Ethical issues in language testing:

Research into Sexism in Language Testing

& its Implications to Language Testing in China

Tao Baiqiang, Independent Researcher，taobq@yahoo.com

Paper presented at The 2007 International Conference on Language Testing, Nov. 3-4, 2007, Beijing,China

Note: Citation model: Tao, Baiqiang. (2007). *Research into sexism in language testing & its implications to language testing in China.* Paper presented at The 2007 International Conference on Language Testing, Nov. 3-4, 2007, Beijing, China.

Abstract: This paper reviews foreign and domestic sexism research and practice in language testing and reveals that China lags behind in this sociolinguistics perspective in both theoretical study and practice. The paper indicates that sexism is represented in the listening comprehension section in National Matriculation English Test (NMET) after a case study into the listening comprehension part in NMET 2000-2005. It is hoped this tentative research will trigger more studies, ameliorate language testing practice in the perspective of gender equality and reduce DIF (from which validity will suffer) in especially large scale language testing programs.

Key words: language testing; NMET; sexism; test bias

1. Introduction

In the mid of the 19th century, especially in 1960s, the feminist campaigns in western countries triggered research into gender inequality in education, and they demanded that gender inequality should be eliminated from educational system including curriculum planning, instruction materials, and classrooms (Shi, 2000). Research into gender inequality in textbooks prevailed in early 1970s and since then relevant studies of sexism in ESL/EFL textbooks have become popular. Thereafter western scholars conducted a profusion of empirical research into sexism in classroom and assessment. Chinese scholars initially introduced the study of “language and gender” into China academia in 1980s and in 1990s there appeared lots of articles on the English language and gender. Research into sexism represented in China’s textbooks was initiated in 1990s and has become prevalent since the year 2000 (Tao, 2006a).

Testing, as one of the assessment tools, plays a significant role in curriculum design and instruction, mainly because some test results are used in policy-making; The larger a test scale is, the greater impact it will produce; Moreover, the impact of testing upon teaching and learning even ripples to educational system and the society as a whole (Zou, 2005). Sexism in the present one-shot affair—NMET, labeled as high-stakes testing, not only presents a challenge to equal admission to tertiary education, but also impedes test candidates developing positive gender schemas.

Gender differences in large-scale assessment are considered by many to be the most carefully examined aspect of test fairness (Ryan & Demark, 2002). Testing strongly affects our life and work. Although even item flaws were detected in NMET (Tao, 2007), sociolinguistic perspectives should be considered to improve tests. The significance of using fair test items cannot be overemphasized especially in China because test scores from the national college entrance exam is the most critical determinant of candidates’ admission to postsecondary institutions. Those items that may stereotype and under-represent a particular subgroup of examinees may lead the subgroup to be less motivated to do well in the test (Clauser & Mazor, 1998; Tittle, 1982, as cited in Tae-Il Pae, 2002: 98). Gershuny (1977: 143) states that stereotypes limit behavior and understanding by constructing a static image of both sexes and establishing a false impression of male and female characters as an alternative to their socio-cultural origins. In addition, stereotypes are seen as socio-political hierarchies in which one sex is considered superior or dominant over the other inferior sex.

2. Sexism into LT research in foreign countries

Test bias first received attention as early as in 1911, when Binet had to revise his earlier IQ tests after finding that children of high socio-economic status scored much higher than those of low socio-economic status (Eels et al., 1951). Scheuneman & Slaughter (1991) list five sources of bias: historical, cultural, biological, educational, psychometric sources and biological bias refers to gender-related bias.

As early as in 1960s LT was reviewed in the perspective of sociolinguistics when Coffman (1961, as cited in Wood, 1993:169) found there existed male and female topics in SAT. Dwyer (1976) has identified two types of sex bias-inducing factors in the content of test instruments: stereotyped sexes role and under-representation of the sexes, biased item types or technical aspects. In 1980s as LT evolved into a discipline (Bachman，L. F. personal communication），a deluge of researches on testee characteristics and language performance appeared (Farhady, 1982, as cited in Bachman, 1990:278; Locke 1984，as cited in Brown & McNamara, 2004; Zeidner 1987, as cited in Bachman, 1990:278). In 1990s, gender and LT remained a hot topic. Bachman (1990: 291) suggests in developing and using language tests sex, cultural background, prior knowledge of subject matter, language learning styles, native languages, race, age etc, which may lead to test bias, should be considered. Childs (1990) suggests test questions should be checked for: material or references that may be offensive to members of one gender, references to objects and ideas that are likely to be more familiar to men or to women, unequal representation of men and women as actors in test items or representation of members of each gender only in stereotyped roles. Language and gender expert Sunderland (1994, 1995) published two widely-cited articles in *Language Testing Update: Gender and Language Testing: Any Connection? Preliminary Explorations* and *Gender and Language Testing.* In the 21st century, conversation analysis and discourse analysis were introduced into research on oral language test and gender (O’Sullivan, 2002, as cited in Brown & McNamara, 2004). O’Sullivan found that candidate speech was more accurate when the interlocutor was female, but he found no gender effect for complexity; overall, gender did not have a significant effect on candidate speech.

Gender bias may be checked at least in four aspects in language testing: general verbal ability difference, the topic or content, task format, and the rating process which are thereafter elucidated.

2．1 Gender difference in LT scores

Early researches focused on language proficiency differences between males and females. Many researches revealed that females outperform males in overall language ability, but Hyde & Linn (1988) concluded that generally females were found to have slight advantages in reading, speaking, writing, and general verbal ability, but the differences were so small that Hyde and Linn argued that gender differences in verbal ability no longer exist (as cited in Jie & Fenglan 2003). Conversely, a study involving 400 tests and millions of students conducted by the Educational Testing Service (ETS) yielded contrary results that a language advantage for females had remained unchanged compared with 30 years ago (Cole, 1997, ibid). Takala & Kaftandjieva (2000) analyzed gender differential item functioning (DIF) in a second language (L2) vocabulary test with the tools of item response theory and found that the test as a whole is not gender-biased although some item composites proved gender-biased.

It is debatable whether females outperform males in general verbal ability. Teachers may believe that girls are better language learners and such beliefs may influence teachers' expectations and treatment of students, then the 'Pygmalion effect' may facilitate female language excellence.

2．2 Test contents/ topics

Test contents or topics significantly influence test performance between sexes and there exist so-called female and male topics. Coffman (1961, as cited in Wood, 1993:169) noticed that women did better on SAT-verbal items geared towards human relationships and personalities while men performed better on items on technical or mechanical terms. Donlon (1973）demonstrated that items on the SAT-Verbal section drawn from the world of practical affairs or science are easier for males, while items associated with human relations, humanities or aesthetics are easier for females. Doolittle & Welch (1989，as cited in Brantmeier, 2003) found notable gender differences for items associated with specific passages, reporting that females scored higher than males with humanities-oriented reading passages, but lower than males with science-oriented passages.

Bugel & Buunk (1996, as cited in Brantmeier, 2004) found that males scored significantly better on the multiple choice comprehension items for essays about laser thermometers, volcanoes, cars, and football players. Females achieved significantly higher scores on the comprehension tests for essays on text topics such as midwives, a sad story, and a housewife's dilemma. The researchers concluded that the topic of a text is an important factor in explaining gender-based differences in second language reading comprehension.

Female topics	human relationships，humanities-oriented, aesthetics, topics on midwives, a sad story, a housewife's dilemma
Male topics	technology，mechanics，science，war，laser，volcano, automobile，football，boxing

2．3 Test formats/Response types

Test method facet has been considered as an important factor affecting the testee's performance on a test. That is, a test used to assess a particular ability would yield different results when different test methods are used to gauge the same trait.

Impact of test formats on performance focuses on multiple choice (MC) items and essay writing. Murphy (1978, 1982, as cited in Wood, 1993:171-172) disclosed that MC both favors boys and disadvantages girls and that essay test has the opposite effect. Bolger & Kellaghan （1990）found an advantage for males on multiple-choice tests and offered explanations based on other research, such as that males tend to guess more on multiple-choice exams, an advantageous strategy relative to omission, which is practiced more by girls.

The overall findings of Brantmeier’s research（Brantmeier, 2004）indicate that females may have an advantage over males in the free written recall procedure. These findings raise issues for test developers and educational policymakers pertaining to the choice of measurement in high-stakes tests.

2．4　Rating

The performance test is generally conducted by observing testees’ task performance and ranking testees with a holistic rating tool (Han, 2003). Performance tests, though with a comparatively high validity, requires reasonable rater reliability. The tester may affect the test results in three ways: (a) by marking female or male candidates preferentially (Carroll 1991, as cited in Brown & McNamara, 2004; Eckes, 2005; O'Loughlin, 2002; Spear 1984, as cited in Greatorex & Bell, 2002), (b) if male and female testers have different standards from each other, and (c) if, especially on an oral test, the sex of testers influence candidates score or ranking (Locke 1984，Carroll 1991, as cited in Brown & McNamara, 2004; O’Neil 1985, as cited in Greatorex & Bell, 2002; Porter & Shen, 1991, as cited in Sunderland, 2000).

3．Sexism into LT research in China

In terms of sexism into LT research, China academia focus on English language itself and textbooks and yields no systematic study on sexism elimination strategies in language testing despite some researches on English test score differences (Xiao & Xiang, 2005; Zhang & Du, 2004).

In China’s English language testing books no chapter deals with gender bias in language testing (Li, 2001; Wu, 2002; Zou, 1998). Some papers have mentioned avoiding gender bias or sexism in language testing (Liu, 1993; Zhou & Fu, 1994). Tao Baiqiang appealed to LT administrators and practitioners to check and eliminate sexism representations in NMET (Tao, 2006b) and proposed that gender equity should be observed in curriculum design, textbook compiling, classroom instruction and evaluation.

4. Research methodology

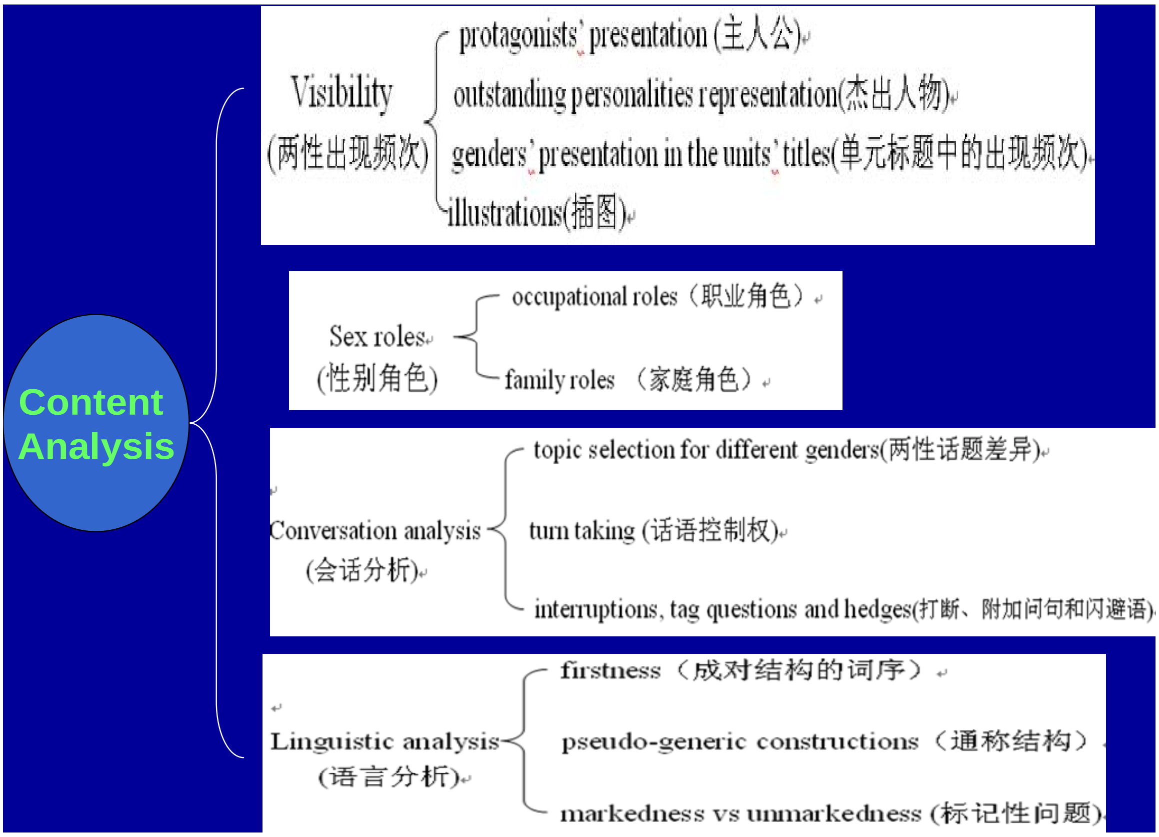
Subjects of the study: Listening comprehension section in NMET 2000-2005. There are two parts in the listening section:

Part 1: Listen to 5 short dialogues with 5 MC test items;

Part 2: Listen to 5 longer dialogues or monologues with 15 MC test items. In all 50 dialogues or monologues have been examined.

Note: the provincial NMET listening sections in years 2004 and 2005 are not included in the sample.

Research methods—content analysis: The potential sexism representation will be examined with the tool—content analysis (Tao, 2006a) which is employed in Tao’s thesis.


Content Analysis
Visibility
(两性出现频次)
protagonists' presentation (主人公)
outstanding personalities representation(杰出人物)
genders' presentation in the units' titles(单元标题中的出现频次)
illustrations(插图)
Sex roles
(性别角色)
occupational roles（职业角色）
family roles （家庭角色）
Conversation analysis
(会话分析)
topic selection for different genders(两性话题差异)
turn taking (话语控制权)
interruptions, tag questions and hedges(打断、附加问句和闪避语)
Linguistic analysis
(语言分析)
firstness（成对结构的词序）
pseudo-generic constructions（通称结构）
markedness vs unmarkedness (标记性问题)

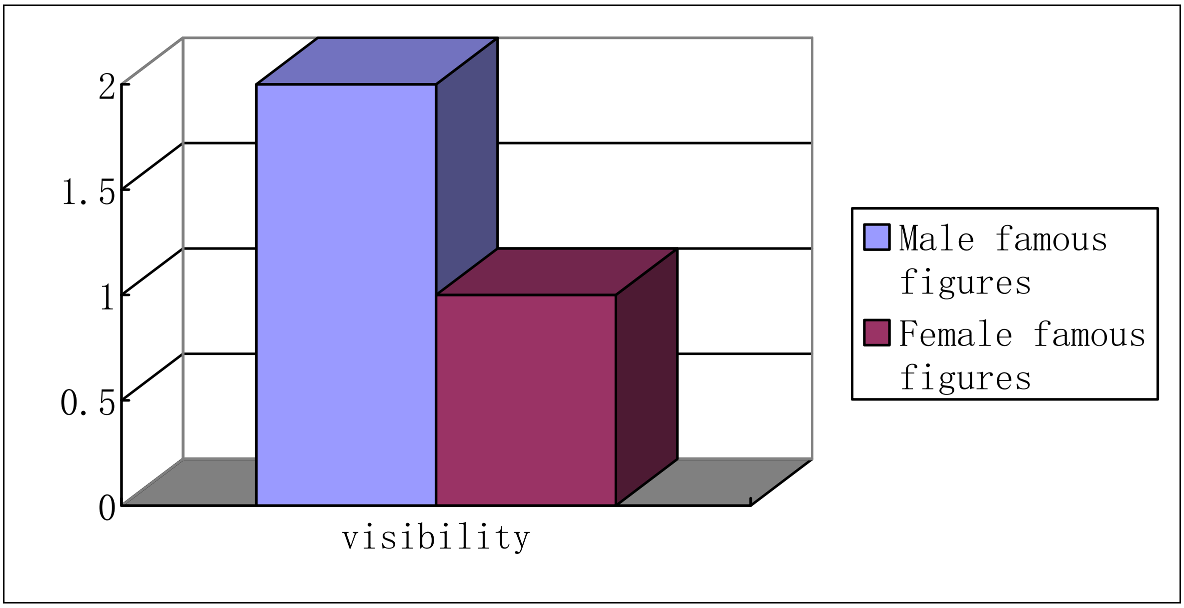
(Source: Tao, 2006a)

5. Data collection and analysis

5.1 Visibility/ Famous figures

Visibility refers to the frequency of female representation against male representation. Visibility is embedded in the following four facets: protagonists, historical figures assuming protagonists' roles or being mentioned, female /male appearance in texts and illustrations. Historical figures in textbooks facilitate students' sex role socialization and act as role models for students. Social psychology studies show that learners' imaginations are limited by the models they are presented with through texts. Porreca (1984: 706) states: 'When females do not appear as often as males in the text, the implicit message is that women's accomplishments … are not enough to be included.' She cites an earlier study by Hartman & Judd (1978) whose findings support this as well. 'In several of the texts reviewed, women suffered most obviously from low visibility (p. 384). Test materials to some extent serve as hidden curriculum for candidates.

In the sample, male famous figures have appeared twice, i.e. the writer Chris Paine and football star Michael Owen. In sharp contrast, female famous figure only appears once, i.e. the author of Cottage Garden Flowers, Margery Fish.



5. 2 Sex roles

5.2.1 Family roles

Sexism may be represented in family role distribution, for instance, the females are confined to family roles only as babysitter, dish washer, clothes washer, cook in the kitchen, etc. Women's stereotypical roles are related to housework and childcare. Women's household chores were exemplified by cooking, changing diapers, doing laundry whilst men were depicted fixing the car, changing electrical bulbs and/or mowing the lawn.

In the listening comprehension section some traditional and stereotyped female family roles have been frequently mentioned, such as housewife, shopper and laundry operator.

In Text 4, NMET 2002, a sexist word "housewife" is used. The use of *housewife* for people who worked in the home should be avoided. The *–wife* component refers only to married women, a spousal term suggesting marriage to the home, and as such derogates the nature of the occupation role and effectively excludes men. "Housewife" is suggested to be replaced by "stay at home parent/mother/mom or stay at home father/dad".

5.2.2 Occupational Roles

Male occupational roles cover a much wider range which are demanding, adventurous, high-paying, respectable, etc., such as explorer, physicist, politician, writer, poet, agriculturist, athlete, famous singer, film producer, musician, actor, etc. Conversely, female occupational roles are confined to such office-related clerks as typist, doctor, nurse, teacher, cook, cleaner, shop assistant, waiter, etc. and females appear in more derogatory roles than males.

In the sample, mentions of occupational roles for male and female are as follows:

Male occupations mentioned	Professor, teacher, driver, policeperson, correspondent, lawyer, manager, waiter, secretary, office worker
Female occupations mentioned	Shop/hotel waiter, ticket seller, shop assistant, teacher, gardener, manager

In conclusion, females are bestowed more waiter roles, although female manager and male secretary appeared in tapescripts in NMET 2003; and male clothes shop assistant in NMET 2005.

In addition, attributions refer to characteristics imposed upon females or males. Some feminine traits may appear as sexist attitudes against females, who are often described as incapable, dependent, gentle, effeminate, sissy, appearance sensitive, gossipy.

It is revealed in the sample that females were depicted uncertain, non-affirmative, sentences showing uncertainty (*I'm not sure; I haven't decided yet*) appeared three times, respectively in NMET 2002, 2003 and 2004 with no occurrences for males.

5. 3 Conversation analysis

Research into sexism with the conversation analysis tool refers to topic selection for different sexes, turn taking or conversational dominance, interruptions, tag questions and hedges, e.g. "…when males and females are together males interrupt women far more often than the other way round (Zimmerman & West, 1975, quoted in McCormick, 1994: 1375), one interpretation of these findings is that males use interruptions in order to assert their dominance, and in the absence of a better alternative we have to accept this (as cited in Hudson, 1996: 142) .

The survey of topic selection preference targets same-sex conversations but NMET listening segments are designed as cross-sex dialogues, then no results can be yielded in terms of topic differences. In addition, NMET listening materials are almost quasi-authentic instead of natural conversation samples with alterations to facilitate the testing purpose (Tao, 2006c), this study fails to find occasions of interruptions which are natural and frequent in authentic conversations. Turn taking principle is strictly observed in the NMET listening section.

5. 4 Linguistic analysis

Sexism is imperceptible in the language itself and is consequently, vital to analyze. Brown (1994: 240) asserts:

> "In English, another twist on the language and gender issue has lately focused us on "sexist" language: language that either causes unnecessary attention to gender or is demeaning to one gender—almost always women in today's only male dominated world. Writers are cautioned to refrain from using what we call the generic *he* and instead to pluralize or to use *he* or *she*. What used to be stewardesses, chairman and policeman are now more commonly referred to as flight attendants, chairs, and police officers."

Sexist language may be examined from the following three perspectives.

5.4.1 Firstness

Porreca (1984: 706) defined "firstness" as "Given two nouns paired for sex, such as male/female, the masculine word always came first, with the exception of the pair ladies/gentlemen" or other feminine domains such as *bride and groom, mother and father*. By placing the masculine pronoun in front of the female, male dominance may be displayed. This "reinforces the second place status of women and could, with only a little effort, be avoided by mixing the order" (Hartman & Judd, 1978: 390).

The NMET 2000 sample includes quite a lot of constructions representing firstness such as "a man and his wife, a man and his sister, a man and his girlfriend，the man and woman, John and his wife". The semantically asymmetric constructions "a man and his wife, John and his wife" blatantly hint that females are considered to be appendages to males. In NMET 2005 sample emerged “his and her”.

Ironically, in Dialogue No. 6, NMET 2003, the item writer put the female with a higher SES first in the MC stem: What is the possible relationship between ***the woman and the man*?** (Note: The dialogue is occurred between a female manager and a male employee) The order arrangement indirectly validated the stereotyped “firstness” principle.

The suggested solution is to mix the order or avoid firstness by using plural form.

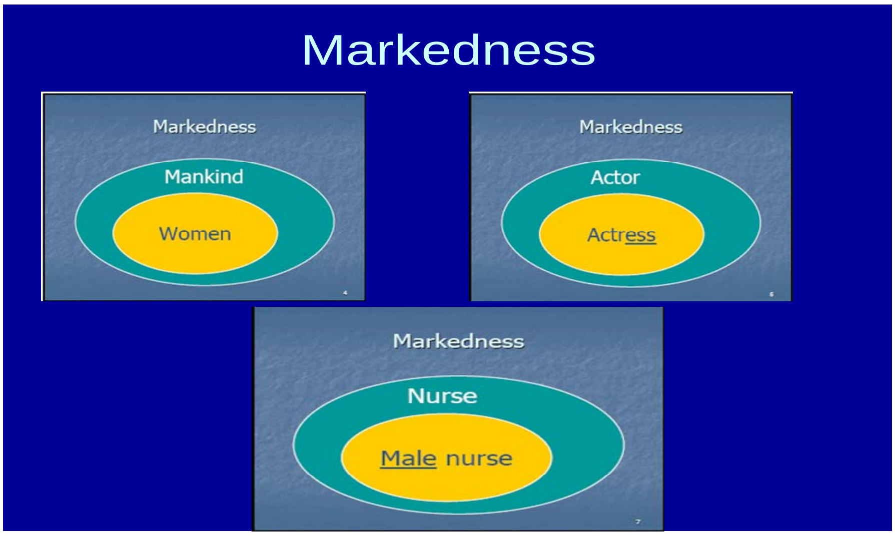
5. 4. 2 Pseudo-generic constructions

The term refers to those constructions which employ male-marked words when referring to both sexes with two typical examples, “he/him/himself” and “man”, for instance, “But man is not made for defeat”, “A man can be destroyed but not defeated” (Cited in the Nobel Prize winning novel *The Old Man and the Sea* by Ernest Hemingway). Peudo-generic constructions ignores the existence and contributions of females.

The survey has found one such generic noun “policeman” in Listening Text 10, NMET 2001, which should have been replaced by “policeperson”.

5. 4. 3 Linguistic asymmetry: Marked and unmarked

Professional titles mostly reflect males as the default or unmarked sex but are marked for females，for instance, woman driver is a marked professional title. The solution strategy includes the elimination or avoidance of *–man* compounds (in generic and sometimes in gender-specific contexts) by replacing them with existing non-gender-marked alternatives, e.g. *police officer* for *policeman*; or with a new *–person* compound, e.g. *chairperson* for *chairman*; or new alternatives, e.g. *flight attendant* for *hostess*/*steward*. Thus the *-person* compounds are examples of a gender-neutralization strategy.



（Figure 5.4.3）

The sample shows no representation of such marked nouns as actress , waitress.

6. Implications and counter measures against sexism in LT

Foreign theoretical research on sexism in LT has gradually been put into LT practice to ensure test validity and reliability. Kunnan (2000:3) proposed a new concept—equal construct validity intended to eliminate construct-irrelevant factors such as gender, race/ethnicity, filed of specialization, native language and culture. World-known testing services have stipulated test guidelines with a gender perspective, such as *ETS Standards for Quality and Fairness* (2002), *ETS Fairness Review Guidelines* (2003), *ETS*

International Principles for Fairness Review of Assessment (2004), *IELTS Handbook 2005*. International Language Testing Association (ILTA) passed *Code of Ethics for ILTA* in Canada in the year 2000 and *The Draft ILTA Code of Practice* in Australia in 2006. The author proposes that three measures be taken to step towards sexism-free and fairer language testing.

(a) Legal measures: The legislature should enact laws to defense gender equality in micro-education environment including curriculum design, EFL textbook compiling, classroom instruction and language evaluation. There is no legislation stipulating national examinations and eliminating sexism in society as a whole.

(b) Professionalization of language testing: Item writing quality in NMET is worrisome (Tao, 2007) and needs improving, ensuring from correctness to fairness. Professionalization of LT requires a code of language testing ethics and practice, experiences and expertise in test item writing and a scientific test review process. There is conflict between pretests for statistics to help maximize reliability, validity and fairness on the one hand and the need for test security on the other for very high stakes testing and then a defensible test review process is crucial and urgently needed to obliterate problematic items and remove sensitivity including sexism. External review process is preferred for objectivity.

ETS has developed a comprehensive test review process (Peirce，1994: 45-47): a series of test reviews by approximately six different test development specialists, after the Test Specialist Reviewer (TSR) is satisfied with the test, the test goes to TOEFL coordinators, two of whom are responsible for its style and one is responsible for any potential sensitivity including sexism. Peirce once changed the word "businessman" in the following sentence into "businessperson" ("Therefore many farmers are tenants and much of the land is owned by banks, insurance companies, or wealthy *businessmen*".)

We should establish comprehensive and sound test review processes to ensure test quality based on foreign counterparts' practice, our own research and EFL context which requires including native speakers of English and NS of English LT experts in the test review panel.

(c) Training: Training is a must for LT administrations and item writers especially in China's present LT operation practice. For the sake of security requirement and the extreme high stakes of some English tests in China, high expectations of test quality is demanded. Quite a lot of item writers (mainly university professors) are temporarily organized about two months before the administration of the test to prepare a large scale test. Time constrains, LT expertise and English language proficiency inadequacy and inexperience may lead to test quality disaster, which need to be ameliorated urgently.

★**End note: Definition of sexism:** Sexism includes blatant, covert, and subtle sexism (Benokraitis & Feagin, 1999, as cited in Swim, Mallett & Stangor, 2004):

1. Blatant sexism: obviously unequal and unfair treatment of women relative to men;

2. Covert sexism: unequal and unfair treatment of women that is recognized but purposefully hidden from view. Both blatant and covert sexism are intended, but only covert sexism is hidden;

3. Subtle sexism: subtle sexism represents unequal and unfair treatment of women that is not recognized by many people because it is perceived to be normative, and therefore does not appear unusual. Thus, like covert sexism, subtle sexism is hidden but

unlike covert sexism, subtle sexism is not intentionally harmful (Swim, Mallett & Stangor, 2004).

Holmes (1996:336) defines *sexism* as follows "… the ways in which language conveys negative attitudes to women", in this sense Holmes defines sexism only in terms of language.

References

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

Bolger, N. & Kellaghan, T. (1990). Method of measurement and gender differences in Scholastic Achievement. *Journal of Educational Measurement*, 27(2), 165-174.

Brantmeier, C. (2003). Does gender make a difference? Passage content and comprehension in second language reading. *Reading in a Foreign Langua*ge, 15(1), 1-23.

Brantmeier, C. (2004). Gender, violence-oriented passage content and second language reading comprehension. *The Reading Matrix*, 4 (2), 1-19.

Brown, A. & McNamara, T. (2004). "The devil is the detail": Researching gender issues in language assessment. *TESOL Quarterly,* 38 (3), 524-538.

Brown, H. D. (1994). *Principles of Language Learning and Teaching* (3rd ed.). NJ: Prentice Hall.

Childs, R.A. (1990). Gender Bias and Fairness [Electronic version]. Washington DC: ERIC Clearinghouse on Tests Measurement and Evaluation. Retrieved August 16, 2007, from http://www.ericdigests.org/pre-9218/gender.htm.

Donlon, T. F. (1973). Content factors in sex differences on test questions. *Research Memorandum*. Princeton, NJ: Educational Testing Service.

Dwyer, C. A. (1976).Test content and sex differences in reading. *The Reading Teacher*, 29: 20-24.

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly*, *2*(3), 197–221.

Eels, K., Havighurst, R.J., Herrick, V.E., & Tyler, R.W. (1951). *Intelligence and cultural differences.* Chicago: University of Chicago Press.

Gershuny, H. L. (1977). Sexism in Dictionaries and texts: Omissions and commissions. In A. P. Nielsen et al (Ed.), *Sexism and Language* (pp.161-179). Illinois: National Council of Teachers of English.

Greatorex, J. & Bel, J. F. (2002). Does the gender of examiners influence their marking? Paper presented at the conference *Learning Communities and Assessment Cultures: Connecting Research with Practice*. Organised by the EARLI Special Interest Group on Assessment and Evaluation and the University of Northumbria.

Han, Baocheng. (2003). On Task-based language assessment. *Foreign Language Teaching and Research*, 5, 352-358. (In Chinese)

Hartman, P.L. & Judd, E.L. (1978). Sexism and TESL materials. *TESOL Quarterly*, 12, 383-393.

Holmes, J. (1996). *An Introduction to Sociolinguistics.* England: Longman Group UK Limited.

Hudson, R. A. (1996). *Sociolinguistics* (2nd edition). Cambridge: Cambridge University Press.

Hyde, J. S. & Linn, M. C. (1988). Gender differences in verbal activity: A meta-analysis. *Psychological Bulletin*, 104: 53-69.

Jie, L. & Fenglan, W. (2003). Differential performance by gender in foreign language testing. Chicago: Paper presented at the 2003 annual meeting of NCME.

Kunnan, A. J. (2000). Fairness and justice for all. In A. J. Kunnan (Ed.), *Fairness and Validation in Language Assessment: Selected Papers from 19th Language Testing Research Colloquium, Orlando, Florida* (pp.1-14). Cambridge: Cambridge University Press.

Li, Xiaoju. (2001). *The science and art of language testing*. Changsha: Hunan Education Press. (In Chinese)

Liu, Qingsi. (1993). Reading comprehension in national matriculation English tests. *Research on Tests*, 382, 1-3. (In Chinese)

O'Loughlin, K. (2002). The impact of gender in oral proficiency testing. *Language Testing,* 19（2）, 169-192.

Peirce, B. N. (1994). The Test of English as a Foreign Language: developing items for reading comprehension. In Hill C & Parry K (Eds.), *From Testing to Assessment: English as an International Language* (pp. 39-60). New York: Longman.

Porreca, K. L. (1984). Sexism in current ESL textbooks. *TESOL Quarterly*, 18(4): 705-724.

Ryan, J.M., & Demark, S. (2002). Variation in achievement scores related to gender, item format, and content area tested. In G. Tindal & T.M. Haladyna (Eds.), Large-scale assessment programs for all students: Validity, technical adequacy, and implementation (pp67-88). Mahwah, New Jersey: Lawrence Erlbaum Associates, Publishers.

Scheuneman, J. D., & Slaughter, C. (1991). Issues in text bias, item bias, and group differences and what to do while waiting for answers. Report: Evaluative/feasibility. *ERIC* ED 400 294.

Shi, Jinghuan. (2000). Feminist Pedagogy and Its Practice in Western Countries. *The Journal of Shanxi Teachers University*, 27(3), 5-10. (In Chinese)

Sunderland, J. (1994). Gender and language testing: any connection? preliminary explorations. *Language Testing Update*, (16, 46-56.

Sunderland, J. (1995). Gender and Language Testing. *Language Testing Update*, 17, 24-35.

Sunderland, J. (2000). State of the art review article: Gender, language and language education. *Language Teaching*, 33（4）, 203-223.

Swim, J. K. , Mallett, R. & Stangor, C. (2004). Understanding subtle sexism: detection and use of sexist language. *Sex Roles,* 51(3-4), 117–128.

Tae-Il Pae. (2002). *Gender Differential Item Functioning on a National Language Test (Korea).* Doctoral dissertation of Purdue University (Advisor: Susan J. Maller and Beverly E. Cox)

Takala, S. & Kaftandjieva, F. (2000). Test fairness: a DIF analysis of an L2 vocabulary test. *Language Testing,* 17(3), 323-340.

Tao, Baiqiang. (2006a). *A Probe into Sexism in China's EFL Textbooks for Senior High School Students*. Chongqing: China Southwest university MA thesis (Advisor: Professor Chen

Zhi'an).

Tao, Baiqiang. (2006b). Revision suggestions on National English Curriculum for High School Students of China. *Foreign Language Teaching & Research in Basic Education*, 3, 56-57. (In Chinese)

Tao, Baiqiang. (2006c). National Matriculation English Test (NMET) listening evaluation and item writing. *The Journal of Test Research*, 8, 1-4. (In Chinese)

Tao, Baiqiang. (2007). On improving test item writing art in NMET—A case study of NMET 2006 multiple choice test items. *English Teaching & Research Notes*, 3, 45-51. (In Chinese)

Wood, R. (1993). *Assessment and Testing: A Survey of Research.* Cambridge: Cambridge University Press.

Wu, Zunmin. (2002). *English language testing: from theory to practice*. Beijing: Foreign Language Teaching and Research Press. (In Chinese)

Xiao, Defa & Xiang, Ping. (2005). Gender Differences in PETS Oral Test. *Shandong Foreign Languages Teaching Journal*, 1, 54-56. (In Chinese)

Zhang, Bin & Du, Cuiqin. (2004). Correlation between gender and English learners' performance. *Social Sciences*, 19(1), 127-128. (In Chinese)

Zhou, Sheng & Fu, Jihua. (1994). Test and anxiety. Foreign Language Teaching and Research, 3, 43-46. (In Chinese)

Zou, Shen. (1998). *English Language Testing–Some Theoretical and Practical Considerations*. Shanghai: Shanghai Foreign Language Education Press. (In Chinese)

Zou, Shen. (2005). Understanding the washback effect of tests—With special reference to the revision of the TEM 4/8 test battery. *Foreign Language World*, 5, 59-66. (In Chinese)