Pergamon

PII: S0346-251X(96)00040-1

System, Vol. 24, No. 4, pp. 437–448, 1996

0346-251X/96 $15.00 + 0.00

I DIDN'T GET THE GRADE I NEED. WHERE'S MY SOLICITOR?

GLENN FULCHER* and RON BAMFORD†

**English Language Institute and †Surrey European Management School, University of Surrey, Guildford, Surrey, GU2 5XH, U.K.*

This article looks at standards in language testing, in the context of the legal framework of the United States and the United Kingdom. We argue that testing bodies in the U.S. conduct appropriate research into reliability and validity partly[1] because of a legal requirement to ensure that all tests meet certain technical standards. These grew out of litigation involving employment testing, and now apply to all psychological and educational tests too. In the United Kingdom litigation still only relates to employment issues. However, principles have been laid by which successful challenges to educational assessment could be mounted. We look at these principles, and highlight areas in which examination boards offering EFL tests may be challenged in the courts. Alderson and Buck (1993) argued that British examination boards would only begin to meet appropriate technical standards once EFL testing was integrated with mainstream school testing. We agree with the analysis of the situation provided by Alderson and Buck, but conclude that this end can only be achieved once examination boards have been taken to court on the grounds that their tests are unreliable, invalid, or biased. Copyright © 1996 Elsevier Science Ltd

INTRODUCTION

Many U.K. based examining boards offer tests in EFL/ESL to international students (see the Appendix), and some of these are "high stakes" tests in that entry to a college or university course may depend upon the results. Companies and organisations also frequently use external tests, provided by examining boards, as screening instruments for appointments or promotions. The question which we pose in this paper is: if a candidate for a British examination were to get a grade which was lower than required, which resulted in his/her being denied access to higher education, a job position or promotion, on what grounds could he/she take the examining board to court? As such, this paper is particularly concerned with testing English as a foreign language as practiced in the U.K., although testing related legal issues in the U.S. is discussed. There is no attempt to provide a comprehensive description of litigation on either side of the Atlantic. Rather cases have been selected for discussion on the basis of how well they represent the type of concerns which are relevant to language testing.

BACKGROUND

Standards in testing and assessment has been an important issue since the 1960s, but it has only been recently that language testing bodies in the U.K. have begun to respond to the need to ensure that certain standards are maintained. What do "standards" mean? Stansfield (1993: p. 190) says:

> "*Standards* is used to refer to the procedures language testers follow in developing tests, operating test programs, or administering and interpreting tests and test scores."

It would be fair to say that the notion of *standards* is considerably more developed in the United States than in the U.K. (see Alderson *et al.* 1995, for a description of the failings of Examination Boards in the U.K.), and one reason for this is that testing bodies (Examination Boards) may be held legally accountable for the tests they develop, the programmes they run, and possibly any negative uses to which test scores are put. It is for this reason that standards such as the American Psychological Association's (APA, 1985) have been used to guide test developers in all fields of measurement design and score interpretation, and are increasingly being used in litigation.

Although the development of standards and codes of practice is becoming more common in the U.K., we feel that some U.K. Examining Boards will only be obliged to implement these standards, employ professional testing specialists, conduct reliability and validity studies, and publish test information, when there is a real threat of court action against them.

THE SITUATION IN THE UNITED STATES

Litigation regarding tests in the U.S. did not begin because of disillusionment with the tests themselves or the educational system, but as a way of overcoming the segregation of blacks and whites in the 1960s within a system which often used tests to allocate students to special programmes (Bersoff, 1979). The first case was that of Hobsen versus Hansen (1967, 1969) in which the state of Washington was charged with placing a disproportionate number of black children in lower tracks of the educational system based on a test which was biased against them (Reschly *et al.*, 1988a: pp. 16–17).

This case set the scene for many more, in which the key issue to develop was that of test bias, which may be defined as occurring:

> "....when groups of individuals with certain characteristics (e.g. male vs female, high vs low socioeconomic status) consistently obtain different scales on a specific instrument." (Taylor, 1991: p. 3).

Cases concerning perceived bias against ethnic groups have appealed to the 14th Amendment of the constitution of the United States, which protects the rights of minorities, and Title VII. The latter was the 1964 Civil Rights Act (Section 701 of the Act 42 U.S.C. 2000e) passed by Congress to address discriminatory employment practices (Hood and Parker, 1991). For issues in gender discrimination, see Childs (1990).

One of the most important cases of this kind was that of Larry P. versus Riles, first filed in 1971 and before the courts in 1972 (ERIC, 1985). The plaintiff charged that an IQ test used in San Francisco allocated a disproportionate number of blacks to programmes for the mildly mentally retarded. The result of the trial was that IQ tests were banned in the United States for the purpose of testing black students (Prasse and Reschly, 1986; Reschly *et al.*, 1988a, 1988b, 1988c; Reschly, 1990; Taylor, 1991; Macmillan and Barlow, 1991). In this case, and all cases where the plaintiffs were successful, the issue placed before the court was test bias. In cases where the defendants were successful, the issue was one of misclassification rather than bias. That is, there was no systematic bias in the test, but some error of measurement.

In a brief review of important cases which relate to issues of test bias, the areas in which litigation is likely to be successful will be highlighted.

In 1984, an out-of-court settlement in the Golden Rule Insurance Company versus Mathias, ensured that in the United States all test results must be reported and analysed by race, so that bias would be seen not to exist in test results (Hood and Parker, 1991: p. 611). This is in line with the APA Standard 3.10 (APA, 1985) which states that:

> "When previous research indicates the need for studies of item of test performance differences for a particular kind of test for members of age, ethnic, cultural and gender groups in the population of test takers, such studies should be conducted as soon as is feasible. Such research should be designed to detect and eliminate aspects of test design, content, or format that might bias test scores for particular groups."

However, it should be noted that as a result of cases such as Wards Cove versus Antonio (Hood and Parker, 1991) and Debra P. versus Turlington (Phillips, 1991) providing proof of statistical bias alone cannot ensure success for a plaintiff. In the former case, regarding employment, the court ruled that the plaintiffs would have to show that "specific practices by the state or the testing company caused the discrimination and had a specific impact on minorities (Hood and Parker, 1991: p. 604). In the latter case, it was ruled that the State Student Assessment Test (SSAT II), a test of functional literacy, could be used in the State of Florida, even though there existed statistical differences between the number of blacks and whites who passed the test, and therefore received a high school diploma. The view of the court was that ensuring equal pass rates was unequal treatment to whites, as black students would gain a diploma whether it had been earned or not, making the diploma worthless (Phillips, 1991: p. 195).

In both of these cases, won by defendants, the issues on which the cases were assessed, and considered to be of more importance than statistical bias, were:

Wards Cove versus Antonio

- *Predictive validity:* evidence of the relationship of test scores to future success.
- *Content validity:* the presence of a set of tasks which adequately represents the content domain of the subject.

Debra P. versus Turlington

- *Introduction of tests/new syllabuses:* there must be a significant delay between the publication of new tests/syllabuses to allow students to adequately prepare over a

number of years, in the new skills to be tested, as these are normally acquired over a period of time.

- *Curricular validity:* What is tested (skills, not test format) should be adequately represented in available texts, workbooks and other teaching materials.

This means that there should be equal opportunities, but not necessarily equal outcome for different groups. However, should the test be poorly developed or have technical flaws, a plaintiff would have a strong case. This is confirmed by discussions of other cases, in which potential grounds for litigation are presented.

In the Abermarle Paper versus Moody (1975) a test was declared invalid for a particular purpose because it was not designed to the standards laid down by American Educational Research Association (AERA) in the APA Guidelines. United States versus State of North Carolina (1975) and United States versus South Carolina (1977) established the principle that if pass scores are to be set for tests, then proper validation studies must be conducted to establish fairness (McDonough and Wolf, 1988). Kleinman and Faley (1985: p. 823) relate the issue of cut scores to the APA Guidelines Standard 6.9 (APA: p. 43), which states that:

> "When a specific cut score is used to select, classify, or certify test takers, the method and rationale for setting that cut score, including any technical analyses, should be presented in a manual or report. When cut scores are based PRIMARILY on professional judgement, the qualifications of the judges should also be documented."

Finally, in criterion-referenced tests, Kleinman and Faley show that in litigation it is important for defendants to be able to demonstrate that rating scales are not ambiguous, and that raters are properly trained in the use of rating scales. A failure in either of these areas would violate technical requirements of criterion-referenced tests as stated in the APA Guidelines (APA, 1985).

The following further possible challenges to tests and test scores may be listed:

- inadequate methods of establishing cut scores for decision making;
- arbitrary and capricious development or implementation of a test;
- tests which lack statistical and conceptual validity;
- failure to provide unambiguous rating scales in criterion referenced tests;
- failure to train raters in criterion referenced testing.

If all evidence collected to demonstrate that a test is valid for its purpose contributes to construct evidence (Messick, 1989), and if Shimberg (1990: p. 13) is correct when he says that "the collection of construct evidence, if done at all, is usually done after the test has been used in decision making", it is actually surprising that there has been only limited litigation over language tests which are used to admit or exclude overseas students from higher education.

THE SITUATION IN THE UNITED KINGDOM

Although codes of practice have been common in the United States (ETS 1983; APA, 1985; NCME, 1990), and organizations such as Educational Testing Service have published

research reports, including detailed studies of test reliability and validity, this trend has only just begun in the United Kingdom.

The most notable attempt to introduce standards to language testing has been conducted by the Association of Language Testers in Europe (ALTE), a consortium made up of the University of Cambridge Local Examinations Syndicate (UCLES), Generalitat de Catalunya, Danish Language Testing Consortium, Goethe-Institut, Instituto Cervantes and Universidad de Salamanca, Alliance Française, Universitá per Stranieri, Perugia, CITO, Instituut voor Toetsontwikkeling, and Universidade de Lisboa (ALTE, 1994a, 1994b, 1994c, 1994d). In the ALTE code of practice (ALTE, 1994c), UCLES, the largest EFL/ESL testing organization in the United Kingdom, commits itself to certain standards. These are provided under 4 headings: developing examinations, interpreting examination results, striving for fairness, and informing examination takers. Included under these headings are the following commitments:

- define what each examination assesses and what it should be used for. Describe the population(s) for which it is appropriate;
- describe the process of examination development;
- provide either representative samples or complete copies of examination questions, instructions, answer sheets, manuals and reports of results to users;
- describe the procedures used to ensure the appropriateness of each examination for groups of different racial, ethnic or linguistic backgrounds who are likely to be tested;
- describe the procedures used to establish pass marks and/or grades;
- if no pass mark is set, then provide information that will help users follow reasonable procedures for setting pass marks when it is appropriate to do so;
- warn users to avoid specific, reasonably anticipated misuses of examination results;
- enact procedures that help to ensure that differences in performance are related primarily to the skills under assessment rather than to irrelevant factors such as race, gender and ethnic origin.

Lack of published data regarding UCLES examinations has been commented upon in the past (Hamp-Lyons, 1987: p. 19), although this is beginning to change as research is not only carried out but also published (Davidson and Bachman, 1990; Bachman *et al.*, 1995; Kunnan, forthcoming). It is to be hoped that UCLES will make information and data available in all of the areas to which it has committed itself in the near future, on a regular basis.

The other organization which provides standards by which its members must operate is the Association of British ESOL Examining Boards (ABEEB) (Arnold, 1990). An altogether less substantial document (two sides of photocopied information) is available, entitled "Code of Practice for ESOL Examinations". Standards are mentioned in the areas of information to users, standardization of question setting, standardization of marking, awarding grades, providing "appropriate" appeals procedures, and administration. It is notable that the words "reliability" and "validity" do not appear on the document. Although standard 3(v) does read "the work of all examiners is monitored to verify the consistent application of the mark scheme throughout the process" it is not

said how this should be done, and no mention is made of the calculation and publication of reliability estimates. The concept of validity does not enter into this document at all. If it appears anywhere, it is in the faith or belief that all examinations of ABEEB members have been validated in the ESU Framework Project (Carroll and West, 1989). However, it is certainly the case that the Framework project has never been verified, and many of the placings may be invalid, if only because many of the examining bodies refused to allow the researchers to calculate the reliability of the ratings of pieces of work used to create the framework in the first place (personal communication). This situation is clearly unacceptable.

POTENTIAL GROUNDS FOR CHALLENGING TEST SCORES IN THE UNITED KINGDOM

The only major survey of the practices of examination boards in the United Kingdom is that of Alderson and Buck (1993). Their research involved an initial survey followed by a detailed questionnaire on standards in testing, as:

> "It seems that there is reason to fear that UK examination boards are failing to live up to the highest standards in educational measurement." (Alderson and Buck, 1993: p. 3)

The first phase of the Alderson and Buck survey was an open ended letter to the examination boards listed in the Appendix, asking (a) whether they had a set of standards to which they adhered; (b) what procedures they used to estimate test reliability; and (c) what procedures were used to ensure validity. Of the 14 examination boards surveyed, only ten responded. Alderson and Buck's results may be summarized as follows:

Table 1. Results of the first phase of the Alderson and Buck survey

Question	No. relevant responses	Content of relevant responses
Standards	4	2 referred to syllabus contents 1 said they were being prepared 1 responded "yes"
Reliability	3	1 referred to "standardizing markers" (sic) 2 provided information
Validity	3	content validity decided by "expert judges"

From the detailed questionnaire in the second phase of the study, to which 12 examination boards responded, the following findings emerged:

1. Syllabus: All examination boards claimed to produce a syllabus to make available to the public the content of their examinations, and a statement of the intended target population. In most cases, it was doubtful how these syllabuses were arrived at, or how they were operationalized in the tests.

2. *Examination construction*: Most test items were found to be constructed by teachers employed on a part-time basis. The test material was frequently checked by a committee or chief examiner, but the thoroughness was suspect. Eight boards said they did not pre-test items, and while 4 said that they did, it was not clear that at least 3 of these understood what pre-testing items meant.

3. *Validation*: Questions were asked about whether concurrent, predictive and/or concurrent validity were calculated. It transpired that half the respondents from the examination boards did not understand the questions. One board, which only produces an oral test, claimed that such questions are not relevant to oral tests. Half the boards claimed to assess construct validity, whilst 4 boards claimed to estimate predictive validity. However, this must be interpreted with some caution, as the extracts from replies from Alderson and Buck (1993: p. 11) clearly show this not to be the case. For example, validity is estimated "in impressionistic and anecdotal ways...."

Alderson and Buck (1993: p. 11) conclude that:

> "The responses show quite clearly that validation procedures as normally understood by language testers and as described in the language-testing literature are not used by most boards."

4. *Administration*: Boards do not train administrators/invigilators.

5. *Markers*: are employed on a part-time basis, and chosen on the grounds of "experience". Ten of the 12 boards organized some form of "standardizing meeting", but it was not clear how these operated. Double marking was found to be very rare. Two boards did not calculate interrater reliability, 4 boards said that they did it from time to time, and less than half said that they were routinely calculated. Cut scores were established at grade-award-meetings by only half of the boards, and establishing the cut scores appeared to be mostly subjective.

6. *Post-hoc analysis*: In only 1 case did a board calculate and make available test statistics other than percentages of passes at certain grade boundaries.

7. *Examination revision*: Examinations were constantly changing to meet market demands in three quarters of the examination boards.

LEGAL PRECEDENT

We believe that this situation leaves British examination boards open to litigation. Although the history of litigation in the U.K. over the use and interpretation of test scores is not as great as that in the U.S., the precedents which do exist are important. The U.K. legal framework for the use of tests in the U.K. has come about indirectly via anti-discrimination legislation introduced by the European Community via the European Community Council Directive 76/207 on Equal Treatment. The Equal Pay Act 1970, the Sex Discrimination Act 1975, and the Race Relations Act 1976, made the discrimination on the grounds of gender, race, colour, nationality or ethnic origin unlawful, and led to the establishment of the Equal Opportunities Committee (EOC) and the Commission for Racial Equality (CRE).

As in the U.S., most of the litigation in the U.K. concerns employment. But these legal concepts could easily be applied to other forms of testing. Most litigation concerns the validity of selection tests which are based on some form of cultural bias built into the test question. The key notion is that of discrimination, which in this context is defined as:

> *A person discriminates against another person if he treats the person less favourably than he treats or would treat other persons.*

In Perera versus Civil Service Commission (No. 2), (1983) ICR 428 Court of Appeal, a Sri Lankan graduate in science and law and a qualified accountant who had practised law in Sri Lanka before coming to England, been called to the Bar, and an Executive Officer in the Civil Service, alleged racial discrimination after his application for a transfer into the legal Civil Service was turned down. The selection process consisted of an interview, and the board were asked to take account of four factors: experience of the U.K., fluency in English, age, and the nationality of the applicant. The plaintiff maintained that these four factors would exclude all people of his racial group. However, the case was lost because the judge ruled that the plaintiff could not prove that the board had singled out any of the above factors in the rejection, their claim being that they relied on "personal qualities". This case does nothing but add to the complaint of all minorities that they have to be twice as good to be accepted. In Hampson versus Department of Education and Science (1989) ICR 179 Court of Appeal, a Hong Kong Chinese with qualified teacher status was rejected for qualified teacher status in the U.K. She claimed indirect discrimination, and lost the case, on the grounds that a discriminatory condition must be balanced by the reasonable needs of the party who applies the condition! On appeal to the House of Lords, this was overturned.

It appears to us, that tests which are standardized or pre-tested on samples which do not represent the population of test takers are likely to discriminate against those of some ethnic backgrounds. Tests in which 80% of black candidates fail whereas only 20% of white candidates fail are likely to be held as invalid by the courts unless it can be shown that the cut-off point is justifiable. In 1990, for example, eight guards at Paddington Station (London) brought an action against British Rail after they failed a train drivers test. The assessment included a psychometric aptitude test, personality questionnaire, an interview and tests of vigilance. A review of test scores indicated that the tests, particularly those of verbal comprehension, discriminated against racial minorities. The case was settled out of court, and British Rail agreed to review its testing procedures. A comparable case is that of the tests for admission to the system of training for barristers, which has come under severe criticism in recent years. In 1991/2, research indicated that 45% of ethnic minorities failed the entry tests as opposed to 17% of whites. A committee of enquiry, chaired by Dame Jocelyn Barrow, recommended, among other things, a system of double marking of all scripts.

DISCUSSION

We are now in a position to consolidate the data presented above, in tabular format. It is believed that legal challenges to test scores could be successfully made in one or more

of the sections of Table 2, given legal precedent and the present unsatisfactory state of British testing, as revealed by Alderson and Buck (1993).

It is clear that some progress is being made in British testing, particularly with the foundation of ALTE. The foundation of the International Language Testing Association (ILTA) which is currently developing its standards document (Davidson, 1994), will ensure that there are individual professional language testers who will implement standards in their work, whether the examination boards subscribe to them or not. However, this is far from institutionalising professional standards in the language testing industry.

It is also clear that many of the examination boards that set English language tests for students from overseas, do not meet the requirements laid out in Table 2. This, in some cases, is in clear contravention of the codes which they voluntarily agree to abide by. There is at

Table 2. Grounds for legal challenges to examination boards

Are there formal statements of standards?	Does the test have demonstrated technical qualities? procedures?	Does the examination board have appropriate quality control
Which includes:	Which includes:	Which includes:
Are these published?	Is this test reliable?	Did the examination board ensure that the administration of the test was fair?
Do examination boards in consortiums which subscribe to standards meet their publicly stated commitments?	Have cut scores been established empirically?	Is the marking procedure explained?
	Has the test been properly constructed and pre-tested?	Are raters trained?
	Has the examination board made every attempt to eliminate bias?	Does the examination board conduct and make available *post-hoc* analyses and results?
	Has the test been validated?	When the test syllabus last changed, was the institution given enough time to prepare students over a number of years?
	Has construct validity been estimated?	
	Has content validity been established?	
	Has curricular validity been established, where there is a related syllabus?	
	Has predictive validity been estimated, where this is approriate?	

present no way in which such organizations can be made to meet appropriate standards through the application of pressure, or of periodic inspection by so-called validating bodies.

There is scope in the U.K. legal system to allow challenges to test results in EFL examinations, in much the same way as it is possible in the U.S. The trend is for the litigation to begin in the field of employment testing, and to move to other areas, including psychological and educational testing.

CONCLUSION

Alderson and Buck (1993: p. 21) argued that:

> "There has not (yet) been a general call for public accountability on the part of the examining boards, and as long as EFL remains outside the mainstream of school examinations in the U.K., this is likely to remain so. We believe that U.K. examination boards should be publicly accountable, and that this accountability should extend to their EFL tests."

Despite the current move towards the production of "standards" and the publication of a limited number of research projects sponsored by British examination boards, we do not agree with Alderson and Buck that the answer necessarily lies with bringing EFL testing within government control. Indeed, their own comments on SEAC (Alderson and Buck, 1993: p. 5) tend to militate against their own conclusions. Rather, a number of successful legal challenges to test results which cannot be justified would do much to force examination boards to rapidly improve their products and their performance. We feel that only when (most) examination boards in the U.K. face the threat of litigation, as is the case in the U.S., will they begin to conduct the appropriate research, follow necessary procedures, and publish all statistics which external observers need to assess the reliability, validity and fairness of their tests.

NOTE

[1]Other reasons relate to the history of the development of language testing in the two countries. See excellent descriptions of the different traditions in Spolsky (1995a; 1995b).

REFERENCES

ALDERSON, J. C. and BUCK, G. (1993) Standards in testing: a study of the practice of UK examination boards in EFL/ESL testing. *Language Testing* **10**(1), 1–26.

ALDERSON, J. C., CLAPHAM, C. and WALL, D. (1995) *Language Test Construction and Evaluation.* Cambridge: Cambridge University Press.

ALTE (1994a) *European Language Examinations: Descriptions of examinations offered by members of the Association of Language Testers in Europe.* ALTE

ALTE (1994b) *European Examination Systems: Descriptions of examination system administered by members of the Association of Language Testers in Europe.* ALTE.

ALTE (1994c) *The ALTE Code of Practice: The Code of Practice for the Association of Language Testers in Europe.* ALTE.

ALTE (1994d) *The ALTE Framework: A description of the framework of the Association of Language Testers in Europe*. ALTE.

APA (1985) *Standards for Educational and Psychological Testing*. American Psychological Association/ American Educational Research Association/National Council on Measurement in Education.

ARNOLD, J. (1990) The Association of British ESOL Examining Boards. *Language Testing Update*, 7.

BACHMAN, L., DAVIDSON, F., RYAN, K. and CHOI, I. C. (1995) *An investigation into the comparability of two tests of English as a Foreign Language*. Cambridge: Cambridge University Press: Studies in Language Testing 1.

BERSOFF, D. (1979) Regarding psychologists testily: Legal regulation of psychological assessment in public schools. *Maryland Law Review* **29**, 27–120.

CARROLL, B. J. and WEST, R. (1989) *ESU Framework: Performance Scales for English Language Examinations*. London: Longman.

CHILDS, R. A. (1990) *Gender Bias and Fairness*. Princeton, NJ: ERIC Clearninghouse on Tests, Measurement and Evaluation, ERIC Digest ED328610.

DAVIDSON, F. (1994) Task Force on Test Standards. *Language Testing Update* **16**, 8–13.

DAVIDSON, F. and BACHMAN, L. (1990) The Cambridge-TOEFL Comparability Study: an example of cross-national comparison of language tests. In de Jong, J. (ed). *Standardization in Language Testing*, pp. 24–45. AILA Review.

ERIC (1985) *Legal Issues in Testing*. Princeton, NJ: ERIC Clearninghouse on Tests, Measurement and Evaluation, ERIC Digest, ED289884.

ETS (1983) *ETS Standards for Quality and Fairness*. Princeton, NJ.

HAMP-LYONS, L. (1987) Cambridge First Certificate English. In Alderson, J. C., Krahnke, K. J. and Stansfield, C. W. (eds), *Reviews of English Language Proficiency Tests*, pp. 18–19. TESOL Publications.

HOOD, S. and PARKER, L. (1991) Minorities, Teacher Testing, and Recent U.S. Supreme Court Holdings: A Regressive Step. *Teachers College Record* **92**(4), 603–618.

KLEINMAN, L. and FALEY, R. H. (1985) The implications of professional and legal guidelines for court decisions involving criterion-related validity: A review and analysis. *Personnel Psychology* **38**(4), 803–833.

KUNNAN, A. J. (forthcoming) *Test-taker characteristics and test performance: A structural modelling approach*. CUP: Studies in Language Testing 2.

MACMILLAN, D. L. and BARLOW, I. H. (1991) Impact of Larry P. on Educational Programs and Assessment Practices in California. *Diagnostique* **17**(1), 57–69.

McDONOUGH, M. W. and WOLF, W. C. (1988) Court Actions which helped define the direction of the competency-based testing movement. *Journal of Research and Development in Education* **21**(3), 37–43.

MESSICK, S. (1989) Validity. In Linn, R. L. (ed), *Educational Measurement*, pp. 13–104. Macmillan/American Council on Education.

NCME: National Council on Measurement in Education. (1990) Standards for Teacher Competence in Educational Assessment of Students. *Educational Measurement: Issues and Practice* **9**(4), 30–32.

PHILLIPS, S. E. (1991) Diploma Sanction Tests Revisited: New Problems From Old Solutions. *Journal of Law and Education* **20**(2), 175–199.

PRASSE, D. P. and RESCHLY, D. J. (1986) Larry P.: A Case of Segregation, Testing, or Program Efficacy? *Exceptional Children* **52**(4), 333–346.

RESCHLY, D. (1990) The Effects of Placement Litigation on Psychological and Education Classification. *Diagnostique* **17**(1), 6–20.

RESCHLY, D. J., KICKLIGHTER, R. and McKEE, P. (1988a) Recent Placement Litigation, Part I, Regular Education Grouping: Comparison of Marshall (1984, 1985) and Hobson (1967, 1969). *School of Psychology Review* **17**(1), 9–21.

RESCHLY, D. J., KICKLIGHTER, R. and McKEE, P. (1988b) Recent Placement Litigation, Part II, Minority EMR Overrepresentation: Comparison of Larry P. (1979, 1984, 1986) with Marshall (1984, 1985) and S-1 (1986). *School Psychology Review* **17**(1), 22–38.

RESCHLY, D. J., KICKLIGHTER, R. and McKEE, P. (1988c) Recent Placement Litigation, Part III: Analysis of Differences in Larry P., Marshall and S-1 and Implications for Future Practices. *School Psychology Review* **17**(1), 39–50.

SHIMBERG, B. (1990). Social Considerations in the Validation of Licensing and Certification Exams. *Educational Measurement: Issues and Practice* **9**(4), 11–14.

SPOLSKY, B. (1995a) Introduction: A not too special relationship. In Bachman, L. F., Davidson, F., Ryan, K. and Choi, I. C. (1995). *An investigation into the comparability of two tests of English as a Foreign Language.* pp. 1–13. Cambridge: Cambridge University Press: Studies in Language Testing 1.

SPOLSKY, B. (1995b) *Measured Words.* Oxford: Oxford University Press.

STANSFIELD, C. W. (1993) Some Ethical Issues and Problems in Language Testing. *Issues in Applied Linguistics* **4**(2), 189–205.

TAYLOR, R. L. (1991) Bias in Cognitive Assessment: Issues, Implications, and Future Directions. *Diagnostique* **17**(1), 3–5.

APPENDIX

ARELS Examination Trust
Associated Examining Board
City and Guilds of London Institute
English Speaking Board
Institute of Linguists Examination Board
Joint Matriculation Board
London Chamber of Commerce and Industry Examinations Board
National Board of Speech and Drama Studies
North West Regional Examinations Board
Pitman Examinations Institute
Trinity College London
University of Cambridge Local Examinations Syndicate
University of London School Examination Board
University of Oxford Delegacy of Local Examinations

Bibliographical Note on the Authors

Glenn Fulcher is Director of the English Language Institute at the University of Surrey. He holds an MA in Applied Linguistics from the University of Birmingham, and a Ph.D. from the University of Lancaster. After qualifying as a teacher at Cambridge University, he has been a teacher and teacher trainer for thirteen years.

Ron Bamford is Deputy Director of the Surrey European Management School. He qualified as a lawyer at the University of London, and is a Barrister-at-law of Lincoln's Inn.