachers To the extent that either tests or teachers uniquely tend to nominate pupils with certain characteristics, we have identified a relative bias between the two

Literacy, of course, is not readily defined This is partly because the skills involved are relative to the demands which individuals have to meet and may change over time Our approach in this study was a relatively simple one, focusing on the reading and writing abilities of pupils at the stage when they are in their last year in elementary school (at about age 12 years) (see Fontes & Kellaghan, 1977) We asked sixth-grade teachers to nominate pupils they perceived as having literacy problems Pupils' performance on a standardized-reading test was also assessed Thus we have two separate sources of information — teachers' judgment and test performance — which we can relate to the criterion of literacy We also obtained information on a range of other variables, our consideration of possible sources of bias suggested that gender, age, social class, and personal characteristics of the pupils might all be relevant We use multivariate analysis of variance (MANOVA) to determine whether or not differences exist between pupils who were identified as having literacy problems by both teacher and test, pupils who were identified by the test only, and pupils who were identified by the teacher only Canonical discriminant-function analysis was used to specify differences in the characteristics of pupils forming the three groups

METHOD

Sample

A representative sample of 105 schools was drawn from the population of primary schools (excluding private, Protestant, special, and one-teacher schools) in the Republic of Ireland Of these 97 agreed to participate in the study However, only 63 of the schools returned complete data Information collected was used in analyses only if all sixth-grade teachers within a school completed all questionnaires and carried out testing Altogether, 83 teachers met these requirements

Instruments

Literacy questionnaire. The literacy questionnaire was a document in which four lists of pupils' names were sought from teachers by directing them as follows: (i) Please name the pupils in your class who, in your opinion, if they were to leave school now, would *not* be able to cope with the *everyday demands of our society* in (a) reading (e.g., reading notices, official forms, newspapers); (b) writing (e.g., writing letters, applications for jobs). (ii) Please name the pupils in your class who, in your opinion would *not* be able to cope with the *demands of education in a post-primary school* in (a) reading (e.g., reading text-books); (b) writing (e.g., writing essays).

Standardized tests. The standardized-reading test, the Drumcondra English Test, Level III, Form A (Educational Research Centre, 1976), is a group multiple-choice test which yields a total reading score made up of sub-test scores for reading vocabulary and reading comprehension.

Ratings of personal characteristics of pupils. Ratings of pupils were obtained on a Pupil Evaluation Form completed by teachers. Each pupil was rated on a 5-point scale (5=very good, 4=good, 3=average, 2=fair, 1=poor) for the following 12 personal-social characteristics: participation in class, behaviour in school, personal appearance and dress, attention span/concentration, persistence in school work, keenness to get on, speech/use of language, neatness in school work, manners/politeness, getting along with other children, working with limited supervision, and attendance. Teachers were also asked to state the occupation of each pupil's father or guardian, giving sufficient detail to enable classification of occupational status to be made.

Procedure

Administration of instruments. The standardized-achievement test was administered to pupils by their own teachers during the first three months of the school year. Around the same time, and before the results of the test were available to teachers, each teacher was asked to complete the Pupil Evaluation Form for each pupil in his/her class. The literacy questionnaire was administered to teachers towards the end of the school year by a field worker; the questionnaire was completed in the presence of the field worker who was available to give assistance in interpretation.

Classification of pupils identified by teachers and by test. Of the 2,762 pupils in the classes from which judgments were obtained, 303 (11%) were rated by their teachers as having at least one of the problems with literacy described in the literacy questionnaire. Since the standardized-reading test does not have an obvious cut-off score which could be used to define lack of literacy skills, a

cut-off score of 35 was selected since the same proportion (11%) of pupils fell below this score as the proportion assigned to the reading difficulty category by teachers There were 274 children with a score below 35 The final step in the classification of pupils was to determine how many had been identified by both their teacher and the test, by their teacher only, and by the test only, 108 pupils were identified by both, 102 by their teacher only, and 107 by the test only

Measures of possible distinguishing characteristics The age (in months) and the gender (1=boy, 2=girl) of each pupil in each of the three groups was identified The pupil's socioeconomic status was recorded on the basis of parental occupation, using four dummy variables — high SES (professional/ managerial and white collar workers), middle SES (skilled workers and farmers with 50 or more acres), low SES (unskilled workers and farmers with fewer than 50 acres), and unknown SES (including unemployed) A score of 1 indicated pupil membership in a category and a 0 non-membership

A factor analysis of the teacher ratings of pupils' personal-social characteristics identified two factors (Airasian, Kellaghan, & Madaus, 1977) The first was termed a classroom-behaviour factor and included such items as participation in class, persistence in school work, keenness to get on, and neatness in school work The second was called a social-behaviour factor, characteristics which loaded highly on it were behaviour in school, personal appearance and dress, manners/politeness and getting along with other children In our analyses, pupils were assigned a classroom behaviour score and a social-behaviour score, each score being the sum of the teacher ratings for the characteristics loading on each factor

Analysis

A multivariate analysis of variance (MANOVA) was carried out on eight variables (see Table 1) to test the hypothesis of an overall difference between the group identified by the teacher only, the group identified by the test only, and the group identified by both teacher and test Subsequent to the MANOVA and contingent upon a significant difference between groups being found, univariate analyses of variance (ANOVA) were carried out for each variable If these analyses yielded significant F-values Scheffe post-hoc analyses were carried out to examine the significance of differences between pairs of groups Subject to the same conditions as the ANOVAs, a discriminant-function analysis, using a stepwise-inclusion method, was carried out to determine what weighting of the available variables served to discriminate most clearly among the three groups Finally, the differences between the pairs of groups on any

significant weighted combination(s) obtained were themselves tested for significance, using Scheffé contrasts

RESULTS

Means and standard deviations of the variables for each group are presented in Table 1 The *lambda* value was 883, for which $p < 001$ ($\chi^2 = 56\ 55$, $df = 16$), indicating overall significant differences between the groups ANOVAs indicated the presence of significant differences for pupils' age, gender, classroom behaviour, and social behaviour Scheffe contrasts revealed that differences in pupil age and classroom behaviour existed between the groups identified by both the test and the teacher on the one hand and the group identified only by the test and the group identified only by the teacher on the other; pupils identified by both techniques were older and received lower classroom-behaviour ratings Social-behaviour ratings of the group identified by both techniques were also significantly lower than those of the group identified only by teachers Finally a significant contrast was found between the group identified only by the teacher and the group identified only by the test, the former included a higher proportion of girls, the latter a higher proportion of boys

TABLE 1

STUDENT CHARACTERISTICS MEANS AND STANDARD DEVIATIONS OF GROUPS IDENTIFIED BY BOTH TEACHER AND TEST, BY TEACHER ONLY AND BY TEST ONLY

	Teacher & Test (N=108)		Teacher Only (N=102)		Test Only (N=107)		*F*	*p*
	M	SD	M	SD	M	SD		
Age (months)	149 69	9 21	145 29	7 87	145 58	9 00	8 45	001
Gender	1 49	50	1 58	50	1 36	48	5 46	01
High SES	05	21	14	35	10	31	2 61	NS
Middle SES	25	44	34	48	26	44	1 31	NS
Low SES	31	46	28	45	35	48	47	NS
Unknown SES	34	48	22	41	26	44	2 20	NS
Classroom behaviour	13 75	5 65	16 83	6 36	17 36	6 25	10 95	001
Social behaviour	16 77	4 78	18 72	4 33	17 77	4 56	4 78	01

The canonical discriminant function selected seven variables for which the *lambda* value of the MANOVA was 839 ($\chi^2 = 54\ 76$, $df = 14$, $p < 001$) Two significant functions, which accounted for 15 6% of the total variance between

the groups, were derived from the seven variables The functions to a considerable extent reflect the distinction between teachers' ratings of classroom and social behaviour, with the incorporation of additional variables related to gender and SES (see Table 2) The first function, to which 61 77% of the explained variance is attributable, may be regarded as describing achievement-related behaviour Classroom behaviour (attention span, persistence), SES (high), and gender (girls scoring higher than boys) have positive weightings on the function and age has a negative weighting The second function, accounting for 38 23% of explained variance, has a strong sociability component, though classroom behaviour also figures on it High scores are associated with gender (girls scoring higher than boys) and social behaviour (manners/politeness, getting along with other children, personal appearance and dress) while low scores are associated with low SES and aspects of classroom behaviour

On the first function, the centroid for the group identified by both teacher and test differed significantly from the centroids of the two other groups On the second function, the group identified only by the test differed significantly from the centroids of the other groups

Prediction of group membership on the basis of function scores yielded 57 4% accuracy for the group identified by both teacher and test, 45 1% for the group identified only by the teacher, and 49 5% for the group identified only by the test

TABLE 2

DISCRIMINANT FUNCTION ANALYSIS COEFFICIENTS AND CENTROIDS

Variables	Function I	Function II	Variables
Classroom behaviour	55	75	Gender
High SES	29	60	Social behaviour
Gender	29	05	High SES
Low SES	11	15	Age
Social behaviour	05	44	Low SES
Unknown SES	15	51	Unknown SES
Age	57	77	Classroom behaviour
Group centroids			
Identified by teacher & test	46	07	
Identified by teacher only	31	29	
Identified by test only	16	35	

DISCUSSION

If a teacher's judgment and a standardized test are used to identify pupils with literacy problems then approximately one-third of pupils will be identified by both teacher and test, a further third only by the teacher, and the final third only by the test Hence the method used results in the identification of different sets of pupils

Further, the characteristics of pupils identified by one method differed from those identified by another The main differences occur between the group identified by both methods and the groups identified by only the teacher or the test A pupil identified by both methods by comparison with one identified by only one method is likely to have relatively low achievement-oriented behaviour The results of our ANOVAs indicate that such a pupil receives poor classroom-behaviour ratings and is relatively old (older children in a class are likely to have experienced grade retention because of low achievement) The discriminant-function analysis suggests that the pupil is also likely to come from a low SES background and to be a boy

The discriminant analysis also indicates that, in their assessments, tests differ from teachers and from the combined assessment of test and teacher in their tendency to identify pupils that are low in sociability (sociability being defined in terms of the social behaviour of pupils), gender, and SES This interpretation is supported in the ANOVAs for social behaviour and gender, tests, compared to teachers, identified more boys than girls, while both tests individually and in combination with teachers (compared to the unique identifications of teachers) were more likely to identify pupils rated low in social behaviour

These findings are somewhat surprising In the case of gender, they run counter to the findings of studies in the United States (e g , Carter, 1952) and in Britain (Morrison, McIntyre, & Sutherland, 1965) that teachers accord higher grades to girls than to boys even when no differences are detectable on a standardized test An earlier Irish study revealed no differences in teachers' assessments of the general scholastic progress of boys and girls (Kellaghan, Macnamara, & Neuman, 1969) It may be that teachers behave differently when assessing literacy problems than when assessing general scholastic progress

Teachers' relative lack of bias regarding pupils' social behaviour (e g , personal appearance and dress, manners/politeness, ability to get along with other children) is also surprising in the light of findings of earlier studies (e g , (Morrison, 1970, Rist, 1970) and the social nature of the interaction between teachers and pupils Our findings, of course do not indicate that teachers judgments are not influenced by aspects of a pupil's background, what they do

indicate is that teachers' judgments are less related to such factors than is performance on a standardized test It may be that teachers make positive efforts not to be affected by what they might regard as peripheral factors in making judgments about pupils' levels of literacy

Overall, our findings indicate that some degree of bias in terms of the variables we considered (pupil's age, gender, SES, classroom behaviour, and social behaviour) is involved whether tests or teacher judgments are used to identify pupils with literacy problems When literacy is identified by both teacher and test, there is a tendency to identify pupils who are low in terms of general achievement-oriented behaviour and related characteristics Tests, by contrast with other methods, tend to identify pupils low in sociability and related characteristics Teachers exhibit the least bias Obviously, they have more information available to them in making a judgment about pupil literacy than do tests and they can behave in a more 'intelligent' way in making their decisions Our findings suggest that they may compensate for some of the factors (e g , social behaviour) which are sometimes considered to play a role in the 'labelling' of pupils (see Archer & Martin, 1980)

While our findings must be taken as evidence that different approaches to the identification of literacy indicate some biases in that the characteristics of pupils identified in different ways are distinguishable, the magnitude of these relative biases should not be overestimated Our discriminant function analysis accounted for less than 16% of the total variance between groups and a prediction of group membership on the basis of the characteristics we examined would be correct for only about 50% of cases

REFERENCES

Airasian, P W , Kellaghan, T , Madaus, G F , & Pedulla, J J (1977) Proportion and direction of teacher rating changes of pupils' progress attributable to standardized test information *Journal of Educational Psychology, 69,* 702-709

Archer, P , & Martin, M (1980) Teachers' ratings of reading attainment as a function of pupils' socio-economic status *Irish Journal of Education, 14,* 33-42

Binet, A (1911) Nouvelles reserches sur la mesure du niveau intellectuel chez les enfant d'école *Annee Psychologique 17* 145-201

Carter, R (1952) How invalid are marks assigned by teachers? *Journal of Educational Psychology, 43,* 218-228

Clifford, M , & Walster, E (1973) The effect of physical attractiveness on teacher expectation *Sociology of Education 46,* 248-258

Cronbach, L J (1963) *Educational psychology* (2nd ed) London Rupert Hart-Davis

Doyle, W , Hancock, G , & Kifer, E (1972) Teachers' perceptions Do they make a difference? *Journal of the Association for the Study of Perception, 7,* 21-30

Educational Research Centre (1976) Drumcondra English Test, Level III, Form A Dublin Educational Research Centre, St Patrick's College

Fontes, P J , & Kellaghan, T (1977) Incidence and correlates of illiteracy in Irish primary schools *Irish Journal of Education, 11,* 5-20

Frender, R , Brown, B , & Lambert, W E (1970) The role of speech characteristics in scholastic success *Canadian Journal of Behavioral Science, 2,* 299-306

Gordon, E M , & Thomas, A (1967) Children's behavioral style and the teacher's appraisal of their intelligence *Journal of School Psychology, 5* 292-300

Ironson, G H , & Subkoviak, M J (1979) A comparison of several methods of assessing item bias *Journal of Educational Measurement 16* 209-225

Jensen, A R (1980) *Bias in mental testing* New York Free Press

Kellaghan, T , Macnamara, J , & Neuman, E (1969) Teachers' assessments of the scholastic progress of pupils *Irish Journal of Education 3* 95-104

Kellaghan, T , Madaus, G F , & Airasian, P W (1980) *Standardized testing in elementary schools Effects on schools, teachers, and pupils* Washington DC National Institute of Education,US Department of Health, Education, & Welfare

Kellaghan, T , Madaus, G F , & Airasian, P W (1982) *The effects of standardized testing* Boston Kluwer-Nijhoff

McCandless, B R , Roberts, A, & Starnes, T (1974) Teachers' marks, achievement test scores, and aptitude relations with respect to social class, race and sex *Journal of Educational Psychology, 63* 153-159

Morrison, A (1970) Some aspects of assessment in the classroom *Scottish Educational Studies, 2,* 95-101

Morrison, A , McIntyre, D , & Sutherland, J (1965) Teachers' personality assessments of primary school pupils *British Journal of Educational Psychology, 35,* 306-319

Pedulla, J J , Airasian, P W , & Madaus, G F (1980) Do teacher ratings and standardized test results of pupils yield the same information? *American Educational Research Journal 17,* 303-307

Quinto, F (1977) Why standardized tests fail the accountability test In R M Bossone & M Weiner (Eds), *Proceedings from the National Conference on Testing Major issues* Center for Advanced Study in Education, Graduate School and University Center of the City University of New York

Rist, R C (1970) Student social class and teacher expectations The self-fulfilling prophecy in ghetto education *Harvard Educational Review 40,* 411-451

Rudner, L M , Getson, P R , & Knight, D L (1980) A Monte Carlo comparison of seven biased item detection techniques *Journal of Educational Measurement, 17* 1-10

Samuda, R J (1977) Critical concerns for the testing of minorities Time for new initiatives In R M Bossone & M Weiner (Eds), *Proceedings from the National Conference on Testing Major issues* Center for Advanced Study in Education, Graduate School and University Center of the City University of New York

Scheuneman, J (1979) A method of assessing bias in test items *Journal of Educational Measurement 16* 143-152

Seligman, C R , Tucker, G R , & Lambert, W E (1972) The effects of speech style and other attributes on teachers' attitudes towards pupils *Language in Society, 1* 131-142

Stockard, J (1980) Sex inequities in the experiences of pupils In J Stockard et al, *Sex equity in education* New York Academic Press

Terman, L M (1919) *The measurement of intelligence* London Harrap

Thompson, G (1936) The value of intelligence tests in an examination for selecting pupils for secondary education *British Journal of Educational Psychology, 6,* 174-179

Tyler, R W , & White, S H (1979) *Testing teaching and learning* Washington, DC National Institute of Education, US Department of Health, Education, & Welfare

Varner, G R (1923) Improvement in rating the intelligence of pupils *Journal of Educational Research 8* 220-232