Studies in Language Testing 9

Fairness and validation in language assessment

Selected papers from the 19th Language Testing Research Colloquium, Orlando, Florida

Edited by
Antony John Kunnan

Series Editor
Michael Milanovic

UNIVERSITY *of* CAMBRIDGE
Local Examinations Syndicate

CAMBRIDGE
UNIVERSITY PRESS

Published by the Press Syndicate of the University of Cambridge
The Pitt Building, Trumpington Street, Cambridge CB2 1RP, UK
40 West 20th Street, New York, NY 10011–4211, USA
10 Stamford Road, Oakleigh, Melbourne 3166, Australia

First published 2000

Printed in Great Britain at the University Press, Cambridge, UK

British Library cataloguing in publication data

University of Cambridge, Local Examinations Syndicate

Fairness and validation in language assessment: Selected papers from the 19th Language Testing Research Colloquium, Orlando, Florida.

Antony John Kunnan

1. Education. Assessment 2. Education. Tests. Setting

ISBN 0 521–651034 hardback
0 521–658748 paperback

Contents

1 Fairness and justice for all

Antony John Kunnan
California State University, Los Angeles

Introduction

Although it has been argued that language test developers and researchers are concerned with the concept of fairness when they investigate tests for technical qualities like validity and reliability, the primacy of fairness has not been considered or acknowledged. Furthermore, fairness as a concept within a framework of social justice has not been developed and debated. I hope to make a beginning on these matters in this short chapter by discussing a possible definition of fairness and connections between fairness and four critical areas in language testing or assessment: research, test development, legal challenges and test developers. As a concept, fairness is seemingly clear but quite complex and thus often lends itself to dangerous misunderstandings. Moreover, often it is said that fairness is in the eye of the beholder and such discussions of fairness are obviously interminable. So, a clarifying definition seems to be difficult and elusive. One document that provides direction on this matter is the *Code of Fair Testing Practices in Education* (*Code* from now on) prepared by the Joint Committee on Testing Practices (1988). It presents standards for educational test developers and users in four areas: developing and selecting tests, interpreting scores, striving for fairness and informing test takers.

Here is the excerpt from Section C, Striving for Fairness, of the *Code*:

Test developers should strive to make tests that are as fair as possible for test takers of different races, gender, ethnic backgrounds, or handicapping conditions.

Test developers should:
14 Review and revise test questions and related materials to avoid potentially insensitive content or language. 15 Investigate the performance of test takers of different races, gender, and ethnic backgrounds when samples of sufficient size are available. Enact procedures that help to ensure that differences in performance are related primarily to the skills under the assessment rather than to irrelevant factors. 16 When feasible, make appropriately modified forms of tests or administration procedures available for test takers with handicapping conditions. Warn test users of potential problems in using standard norms with modified tests or administration procedures that result in non-comparable scores.

Test users should select tests that have been developed in ways that attempt to make them as fair as possible for test takers of different races, gender, ethnic backgrounds, or handicapping conditions.

Test users should:
14 Evaluate the procedures used by test developers to avoid potentially insensitive content or language. 15 Review the performance of test takers of different races, gender, and ethnic backgrounds when samples of sufficient size are available. Evaluate the extent to which performance differences may have been caused by inappropriate characteristics of the test. 16 When necessary and feasible, use appropriately modified forms of tests or administration procedures for test takers with handicapping conditions. Interpret standard norms with care in the light of the modifications that were made.

(Code 1988,p.2-3)

Towards a definition

Using the *Code* as a set of guiding principles, a definition of fairness for language assessment can be attempted. In general, the *Code* urges both test developers and test users to strive for fair tests and testing practices as far as possible for all test takers. Specifically, in the three points the *Code* urges test developers and test users to review and revise insensitive test content or language, investigate differential test performances and ensure construct irrelevant factors are not being assessed, and provide accommodations for test takers with disability. In addition to these three points, two other main

concerns such as access to tests and impact of testing practice have been of considerable recent interest and therefore need to be added to the list. Table 1.1 summarizes the main concerns of fairness and their specific focuses.

Table 1.1
Main concerns of fairness

Main concern	*Specific focus*
Validity	construct validity content and format bias Differential Item/Test Functioning insensitive language stereotyping of test taker groups
Access	financial: affordability geographical: location and distance personal: accommodations for disabled persons educational: opportunity to learn equipment and test conditions
Justice	societal equity legal challenges

Validity

The focus of this concern is on whether test-score interpretations have *equal construct validity* (and reliability) for different test-taker groups as defined by salient test-taker characteristics such as gender, race/ethnicity, field of specialization and native language and culture. Construct-irrelevant factors in terms of content bias that might cause unfairness among groups include topical knowledge and technical terminology, specific cultural content and dialect variations. Format bias could include multiple-choice, constructed response, computer-based responses and multi-media materials. The *Code* calls for investigations of test performance of *different test-taker groups* so that test developers and test users are confident that the differences in performances are related primarily to the abilities that are being assessed and not to construct-irrelevant factors. Other key construct-irrelevant factors include *insensitive* or *offensive test materials* and materials that *stereotype* and show certain test-taker groups in unfavourable light.

Access

The focus of this concern is on whether tests are accessible to test takers from various aspects such as financial, geographical, personal, and educational access and familiarity of test conditions and equipment. *Financial* access in terms of affordability is a key concern as the consequences of unaffordable tests in all regions should be known to test developers and test users. Similarly, *geographical* access to test sites is critical too and this also varies from context to context. Once again, what is considered accessible in one region may not be so in another. Another focus is *personal access.* The focus here is on providing where feasible appropriate accommodations in test administration procedures for test-takers with disability or impairment. *The Code* calls for this modification in order that test takers who are disabled are not denied access to tests that can be offered without compromising the construct being measured. *The Code* also indicates that test users should be warned of the type of accommodations provided so that test-score interpretations can be made in the light of the accommodations. In terms of educational access, the focus is on opportunity to learn. There is no doubt that opportunity to learn plays a major role in test-takers' success on tests when test-takers have had the opportunity to learn the material on which they are assessed. Further, if test-taker groups have differential opportunities to learn, then group performance on a test will most certainly differ significantly. In large-scale assessment programs, in many cases differential opportunities to learn among test takers is common, and therefore, unequal advancement may result. Yet another focus is on whether test takers have had prior *access to test-taking equipment and test-taking conditions* so that they are familiar with these conditions. Relevant examples here are the use of computers in computer-based tests and the use of multi-media in web-based testing.

Justice

The focus of this concern is on justice in terms of *societal equity* and *legal challenges*. Specifically, the notion of societal equity goes beyond equal validity and access and focuses on the social consequences of testing in terms of whether testing programs contribute to social equity or not and in general, whether there are any pernicious effects due to them. For example, if a test taker group (defined by political ideology, native language, race/ethnicity, gender, national origin or socioeconomic status) as a result of a testing program does not gain equal access to college or promotion on the job in the same proportion when compared to other test-taker groups, there could be legitimate concern that the testing program is causing the inequity rather than that the inequity among the groups actually exists. The focus of this concern would be to devise a mechanism that can investigate the burden on the testing program to show that the societal inequity is not an artifact of the testing program.

Related to societal equity and assessment is the issue of standards in assessment practices which have not been clearly formulated and this has led to *legal challenges* particularly in the US and UK. In the US, Title VII of the Civil Rights Act of 1964 (and subsequent related legislation) provides remedies for persons who feel they are discriminated against due to their gender, race/ethnicity, native language, national origin and so on. This Act has been used broadly; for example, to challenge the use of test scores, the curricular validity and predictive validity of tests in school and in employment contexts.

In summary, the way the different concerns of validity, access and justice contribute to the multi-faceted definition of fairness indicates that the concept is an interdisciplinary one; not only based on the psychometric view of tests and testing practice but also on social, ethical, legal and philosophical views. A definition of fairness along these lines is stated by Jensen, an unlikely scholar on the subject, who writes that fairness refers

> *'to the ways in which test scores (whether of biased or unbiased tests) are used in any selection situation. The concepts of fairness, social justice, and equal protection of the laws are moral, legal, and philosophical ideas and therefore must be evaluated in these terms.*
>
> (Jensen 1980: 376)

Fairness and research studies

The research studies that have focused on fairness in language assessment over the last 15 years (taken from Kunnan 1998a) are not many in number nor part of a coherent research program either. Table 1.2 presents some of the best examples of such placed within the fairness framework listed in Table 1.1.

Table 1.2
Studies with fairness concerns in language assessment (1985–1999)

Fairness concerns	*Studies*	*Specific focus*
Validity:		
construct validity	Alderson and Urquhart 1985a, b	academic major and reading
	Hale 1988	major field and test content
	Clapham 1996 1998	ESP testing
	Norton and Stein 1998	test taker feedback
	Kunnan 1995	+/- Indo-European languages
	Ginther and Stevens 1998	native language groups
	Kunnan, 1992	standard setting and placement
	Wall and Alderson 1993	washback
	Alderson and Hamp-Lyons 1996	test preparation
DIF	Alderman and Holland 1981	native language
	Chen and Henning 1985	native language
	Zeidner 1986,1987	sex, age and minority bias
	Kunnan 1990	native language and gender
	Ryan and Bachman 1992	gender
content	Lowenberg 1989	different Englishes
format	Shohamy 1984	test method and reading
	Shohamy and Inbar 1991	question type and listening
Access:		
test conditions	Brown 1993	tape-mediated test
	Taylor *et al.* 1998	computer familiarity
Justice	none	

Although these studies may seem like many examples of research focused on fairness, there is clearly a great need for more studies in this area. Also, most of these studies listed are generally post-hoc analyses and independent studies that are not part of a coherent fairness research program that is part of test development, maintenance and research program. Quite obviously more needs to be done. Perhaps, examples of research studies and general articles from the field of general assessment that are relevant to the fairness program could help propel language assessment researchers. For example, many fairness issues in the US have been brought to the forefront in recent years. Among the issues discussed include gender differences in education (Sadker and Sadker 1994), gender differences scores on the SAT-Math section (Wainer and Steinberg 1992), bias in the assessment of bilingual students (Hamayan and Damico 1991), testing African American students (Hilliard, 1991), and bias in reading tests for Black language students (Hoover, Politzer and Taylor 1991). In addition, articles on test sensitivity review (Ramsey 1993), assessment and diversity (Garcia and Pearson 1994), equity issues and American testing policy (Madaus 1994), educational equity and performance assessment (Darling-Hammond 1994) and equitable assessment policies for

English language learners (Lacelle-Peterson and Rivera 1994) can provide an understanding of how fairness concerns are discussed outside the language assessment arena.

Fairness and test development

A framework to focus on the fairness concerns articulated during all stages of the test development, maintenance, and research needs to be developed. Table 1.3 presents the stages and the fairness concerns that need to be focused on for optimum administration of the fairness agenda.

Table 1.3
Fairness concerns and stages of test development

Stages	Fairness concerns	
Thinking	Validity:	construct
		content and format
		scoring and reporting
	Access:	financial: affordability
		geographical: location and distance
		personal: accommodations
		educational: opportunity to learn
		equipment and test conditions
	Justice:	societal equity
Writing	Validity:	tasks, topics, canon
		language standards
		insensitive language review
		stereotyping of societal groups
Piloting	Validity:	norming samples
Analyzing	Validity:	item/task
		analysis
		internal structure
		scoring, raters
		differential item/test functioning
		speededness
	Justice:	societal equity
Maintenance and Research	Validity:	all areas
	Access:	all areas
	Justice:	all areas

As Table 1.3 shows, fairness concerns need to begin with the *thinking* stage which involves thinking about the construct(s), thinking about the content and possible tasks and task methods, and thinking about scoring and reporting issues. In addition, it is critical that issues of access are discussed at this stage and not left until a later stage. In terms of justice, test developers should check to see if the test under development will generally bring about societal equity rather than disharmony. In other words, the question that should be discussed is whether the test will generally do good to society.

Fairness concerns at the *writing stage*, which include decisions about operationalization of constructs into actual written tasks, include discussions regarding the canon from which topics and tasks may be chosen. In other words, the discussion should centre round whether the canon is something that all potential test takers share and learn. In addition, decisions regarding the language standard(s) (or dialects) that are to be adopted for the test need to be made by the developers and writers. Finally, after tasks are written, reviews of tasks for insensitive language and stereotyping of societal groups needs to be conducted.

The third place for fairness concerns is the *piloting* stage in which a test is typically piloted or pre-tested with a norming sample from the intended test-taking population. The sample should be a truly representative sample and not a sample of convenience. This choice is very critical at this stage because how the sample's performance on the tasks is used in making decisions about the tasks.

The fourth place for fairness concerns is the *analyzing* stage in which data collected from test-takers is analyzed. Traditional item analysis, internal structure analysis, rating reliability and rater conduct should be conducted. In addition, investigations of differential item/task or testlet functioning should be conducted in order to be able to state confidently that score differences in performance on the test from different test taker groups are due to relevant construct variance. Further, the issue of speededness needs to be investigated so that the speed of the test is not felt differently by the various test-taker groups (for example, non-native speakers as opposed to native speakers). The analyses should also include how the test might contribute to societal equity.

In the *maintenance and research* stage, all fairness concerns itemized in Table 1.3 should be routinely investigated.

Collectively then, these different fairness concerns at the different developmental stages should uncover any invalidities or unfairness a test might carry, and when follow-up corrective action is taken, it might be clearly possible to minimize or eliminate any invalidites or unfairness.

Fairness and legal challenges

The notion of fairness may be sufficient grounds for challenging a test wherever *equal protection* legislation has been provided by a state constitution or through separate legislation. In addition, whenever a test is in clear violation of a code of standards, if such a code exists, there may be sufficient grounds for a challenge.

A few examples of US Court rulings will be briefly presented in order to provide a flavour of how US courts have viewed legal challenges in the general educational and employment arena. A fuller discussion of relevant court cases is discussed by Bersoff (1981, 1984), McDonough and Wolf (1988), Hood and Parker (1991), Pullin (1994), Fulcher and Bamford (1996)

and Lippi-Green (1997). A selected list of cases with sources from all these discussions is presented in Appendix A.

As an example, one ground for legal challenge has been based on the perception that there is lack of societal equity due to tests that track and classify students in schools. Examples of litigation in the US in this area were *Hobson v. Hansen* (1967), *Larry P. v. Riles* (1971, 1984) and *PASE v. Hannon* (1980). In all three cases, the plaintiffs charged that African American children were being discriminated against as disproportionate numbers of such children were placed based on test scores into a lower-track program (in the first listed case), into a mildly mentally retarded program (in the second case) and into an educable mentally handicapped program (in the last case). The courts found for the plaintiffs in all three cases. In *Debra P. v Turlington* (1981), the ground for legal challenge was curricular after African American students who took a minimum competency test had initially approximately ten times the failure rate of White students. The Court found for the plaintiff stating that 'if the test covers material not taught the students, it is unfair and violates the Equal Protection and Due Process clauses of the US Constitution' (Debra P., at 402).

In *Griggs v. Duke Power Co*. (1971), the ground for challenge was the requirement of a passing test score in addition to a high school diploma for promotion on the job after African Americans working at the company were denied promotion. The Court found for the plaintiff stating that employment tests should be job related: 'What Congress has commanded is that any tests used must measure the person for the job and not the person in the abstract' (Griggs, at 436). In *Albermarle Paper Co. v. Moody* (1975), a test was found invalid as it was not designed to the standards laid down by the American Educational Research Association, particularly referring to the technical quality of employers' validity and reliability studies. In *Golden Rule Insurance Co. v. Mathias* (1984), an out-of-court settlement was agreed upon between the Golden Rule Insurance Company on the one hand and the Illinois Department of Insurance and Educational Testing Service (ETS, the test developer) on the other. All the parties agreed that 'a raw difference, favouring White applicants over Black applicants, of .15 or more in an item's p-values was to be taken as evidence that the item is to be considered biased in the social sense, that is, unfair to the lower-performing group, and identified as an item not normally to be included in the test' (Angoff 1993: 14).

It should be noted here that US Courts have intervened in some contexts but ignored others and have made a few controversial rulings. As Garcia and Pearson (1994) state, '(US courts) have intervened to offset the adverse impact of using test scores to place students of colour in remedial programs' they have not actively constrained the use of the same or similar tests to keep minority students from being placed in gifted programs or college-bound

high-school tracks' (p. 353). Moreover, in employment related cases, they have ruled that 'separate prediction equations and/or lower cut scores must be used to counteract employment discrimination' (ibid: 353). The Golden Rule out-of-court settlement is also an example of court-directed modification in ETS' test development practice for the Illinois insurance licensing examinations.

In summary, challenging a test is possible but until appropriate legislation and a code of standards exists, test takers may have difficulty seeking and obtaining remedies. And, from the test developers' perspective, a test can be challenged because standards and legislation do not exist or are somewhat poorly defined. These issues need to be addressed in every state/province or country where tests are developed and administered so that fair tests are available.

Fairness and test developers

One of the best ways to attain fairness in a test is when test developers (such as thinkers, writers, raters, and researchers) are from a diverse group (in terms of gender, race/ethnicity, native language, etc.) and trained to examine all aspects of a test for its fairness. This would help first, in obtaining different viewpoints concerning the canon, topics, tasks, format, and second, in examining tasks for the specific fairness concerns and third, in setting a research agenda that can enhance fairness.

Conclusion

In conclusion, this paper attempts to present an argument that fairness in language assessment consists of validity, access and justice. The paper also demonstrates that fairness is critically connected to research, test development, legal challenges and test developers. Newer methodologies such as item level exploratory and confirmatory factor analysis (see Bachman and Eignor 1997), structural equation modeling (see Kunnan 1998b), Multidimensional Item Response Theory for DIF (Ackerman 1998), Rule Space (Buck and Tatsuoka, 1998) and verbal protocol analysis (Greene 1997) may provide new avenues for research investigations in these areas. Furthermore, the paper implicitly argues that fairness is a critical central component not just connecting traditional components like validity and reliability (see Kunnan 1997). This conceptualization gives primacy to fairness and in my view if a test is not fair there is little or no value in it being valid and reliable or even authentic and interactive. As Rawls (1971) states, one of the principles of fairness is that institutions or practices must be *just*. Echoing Rawls then, there is no other way to develop tests but to make them such that primarily there is fairness and justice for all.

References

Ackerman, T. (1998) A discussion of measurement direction in a multidimensional latent space and the role it plays in bias detection. In D. Laveault, B. Zumbo, M. Gessaroli, and M. Boss (Eds.) *Modern Theories of Measurement: Problems and Issues:* 105-140).Ottawa, Canada: Edumetrics Research Group, University of Ottawa.

Alderman, D. and P. Holland (1981) Item performance across native language groups on the TOEFL. *TOEFL Research report 9*. Princeton, NJ: Educational Testing Service.

Alderson, J. C. and L. Hamp-Lyons, (1996) TOEFL preparation courses: A study of washback. *Language Testing* 13: 280–297.

Alderson, J. C. and A. H. Urquhart (1985a) The effect of students' academic discipline on their performance on ESP reading tests. *Language Testing* 2: 192–204.

Alderson, J. C. and A. H. Urquhart (1985b) This test is unfair: I'm not an economist. In P. C. Hauptman, R. LeBlanc and M. B. Wesche (Eds.) *Second Language Performance Testing*. Ottawa: University of Ottawa Press.

Angoff, W. (1993). Perspectives on Differential Item Functioning Methodology. In Holland and Wainer, PP.3-23.

Bachman, L. F. (1990) *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

Bachman, L. F. and D. Eignor (1998) Recent advances in quantitative test analysis. In D. Corson and C. Clapham (Eds.) *Encyclopedia of language and Education.* (Volume 7. Language and assessment). Dordrecht: Kluwer Academic Publishers.

Bersoff, D. (1981) Testing and the law. *American Psychologist* 36: 1047–1056.

Bersoff, D. (1984) Social and legal influences on test development and usage. In B. Plake (Ed.) *Social and Technical Issues in Testing*: 87–109. Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Brown, A. (1993) The role of test taker feedback in the test development process: Test takers' reactions to a tape-mediated test of proficiency in spoken Japanese. *Language Testing* 10: 277–304.

Buck, G. and K. Tatsuoka (1998) Application of Rule-space methodology to listening test data. *Language Testing* 15: 118–142.

Camilli, G. (1993) The case against item bias detection techniques based on internal criteria: Do item bias procedures obscure test fairness issues? In (*Holland and Wainer*) 397–413.

Chen, Z. and G. Henning (1985) Linguistic and cultural bias in language proficiency tests. *Language Testing* 2: 155-163.

Clapham, C. (1996) *The Development of IELTS*. Cambridge: Cambridge University Press.

Clapham, C. (1998) The effect of language proficiency and background knowledge on EAP students' reading comprehension. In Kunnan (1988a) 141–168.

Code of Fair Testing Practices in Education. (1988) Washington, DC: Joint Committee on Testing Practices.

Darling-Hammond, L. (1994) Performance-based assessment and educational equity. *Harvard Educational Review* 64: 5–30.

Fulcher, G. and R. Bamford (1996). I didn't get the grade I need. Where's my solicitor? *System* 24: 437–448.

Garcia, G. and D. Pearson (1994) Assessment and diversity. *Review of Research in Education* 20: 337–391.

Ginther, A. and J. Stevens (1998) Language background, ethnicity, and the internal construct validity of the Advanced Placement Spanish language examination. In Kunnan (1988a) 169–194.

Green, A. (1997) *Verbal Protocol Analysis in Language Testing Research*. Cambridge: Cambridge University Press.

Hale, G. (1988) Student major field and text content: Interactive effects on reading comprehension in the TOEFL. *Language Testing*, 5: 49–61.

Hamayan, E. and J. Damico (Ed.) (1991) *Limiting Bias in the Assessment of Bilingual Students*. Austin, TX: Pro-Ed.

Hilliard, A. G. (Ed.) (1991) *Testing African American Students.* Morristown, NJ: Aaron Press.

Holland, P. and H. Wainer (Eds.) (1993) *Differential Item Functioning.* Hilsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Hood, S. and L. Parker (1991) Minorities, teacher testing, and recent US Supreme Court holdings: A regressive step. *Teachers College Record* 92: 603–618.

Hoover, M., R. Politzer and O. Taylor (1991) Bias in reading tests for Black language speakers. In Hilliard, 81–98.

Jenson, H. R. (1980). Bias in mental testing. New York, N.Y: FreePress

Kunnan, A. J. (1990) DIF in native language and gender groups in an ESL placement test. *TESOL Quarterly* 24: 741–746.

Kunnan, A. J. (1992) An investigation of a criterion-referenced test using G-theory, and factor and cluster analysis. *Language Testing* 9: 30–49.

Kunnan, A. J. (1995) *Test-taker characteristics and Test Performance: A Structural Modeling Approach.* Cambridge: Cambridge University Press.

Kunnan, A. J. (1997) Connecting fairness and validation. In A. Huhta, V. Kohonen, L. Kurki-Suomo and S. Luoma, *Current Developments and Alternatives in Language Assessment:* 85–105. Jyvaskyla, Finland: University of Jyvaskyla.

Kunnan, A. J. (Ed.) (1998a) *Validation in Language Assessment.* Mahwah, N.J: Lawrence Erlbaum Associates, Publishers

Kunnan, A. J. (1998b) An introduction to structural equation modeling for language assessment research. *Language Testing* 15: 295–332.

Lacelle-Peterson, M. and C. Rivera (1994) Is it real for kids? A framework for equitable assessment policies for English language learners. *Harvard Educational Review* 64: 55–75.

Lippi-Green, R. (1997) *English with an Accent.* London: Routledge.

Lowenberg, P. (1989) Testing English as a world language: Issues in assessing nonnative proficiency. In J. Alatis (Ed.) *GURT* 1989: 216–227. Washington, DC: Georgetown University Press.

Madaus, G. (1994) A technological and historical consideration of equity issues associated with proposals to change the nation's testing policy. *Harvard Educational Review* 64: 76–95.

McDonough, M. and W. Wolf (1988) Court actions which helped define the direction of the competency-based testing movement. *Journal of Research and Development in Education* 21: 37–43.

Messick, S. (1980) Test validity and ethics of assessment. *American Psychologist:* 35: 1012–1027.

Messick, S. (1989) Validity. In R. Linn (Ed.), *Educational Measurement*, 13–103. London: Macmillan.

Norton, B. and P. Stein (1998) Why the 'Monkeys Passage' bombed: tests, genres, and teaching. In Kunnan: 231–249.

Pullin, D. (1994) Learning to work: The impact of curriculum and assessment standards on educational opportunity. *Harvard Educational Review* 64: 31–54.

Ramsey, P. (1993) Sensitivity review: The ETS experience as a case study. In Holland and Wainer, 367–388.

Rawls, J. (1971) *A theory of justice*. Cambridge, MA: The Belknap Press of Harvard University Press.

Ryan, K. and L. F. Bachman (1992) Differential item functioning on two tests of EFL proficiency. *Language Testing* 9: 12–29.

Sadker, M. and Sadker, D. (1994) *Failing at fairness*. New York, NY: Touchtone/Simon and Schuster.

Shohamy, E. (1984). Does the testing method make a difference? The case of reading comprehension. *Language Testing* 1: 147–170.

Shohamy, E. and O. Inbar (1991) Construct validity of listening comprehensive test of oral proficiency. *Language Testing* 8: 23-40

Taylor, C., Jameison, J., Eignor, D. and I. Kirsch (1998) The relationship between computer familiarity and performance on computer-based TOEFL tests tasks. [*TOEFL Research Report No. 61*]. Princeton, NJ: Educational Testing Research.

Wainer, H. and L. Steinberg, (1992) Sex differences in performance on the mathematics section of the Scholastic Aptitude Test: A bidirectional validity study. *Harvard Educational Review* 62: 323–336.

Wall, D. and C.Alderson, (1993) Examining washback: The Sri Lankan impact study. *Language Testing* 10: 41–69.

Zeidner, M. (1986) Are English language aptitude tests biased towards culturally different minority groups? Some Israeli findings. *Language Testing* 3: 80–95.

Zeidner, M. (1987) A comparison of ethnic, sex and age biases in the predictive validity of English language aptitude tests: Some Israeli data. *Language Testing* 4: 55–71.

Appendix A

Selected list of US Court Cases with sources

Albermarle Paper Co. v. Moody, 422 US (1975)

Debra P. v. Turlington, 474F. Supp. 244 (1979); aff'd in part, rev'd in part, 644 F. 2d 397 (5th Cir. 1981)

Firefighters Institute v. City of St. Louis, 616 F. 2d 350 (8th Cir. 1980)

Golden Rule Insurance Co. v. Mathias, (1984)

Griggs v. Duke Power Co., 401 US 424 (1971)

Hobson v. Hansen, 269 F. Supp. 401 (1967)

Larry P. v. Riles, 495 F. Supp. 926 (1979); aff'd in part, rev'd inpart, 793 F. 2d 969 (9th Cir 1984)

Mandhare v. LaFargue Elementary School, 605 F. Supp 238 (1985), rev'd, (5th Cir. 1986)

McNeil v. Tate, 508 f. 2d 1017 (5th cir. 1975)

PASE v. Hannon, 506 F. Supp. 831 (1980)

Teal v. Connecticut, US, 102 S. Ct. 2525 (1982)

Washington v. Davis, 426 US 229 (1976)