Scoring Fairness in China

Yi Mei

Queen's University

Author Note

Correspondence concerning this article should be addressed to Yi Mei, Faculty of Education, A106 Duncan McArthur Hall, Queen's University, Kingston, Ontario, Canada, K7M 5R7. E-mail: yi.mei@queensu.ca

Abstract

Test fairness is a contested topic in the field of language testing in recent decades. Outside of North America, there is currently little research on scoring fairness in educational contexts. Taking fairness issues involved in scoring in the Chinese context as an exemplar, this study examines how the notion of fairness has evolved, and how test fairness, especially scoring fairness, is defined and empirically investigated in China. The analysis shows that the Chinese notion of fairness can be traced back roughly 2,500 years, to the time of Confucius. Since the era of imperial examinations, test fairness has remained a controversial topic regarding whether it is a purely educational issue or should be considered with its social implications (Yang & Gui, 2007). In China, the modern research on test fairness started from the late 1990s under the influence of Western measurement theories (Cao & Zhang, 1999). The findings also reveal three major sources of influence on how the Chinese academia views test fairness. Apart from the indigenous definition of test fairness, fairness theories in the field of language testing and education have exerted considerable influence on recent scholarship. However, research on scoring fairness is not well supported with empirical evidence, which is consistent with the status in the language testing field elsewhere. This study will have implications for endeavors involving localized research approaches to guiding scoring fairness research and practices in China as well as in other educational contexts.

Key words: test fairness, rater judgment, writing assessment, literature review, China

Introduction

Test fairness is a contested issue in the field of language testing in recent decades (Kunnan, 2008). The definition of fairness and its relationship with validity are among the most controversial topics (e.g., AERA, APA, & NCME, 1999; Gipps & Murphy, 1994; Kunnan, 2000, 2004; Zieky, 2006). Although these researchers and institutions support different perspectives regarding the nature of fairness, they all agree that fairness must be ensured at every stage of test development, from test design through to scoring stage (Kunnan, 2008). Test fairness has been well studied in language testing research on biased test items against test takers with different knowledge backgrounds (Clapham, 1998) or from different language groups (Stricker, Rock, & Lee, 2005), and unfairness in oral interview marking (Brown, 2003. However, fewer studies have been conducted on test fairness in scoring written performance. Guo (2009) and Huang (2011), for example, examined scoring fairness in the North American context, where test takers were grouped into heterogeneous subpopulations from varying cultural and linguistic backgrounds. Whereas fairness is "subject to different definitions and interpretations in different social and political circumstances" (AERA et al., 1999, p. 80), test takers in China tend to share a relatively homogeneous cultural and linguistic background.

China has a history of testing which can be traced back over 2,000 years ago when the first civil examination[1] was held to select officials for the emperor's

1 Civil examination: The test takers of civil examinations were recommended by local governments to the

administration (Zhou & Shen, 2006). This tradition of using tests for selection purposes is still influential in the modern-day educational system in China (Cheng, 2008), such as the national university entrance examinations (also known as Gaokao) for university admission decisions. The high stakes nature of the selection tests has raised considerable concerns over fairness issues in the Chinese testing system (Zhou & Shen, 2006). China's historical and cultural relationship to testing makes it a valuable case to examine scoring fairness issues.

This study aims to understand how scoring is addressed in education in China[2] through a literature review. The paper is divided into three main sections. First, I will describe considerations regarding research method of this study; the fairness discourse then will be analyzed under three themes emerging from the findings; lastly, I will discuss issues identified in relation to test scoring fairness research in China.

Method

The Research Questions

Three research questions were addressed: (1) How does the notion of test fairness evolve in China? (2) What is the nature of test and scoring fairness? (3) How is test and scoring fairness empirically investigated?

central government, and then after passing the exams, the qualified test takers were appointed to a position in the government.

2 In this study, China is the shortened form of People's Republic of China, which is often called Mainland China.

Data Collection

The data was collected from the China Academic Journals Full-text Database (中国期刊全文数据库: www.cnki.net), accessed via Queen's University library system, using the following search words: "test fairness", "assessment fairness", "scoring fairness", "marking fairness", "考试公平" (test fairness), "测试公平" (assessment fairness), "考试公正" (test fairness), "测试公正" (assessment fairness), "评分公平" (scoring fairness), "评分公正" (scoring fairness). Due to the various expressions of fairness in Chinese, a "snowball" approach was adopted to follow up the reference lists of the searched articles to ensure that the issue-related articles were included to the extent possible.

A total of 96 articles, all written in Chinese, were retrieved from the database. Of these, 37 addressed issues of test fairness. The articles feature topics generally consistent with the discussion in the mainstream language testing academia (e.g., absence of bias, equity in test administration and testing outcomes). The other 59 articles feature general discussion of equality in educational testing (e.g., justice in the testing system, minimizing variation in test consequences across regions), which is considered test fairness in a broad sense in the Chinese academia. Scoring fairness is mentioned or addressed in a small number of articles, which is consistent with the underdeveloped status of this research area in the language testing field. All the retrieved articles were published no later than 2011. Among them, 75 articles were published between 2007 and 2011, with two published as early as in 1999.

Due to the unique publication culture of foreign language education journals in Mainland China (Shi, Wang, & Xi, 2005), most of the articles collected for this study were not academic publications. These articles, written by scholars, teachers, policy makers and practitioners in China, are non-research-based, sometimes repeated, general introductions to test fairness issues, comparison between foreign and Chinese interpretations of test fairness, and expressions of personal views; some of them are one to two pages long; very few of them are empirical studies.

Nevertheless, this is a collection worthy of attention for three reasons. First, almost all the mainstream research-based education journals in China have kept silent on the issues probably because test fairness is often considered a socio-political rather than an educational concept in China, which demonstrates the uniqueness of the Chinese context. Approximately 30% of the articles in this study discussed test fairness in Gaokao, which is often analogized to Keju in ancient China because of their shared significance of shaping the test takers' life options. To ensure fairness of Gaokao is regarded a political-oriented accomplishment and a showcase for social equity (Wang & Zhang, 2007). Second, there has been a surge of research interest in test fairness since 2007, which may indicate a precursor of localized test fairness research among Chinese researchers. This collection is reflective of contemporary test fairness research in China. Third, an analysis of Chinese articles into English provides non-Chinese readers with access to key issues on test fairness in China, promoting further dialogue internationally.

The Analytical Procedures

General principles of grounded theory (Charmaz, 2005; Creswell, 2007) are followed in the analysis and the data was coded in a sequence system. Through an iterative process between the research questions and the data, the collected articles were read through several times to locate the thematic categories and relationships between them. First-level codes were given to each article (and sometimes to paragraphs if an article contained more than one theme), leading to five categories:

1. Review of test fairness with reference to foreign models;
2. Review of test fairness with reference to ancient Chinese testing systems;
3. Discussion (mostly from a socio-political perspective) of general fairness issues;
4. Discussion of issues involved in implementing test/scoring fairness;
5. Reports of test/scoring fairness practices in language tests

Based on a frequency count of the recurring themes in the articles, these broad categories were then combined into three second-level categories in relation to the research questions: the first theme, evolution of test fairness in China, was gathered in the original category 2; the second theme, nature of test and scoring fairness, was gathered across the first four original categories, as most of the retrieved articles contained sentences or paragraphs on definitions or illustrations of test and scoring fairness; the third theme, empirical studies on test and scoring fairness, was gathered in the original

category 5. After an integration of the three second-level categories, a general concept of fairness emerges as a focal point, which demonstrates the core of the categories.

Analysis

This section reports on the findings of the three research questions, namely: (1) evolution of test fairness in China; (2) nature of test and scoring fairness; and (3) empirical studies on test and scoring fairness.

Evolution of Test Fairness in China

The notion of fairness emerged as a key term in the socio-political field in China approximately 2,500 years ago. In the Spring and Autumn Period (BC770 - BC 476), the Chinese philosopher Confucius first proposed the concept of using “benevolence”, “righteousness”, “rites”, and “music” to construct a harmonious society and maintain social order (Jin, 2011). Among the four fundamental cornerstones, “benevolence” calls for fairness and justice in establishing social norms. Compared with the western notion of fairness, the notion of fairness in ancient China is integrated with the ideas of “justice” and “harmony” at the institutional level. It does not show much concern on the interests of individuals in the process of establishing social norms and securing fairness, while it focuses mainly on conforming to righteousness of the established institutional norms, which has exerted direct influence on the imperial examination system (Jin, 2011).

The imperial examination system broke the boundaries of social classes, which is a breakthrough for the rigidly stratified feudal China and considered a fair practice at that

time, thus having a far-reaching influence on the Chinese society. The imperial examination allowed free registration, and selection decisions were made based on test takers' performance irrespective of their social classes, which gave test takers from the lower class an opportunity to change their lives through their efforts (Liu, 2007). The imperial examination system developed a set of rigid procedures to ensure test and scoring fairness, such as courtyard locking system[3], concealing test takers' names, transcribing answer sheets, and proofreading[4] (Jin, 2011; Liu, 2007).

Research on test fairness in modern China started in the late 1990s (Dai, Wei, & Liu, 2010) from introducing Differential Item Functioning (DIF) to the Chinese academia (Zeng & Meng, 1999) and examining presence of item bias in a Chinese vocabulary test among Ninth Graders from urban and rural areas (Cao & Zhang, 1999). Fairness issues involved in the national university entrance examinations have attracted considerable attention (e.g., Bai, 2011; Li, 2010; Liu, 2007; Zhou & Shen, 2006). Regional parity and fairness in test outcomes are conflicting stances regarding fairness in the university enrolment system. Advocates of fairness in test outcomes contend that universities should make enrollment decisions based on test takers' performance, while supporters of regional parity protest that measures should be taken to regulate the enrollment quotas

[3] Courtyard locking system: The imperial examiner, once appointed, must be immediately locked in certain places and separated from the outside world for about 50 days, during which period the examiner was not allowed to go home or visit anyone outside the place. This measure was taken to preserve test confidentiality.

[4] Proofreading: After transcribing test takers' answer sheets, highly literate officials were hired to proofread the transcriptions and correct typos.

between regions to avoid great regional disparity.

Nature of Test Fairness and Scoring Fairness

Three recurring sources of defining test fairness are identified in the retrieved articles, which can explain what influences the conceptualization of test fairness in China. First, "all test takers are equal before test scores". Not only is this indigenous perspective in test fairness considered the golden rule of language testing (Gui, 2011), but it is also dominantly portrayed in approximately 31% of the retrieved articles. It is widely believed in China that test fairness must be ensured if test results are fair and objective and test takers are equally treated (Gui, 2011). However, for centuries there has been a controversy over test fairness concerning high-stakes selection decisions. Debate about this issue started to appear in the literature from sources published on Keju since the Northern Song Dynasty (960-1127) and argument continues in current usage of Gaokao (Yang & Gui, 2007; Zhou & Shen, 2006).

Advocates of test fairness in form assert that admission decisions should be made based solely on test results, regardless of geographic regions; however, believers of test fairness in fact argue that given the imbalance of regional development and ethnic diversity in China, each region should have its own quotas for enrollment decisions, so as to promote social equity and narrow the education gap between developed and underdeveloped areas by stabilizing a certain enrolment rate in each region (Bai, 2011). Test fairness in form is in alignment with the notion of "all test takers are equal before

test scores", while test fairness in fact takes regional imparity into consideration and calls for individualized enrollment policies in different regions to achieve general test fairness (Li, 2010). Test fairness in form is a consideration at the individual level, while test fairness in fact is at the collective level.

Second, test fairness defined in the field of language testing in other contexts, especially in the American context, is referenced and taken as exemplars in 20% of the articles reviewed. Among all of the cited definitions and frameworks, the four defining characteristics of test fairness in the *Standards* (AERA et al., 1999) are the most frequently quoted: absence of bias, equitable treatment of test takers in the testing process, equality of testing outcomes for different groups of test takers, and equity in opportunity to learn the measured content. The 1999 *Standards*, various versions of the *ETS Standards for Quality and Fairness* (2002, 2009a, 2009b), the *Code of Fair Testing Practices in Education* by Joint Committee on Testing Practices (2004), Kunnan's (2000, 2004) test fairness framework, and Zieky's (2006) summary of test fairness were well documented (e.g., Jiang, 2007; Lu, 2011; Xie & Wang, 2002). As an aspect of test fairness, the concept of test bias (Brown, 2004; Cole & Moss, 1989) is also introduced (e.g., Lu, 2011; Xie & Wang, 2002). The American history of exploring test fairness is summarized into three focuses: fairness in beneficiary, in test content, and in test outcomes (Dai et al., 2010). Instead of repeating those foreign definitions of test fairness, some researchers interpret the definitions within the Chinese context. The notion of

fairness in test outcomes, which suggests comparable test outcomes across test taker groups (AERA et al., 1999), is enriched by considering fairness in test use and avoiding abuse of test outcomes (Dong & Ma, 2011). Acknowledging ETS' emphases on quality and fairness, fairness is also interpreted as test outcomes reliable enough to be the foundation for distribution of educational resources, showcasing educational and social equity (Jiang, 2007). Although these Chinese researchers show more concerns about the social dimension of test fairness, few argue that fairness should be about testing itself and that maintaining social equity is an unrelated construct (e.g., Xie & Wang, 2002).

Third, theories on educational equality are another source of exploring test fairness in approximately 8% of the issue-related articles. The widely acknowledged theories of Husén's (1975) educational equality and Coleman's (1968) conception of equality in educational opportunity have an impact on understanding test fairness (e.g., Dai et al., 2010). Husén's theories acknowledge equality in access (educational opportunities), procedures (comparable school conditions and equitable opportunities for education process), and results (attainment and life chances) (Dai et al., 2010).

Some Chinese researchers take a similar approach as Husén's (1975) in that they interpret test fairness as equity of procedures, conditions, and equity in fact (e.g., Xie, 2004). Equity of procedures, also called the first fairness, requires that all test takers be equally treated; equity of conditions, the second fairness, requires that in addition to equal treatment, all test takers have equal access to educational resources; equity in fact,

the third fairness, requires that in addition to equal treatment and access all children be granted equal rights to have quality educational resources, which is the ultimate goal of testing professionals (Xie, 2004). The first fairness is similar to Husén's (1975) equality in procedures; the second fairness is similar to equality in access; and the third fairness is similar to equality in results. However, unlike Husén's theories, Xie's classifications of fairness are each achieved on the basis of the previous ones.

Unlike the systematic review and discussion on test fairness, fairness issues involved in scoring written performance are scattered in 12% of the articles reviewed, often found in a few sentences or in a paragraph. In the era of Keju, such measures as concealing test takers' names, transcribing answer sheets, and proofreading were taken to ensure that answer sheets were scored fairly and test results were not influenced by anything other than the content of test takers' responses (Jin, 2011). During the Ming (1368-1644) and Qing dynasties, test takers of the imperial examinations were required to write essays in eight sections, which made the scoring criteria easier to apply and achieved better fairness (Jin, 2011).

In modern educational research, scoring fairness is considered either as a reliability issue, a validity issue, or a rater bias issue. In the issue-related articles, scoring fairness is dominantly believed to be achieved if scoring errors are minimized (Zhou, Ding, Zhang, & Wen, 2010). Fairness is undermined in writing assessment when raters' scoring decisions are influenced by different interpretations of scoring rubrics between

raters and between regions, raters' levels of expertise, rating experience, psychological conditions, test takers' handwriting style, marking equipment, and workplace environment (Wang, 2011), and sometimes by test takers' appearance and personalities in the case of classroom writing assessment (Yang, 2001).

Scoring fairness is considered an interchangeable term with rater bias and an integral part of test validation (Lu, 2011). Rater bias is the interaction between raters and other facets of scoring such as written texts (e.g., text content, structure, rhetoric), writing tasks, rating scales, test taker backgrounds (e.g., gender, ethnicity, education), rater's reading style, scoring focuses, scoring strategy, rater background (e.g., gender, cultural, and linguistic backgrounds), and scoring sessions (Lu, 2011). Zou (2011) believes scoring fairness is guaranteed if scoring validity is ensured. Therefore, scientific, comprehensive, and viable scoring rubrics, representativeness of anchor samples, and rater training and moderation are crucial to improve scoring validity, hence scoring fairness (Zou, 2011).

Empirical Studies on Test and Scoring Fairness

Contrasting with the predominance of review and discussion articles on fairness, very few empirical studies (approximately 6% of all the articles reviewed) have been reported, which is consistent with Dai et al.'s (2010) conclusion that fairness research in China is relatively rich in theoretical and analytical papers but short of empirical evidence. The empirical research focuses on Differential Item Functioning (DIF) analysis of test items to detect potential test bias (Cao & Zhang, 1999); equating two sets of test

scores in one national English examination to determine comparability of test results in different forms of the examination (Han, 2004); comparing enrollment opportunities between regions to probe into regional disparity (Li, 2010); investigating public opinions on fairness issues in educational tests and testing system (Chen, Wang, Song, Yang, Zhang, & Wang, 2009); and reporting fairness issues involved at each stage of a national examination (Zhou et al., 2010).

Discussion

The above review suggests that test and scoring fairness has unique characteristics in China and should be interpreted in terms of the specific social, cultural, and historical influences in the Chinese context. Unlike the efforts to ensure test fairness in North America (e.g., ETS, 2002, 2009a, 2009b; Joint Advisory Committee, 1993; JCTP, 2004), in Europe (e.g., ALTE, 1994, 2001), and in Japan (JLTA, 2006) where fairness issues focus on the test itself, test fairness issues in China focus primarily on the educational testing system, and especially on the enrolment system of Gaokao. In other words, test fairness in China is more than an educational concept and is accompanied with social concerns.

Whether test fairness should remain an educational concept or be considered along with its socio-political implications remains a debatable question for Chinese education researchers. China is a country featuring highly regulated state power with a long history of national testing. The controversy between regional parity and fairness in

test outcomes has never stopped since Keju in Northern Song Dynasty. China is a vast multi-cultural and multi-ethnic nation, where the socio-economic and cultural resources are unevenly distributed. Setting different cut-off scores in favor of economically and culturally developed regions has aroused widespread concerns over inequality in university admission opportunities (Zhou & Shen, 2006). The tilted cut-off Gaokao scores result from the imbalance and exacerbate the tension in turn.

In spite of the history of testing, modern psycho-metrics and standardized language testing theories were not introduced into China until the 1980s (Gui, 1989). Modern language testing research in China is still in its infancy (Gao, 2006). The connotation and nature of test and scoring fairness are in need of continuing exploration with combined considerations of both the Chinese socio-economic situation and cultural traditions, and the modern language testing theories.

In addition to further research on fairness, there is a call for establishing guidelines for fair testing practices. Fairness standards and review system are in urgent need to be established in order to ensure quality and fairness of language tests during test development (Zhou et al., 2010). As language testing practitioners' professionalism is still in need of improvement in the country (Mei & Nie, 2009), such guidelines will facilitate understanding and dissemination of fair testing principles and therefore promote development of systematic theory building and good testing practices. Besides, establishment and implementation of fairness guidelines cannot be successful without

support from authoritative institutions or language testing associations. The blurred boundary between testing itself and its social consequences is responsible for the limited understanding and research on educational testing fairness; the root lies in absence of localized test fairness standards and theories (Dai et al., 2010). This is the major challenge facing educational testing fairness research in China today.

Conclusion

This study examines how scoring fairness issues are addressed in the Chinese context, by reviewing the Chinese publications on test and scoring fairness and summarizing evolution of the concept of test fairness, nature of test and scoring fairness, and empirical research conducted on exploring test and scoring fairness in China. The analysis shows that since Confucius initiated the concept of fairness in China approximately 2,500 years ago, fairness has become an important notion with socio-political implications. In China, test fairness is often more than an educational issue. Whether test fairness should be discussed within the frame of education or be considered with its potential social consequences has remained a controversy since the Keju era for over 1,000 years.

This socio-cultural influence accounts for the Chinese approach to test fairness research, which contrasts that in North America, where test fairness research is measurement-driven and aims to maintain educational equity to different test taker groups regardless of linguistic and cultural backgrounds. Test fairness in China has been given

considerable social meanings as a tool to maintain not only educational equity but also social equity and stabilization, shaped by its unique history and culture for centuries. The findings support the *Standards*' (AERA et al., 1999) claim for the variety of test fairness in different contexts.

In spite of the social acceptance of the indigenous perspective regarding test fairness, in recent years some Chinese researchers have started to adopt fairness theories in the international academia of language testing and education to inform systematic research on test and scoring fairness. The limited evidence available on scoring fairness is consistent with the status of this line of scholarship in the international language testing field. The increasing interest in test and scoring fairness since 2007, however, is indicative of an emerging trend of attention to test and scoring fairness in this country.

Future research on scoring fairness in China should explore localized research approaches to guiding scoring fairness research and practices. This study also encourages examination of scoring fairness issues in other educational, socio-cultural contexts to explore localized interpretations and applications of scoring fairness. Therefore, this study will be of interest to Chinese educational researchers and practitioners as well as researchers who are interested in understanding fairness issues in a localized setting such as China. In spite of the rich information obtained from the literature, discussion of fairness issues is restricted mostly to nonacademic publications, as the publication culture in China tends to shy away from these issues. Nonetheless, this collection reflects the

status of fairness research in China and hence should not be disregarded on the basis of the literature quality.

References

American Educational Research Association (AERA), American Psychological Association (APA), & National Council for Measurement in Education (NCME). (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Association of Language Testers in Europe (ALTE). (1994). *The ALTE code of practice.* Retrieved from Association of Language Testers in Europe website: http://www.testdaf.de/institut/pdf/ALTE/ALTE_code_of_Practice_Einleitung_EN.pdf

Association of Language Testers in Europe (ALTE). (2001). *Principles of good practice for ALTE examinations.* Retrieved from Association of Language Testers in Europe website: http://www.alte.org/attachments/files/good_practice.pdf

Bai, L. (2011). 我国科举录取名额分配制度的历史与反思——兼谈我国高考录取中的考试公平与区域公平 [History and reflection on China's imperial examination enrolment quota distribution system: Fairness in test outcomes and regional parity in Gaokao enrolment system]. *Educational Innovation*, (6), 6-7.

Brown, A. (2003). Interviewer variation and the co-construction of speaking proficiency. *Language Testing, 20,* 1-25.

Brown, D．(2004). What do we mean by bias，Englishes，Englishes in testing, and English language proficiency? *World Englishes, 23,* 317-319．

Cao, Y. & Zhang, H. (1999). Detection of differential item functioning in a Chinese vocatulary test. *Acta Psychologica Sinica, 31,* 460-467.

Charmaz, K. (2005). Grounded theory in the 21st century: Applications for advancing social justice studies. In N. K. Denzin & Y. S. Lincoln (Eds.), *The Sage handbook of qualitative research* (3rd ed., pp. 507–536). Thousand Oaks, CA: Sage.

Chen, Z., Wang, Y., Song, Z., Yang, W., Zhang, S., & Wang, X. (2009). Investigation and analysis on fairness of examination. *Journal of Gansu Normal Colleges, 14*(1), 72-75.

Cheng, L. (2008). The key to success: English language testing in China. *Language Testing, 25,* 15-37. doi:10.1177/0265532207083743

Clapham, C. (1998). The effect of language proficiency and background knowledge on EAP students' reading comprehension. In A. J. Kunnan (Ed.), *Validation in language assessment* (pp. 141-168). Mahwah, NJ: Lawrence Erlbaum.

Cole, N. S., & Moss, P. A. (1989). Bias in test use. In R. L. Robert (Ed.), *Educational measurement* (3rd ed.,pp. 201-219). New York, NY, England: Macmillan Publishing Co, Inc; American Council on Education.

Coleman, J. S. (1968). The concept of equality of educational opportunity. *Harvard Educational Review*, *38*(1), 7-22.

Creswell, J. W. (2007). *Qualitative inquiry and research design: Choosing among five approaches* (2nd ed.). Thousand Oaks: Sage Publications Inc.

Dai, J., Wei, X., & Liu, F. (2010). 教育考试公平性的基本理论研究 [Fundamental theoretical research on educational test fairness]. *China Higher Education Research,* (8), 27-29.

Dong, S., & Ma, S. (2011). Fairness analysis on assessment with score report from measurement perspective. *Examinations Research,* (1), 59-64.

Educational Testing Service (ETS). (2002). *ETS standards for quality and fairness.* Retrieved from Educational Testing Service website: http://www.ets.org/Media/About_ETS/pdf/standards.pdf

Educational Testing Service (ETS). (2009a). *ETS guidelines for fairness review of assessments*. Retrieved from Educational Testing Service website: http://www.ets.org/Media/About_ETS/pdf/overview.pdf

Educational Testing Service (ETS). (2009b). *ETS international principles for fairness review of assessments.* Retrieved from Educational Testing Service website: http://www.ets.org/Media/About_ETS/pdf/frintl.pdf

Gao, S. (2006). Historical development and reform trend of language testing. *Journal of Northwest University (Philosophy and Social Sciences Edition), 36*(4), 155-158.

Gipps, C. V., & Murphy, P. (1994). *A fair test? Assessment, achievement and equity*. Buckingham: Open University Press.

Gui, S. (1989). Language testing: New technology and theory. *Foreign Language Teaching and Research,* (3), 2-10.

Gui, S. (2011). 语言测试的黄金法则 [The golden rule of language testing], *Foreign Language Testing and Teaching,* (1), 6-8.

Guo, F. (2009). *Fairness of automated essay scoring of GMAT® AWA* (GMAC ® research reports: RR-09-01). Retrieved from the Graduate Management Admission Council® website: http://www.gmac.com/NR/rdonlyres/FACE0811-B6F7-45A9-B57D-

ED3703984B9A/0/RR0901_AWAFairness.pdf

Han, K. (2004). 高等教育自学考试的等值研究 [Research on equating of Self-taught Higher Education Examination]. *China Examinations*, (12), 45-48.

Huang, J. (2011). Generalizability theory as evidence of concerns about fairness in large-scale ESL writing assessments. *TESOL Journal, 2,* 423-443. doi: 10.5054/tj.2011.269751

Husén, T. (1975). *Social influences on educational attainment: Research perspectives on educational equality.* Paris: Organization for Economic Co-operation and Development.

Japanese Language Testing Association (JLTA). (2006). *The JLTA code of good testing practice*. Retrieved from Japanese Language Testing Association website: http://www.avis.ne.jp/~youichi/COP.html

Jiang, C. (2007). 应当重视教育考试公正研究 [Educational testing justice should be attached importance to]. *China Examinations, 10,* 4-11.

Jin, R. (2011). The justice and equity of examination. *Examinations Research, 1,* 30-35.

Joint Advisory Committee. (1993). *Principles for fair student assessment practices for education in Canada*. Retrieved from Alberta Education: http://www2.education.ualberta.ca/educ/psych/crame/files/eng_prin.pdf

Joint Committee on Testing Practices (JCTP). (2004). *Code of fair testing practices in education*. Retrieved from American Psychological Association website: http://www.apa.org/science/programs/testing/fair-testing.pdf

Kunnan, A. J. (2000). Fairness and justice for all. In A. J. Kunnan (Ed.), *Fairness and*

validation in language assessment (pp. 1–14). Cambridge, UK: Cambridge University Press.

Kunnan, A. J. (2004). Test fairness. In M. Milanovic & C.Weir (Eds.), *European language testing in a global context: Proceedings of the ALTE Barcelona Conference* (pp. 27–48). Cambridge, UK: Cambridge University Press.

Kunnan, A. J. (2008). Large scale language assessment. In E. Shohamy & N. H. Hornberger (Series Eds.), *Encyclopedia of language and education (2nd Ed.). Vol. 7. Language testing and assessment* (pp. 135-155). New York: Springer.

Li, L. (2010). Research on distributing the entrance to higher education. *Peking University Education Review, 8*(2), 56-70.

Liu, H. (2007). The study of imperial examination and the reform of college entrance examination. *Journal of Xiamen University (Arts & Social Sciences)*, (5), 64-71.

Lu, Y. (2011). Fairness in writing assessment: A survey of factors that affect rater bias. *Foreign Language Testing and Teaching*, (2), 30-36.

Mei, Y. & Nie, J. (2009). A review of ethical issues in language testing. *Foreign Language World, 133*, 91–96.

Shi, L., Wang, W., & Xu, J. (2005). Publication culture of foreign language education journals in China. *TESOL Quarterly, 39*, 765-776.

Stricker, L. J., Rock, D. A., & Lee, Y. W. (2005). *Factor structure of the LanguEdgeTM Test across language groups* (TOEFL monograph series MS-32). Retrieved from http://www.ets.org/Media/Research/pdf/RR-05-12.pdf

Wang, F. & Zhang, B. (2007). Research on the equity in the large-scale national selective education tests from a perspective of political science. *Enrollment and Examination in Hubei,* (24), 29-32.

Wang, J. (2011). Study on Present Situation of Examination Fairness and Countermeasure. *China Examinations,* (5), 53-57.

Xi, X. (2010). How do we go about investigating test fairness? *Language Testing, 27,* 147-170. doi:10.1177/0265532209349465

Xie, X. (2004). 考试如何才能“公平”？ [How can tests be “fair”?]. *The Chinese Language Nursery*, (2), 10-11.

Xie, X. & Wang, Y. (2002). 关于考试公平性的一些思考 [Some thoughts on test fairness]. *Examinations Research*, (2), 1-7.

Yang, H. & Gui, S. (2007). The sociology of language testing. *Modern Foreign Languages*, 30, 368-374.

Yang, Y. (2001). 学业成绩评定的激励作用分析 [Analysis of incentives of attainment assessment]. *Journal of Teaching and Management*, (2), 15-16.

Zeng, X. & Meng, Q. (1999). 项目功能差异及其检测方法 [Differential item functioning and its detection methods]. *Journal of Developments in Psychology, 7*(2), 41-47.

Zhou, H. & Shen, G. (2006). Review and Reflection on the History of Enrolment by Examination in China. *Educational Research*, (4), 43-48.

Zhou, J., Ding, X., Zhang, Q., & Wen, H. (2010). Empirical analysis on the fairness in national uniformed entrance examination in general colleges and universities in

Beijing. *Educational Research*, (10), 46-52.

Zieky, M. (2006). Fairness review in assessment. In S. Downing & T. Haladyna (Eds.), *Handbook of Test Development* (pp. 359–376). Mahwah, NJ: Lawrence Erlbaum.

Zou, S. (2011). On enhancing test fairness. *Foreign Language Testing and Teaching*, (1), 42-50.