Language Testing in Asia
a SpringerOpen Journal

RESEARCH Open Access

A survey of English language testing practice in China: the case of six examination boards

Jinsong Fan[1*] and Yan Jin[2]

* Correspondence:
jinsong_fan@fudan.edu.cn
[1]College of Foreign Languages and Literatures, Fudan University, 220 Handan Road, Shanghai 200433, China
Full list of author information is available at the end of the article

Abstract

English language testing has been developing with great momentum in China in the past two decades. However, little research is existent as to how these English tests are developed, administered, and used. This study reported a survey of English language testing practice in the Chinese context through empirically examining the testing practice of six English as a Foreign Language (EFL) examination boards operating at national, municipal, and university levels. The data in this study were collected through a structured questionnaire, developed on the basis of the framework of good testing practice in *Standards for Educational and Psychological Testing*, and a follow-up semi-structured interview. Though informants' responses indicated dominant uniformity at the general level, the survey identified much variation in the testing practice of the six examination boards at the more specific levels, in particular in the areas of pretesting, marking, test equating, test use, etc. Meanwhile, the survey also identified the problems besetting and challenges facing these examination boards in their testing operations. In conclusion, the survey reiterated the importance and urgency of developing a set of professional standards for EFL testing in China which should be targeted at both test developers and stakeholders.

Keywords: Standards, Good testing practice, EFL testing in China, Quality control

Background

The pursuit for quality and professionalism has become an apparent trend in language testing and assessment, as is evidenced by a host of standards or codes of practice [a] which have been developed, implemented or enforced by testing or research organisations from all over the world (see AERA, APA, and NCME, 1999[b] ; ALTE, 1994; EALTA, 2006; ETS, 2002; ILTA, 2000, 2007). In China, the past thirty years or so have seen the robust development of English as a Foreign Language (hereafter EFL) testing in terms of both theory and practice so as to meet the pressing need of the rapidly expanding EFL learner population. On the one hand, many researches have been conducted in language testing and assessment, touching upon a wide spectrum of important issues such as test validity, washback, and fairness (see Cheng, 2008 for a review of language testing research in the Chinese context); on the other hand, a number of EFL tests have been developed and administered by the educational and examinations authorities (Cheng and Curtis, 2010). A most noticeable feature with many of these EFL tests is their extremely huge scale with millions of EFL learners taking these tests every year. In addition, many of these EFL tests are high-stakes since the results on these tests are often used to make important decisions which may have

significant impact on stakeholders, such as admissions into the university, entry into profession, and job promotion opportunities. Therefore, it is essential that the development, administration and use of these EFL tests follow a set of well-developed professional standards so that test developers can ensure the quality of the EFL tests that they develop and deliver, which, by extension, can help to maintain the validity and fairness in making score-based decisions (Bachman and Palmer, 2010). However, currently in China, there are no standards of such a nature, and stakeholders are often in the dark about the quality of the EFL tests. Similarly, educational and examinations authorities are left at a loss about how to effectively evaluate and monitor the practices of the examination boards. The situation becomes especially worrying if we take into account the high-stakes nature of many of the EFL tests in the Chinese context. Therefore, it is high time that a set of professional standards in language testing was developed and strictly adhered to in language testing operations (Yang and Gui, 2007). An important step towards the development and implementation of such standards is to survey what the current EFL testing practice is like since, as argued by Alderson (2011), preaching good practice without examining current practice is irresponsible.

Though many studies have investigated stakeholders' views and perceptions of language testing practices (see e.g., Cheng, 2005; Gu, 2007; Murray, Riazi, and Cross, 2012; Qi, 2005; Rasti, 2009), very few studies focus on examining the testing practices of the test developers (see Alderson and Buck, 1993 and Alderson, 2010 for two exceptions). To the best of our knowledge, no systematic studies have ever been conducted to demystify the testing practices of the examination boards in the Chinese context. Therefore, this study is intended to fill in this gap by addressing the following two research questions:

RQ1. What is the testing practice of the EFL examination boards in developing, administering, and using the EFL tests in the Chinese context?
RQ2. What are the problems besetting and challenges facing the EFL examination boards in developing, administering, and using the EFL tests in the Chinese context?

Review of literature

Good practice in language testing

Good testing practice has been discussed very extensively in language testing literature, and has been approached from different perspectives by language testing researchers. A common approach to addressing this issue, for example, is to discuss how a language test should be developed, administered, and evaluated (see e.g., Alderson, Clapham, and Wall 1995; Fulcher, 2010; Heaton, 2000; Li, 1997). These discussions are primarily focusing on good practice in each and every step in the testing cycle, including, for instance, test specifications, item writing, test administration, marking, reporting test results, and *post hoc* test data analyses. A good case in point embracing this perspective is the Cambridge Language Assessment Series, though the publications in this series are concentrated on discussing good practice in the testing or assessment of a specific language ability or skill, including reading (Alderson, 2000), listening (Buck, 2001), speaking (Luoma, 2004), writing (Weigle, 2002), vocabulary (Read, 2000), grammar (Purpura, 2004), and English for Specific Purposes (Douglas, 2000). Also included in this series is how to apply statistical analyses in language testing and assessment (Bachman, 2004). These publications are significant in the sense that they help us to

better understand the construct under measurement and the importance of applying relevant language testing expertise and rigorous quality control measures in the process of testing and assessing these abilities and skills. Another common approach to discussing good testing practice is to focus on one particular dimension of language testing and assessment, to develop theoretical models about this particular dimension, and then to apply these theoretical models to language testing practice. For example, Bachman and Palmer (1996) developed a model of 'test usefulness', which, as they argue, was 'the most important consideration in designing and developing a language test' (p. 17). This model was later expanded into the framework of assessment use argument (AUA) (see Bachman and Palmer, 2010), which effectively linked test taker's performance to the decisions made of the test results and the consequences of using an assessment in a particular context. Other examples adopting this approach are Cheng, Watanabe, and Curtis (2004), focusing on test washback, Kunnan (2000, 2004) on test fairness, Shohamy (2001a, b) on use-oriented testing and the power of tests, and McNamara and Roever (2006) on the social dimensions of language testing.

In addition to these varying perspectives mentioned above, good testing practice has also been extensively documented in the standards or codes of practice which have been developed by testing or research organizations from all over the world (see Fan and Jin, 2010 for a review of the language testing standards). For example, *the ILTA Code of Ethics* (ILTA, 2000), which was adopted at the ILTA Vancouver Conference in 2000, prescribes the ethical duties and responsibilities of language testers working in different contexts from all over the world (see also Boyd and Davies, 2002). Later, the ILTA also promulgated *the Guidelines for Practice* (ILTA, 2007) with a view to providing language testers with more concrete professional guidance when developing, administering, and using language tests. Meanwhile, local testing organizations also became very active in developing and enforcing standards among their members. ALTE and EALTA, the two most prestigious testing organizations in Europe, both developed and enforced their own standards. *The ALTE Code of Practice* (ALTE, 1994), for example, stipulates that good testing practice depends on the responsibilities of both the ALTE members (test developer) and the examination users. The Code therefore prescribes the responsibilities of both test developers and examination users in the four broad areas of test development, interpreting test results, striving for fairness, and informing test takers. Compared with the ALTE Code of Ethics, *the EATLA Guidelines for Good Practice in Language Testing and Assessment* (EALTA, 2006) subsumes a broader range of content, including teacher training in language testing and assessment, classroom testing and assessment, and test development in national or institutional testing units or centers. It is worth noting that in the EALTA Guidelines, linkage to the CEFR and test washback are also included as two important components of good testing practice. In the United States, many standards on good testing practice have been developed and published, among which *ETS Standards for Quality and Fairness* (ETS, 2002) and *Standards for Educational and Psychological Testing* (hereafter the Joint Standards, AERA et al 1999) are probably the two most well-known examples. The ETS Standards contain a wealth of useful information about good testing practice, such as test development, administration, use, and fairness. However, the standards are primarily intended to serve the products and services in the ETS itself. In contrast, the Joint Standards, published in 1999 on the basis of its earlier versions, is intended to

serve a much broader context and wider audience. As a very comprehensive set of standards, it contains 15 chapters with a total of 264 guidelines. The 15 chapters are organized into three major components: test construction, evaluation, and documentation (Part I), fairness in testing (Part II), and testing applications (Part III). Like its predecessors, the 1999 Joint Standards has received wide recognition from the educational and psychological testing community, and has been quoted extensively in the literature of language testing and assessment (see, e.g., Bachman and Palmer, 2010; Fulcher and Davidson, 2007). Furthermore, the Joint Standards has informed the development of many other standards such as *ETS Standards for Quality and Fairness* (ETS, 2002), and *Code of Fair Testing Practice in Education* (JCTP, 2004). Due to its wide recognition and extensive influence, the framework from the 1999 Joint Standards will be used as the theoretical basis for the development of the research instrument used in the present study. This framework is displayed schematically in Figure 1.

Empirical studies of language testing practice

Though many surveys of language testing practice have been conducted, very few of them, as we mentioned earlier, are focusing on the test developers or the examination boards probably because of data inaccessibility or the ostensible lack of scientific rigor in such studies (known as 'investigative journalism', see Alderson, 2011). We only managed to retrieve two empirical studies from the language testing literature examining test developer's practices, conducted by Alderson and Buck (1993) and Alderson (2010).

In one such study, (Alderson and Buck 1993, see also Alderson et al., 1995) examined the testing practices of the EFL/ESL examination boards in the UK context. The purpose of their study was to examine what standards were maintained by the EFL/ESL examination boards in the UK, to what extent all EFL/ESL examining authorities followed the same or similar procedures, and whether procedures could be improved. Their study was divided into two phases. In the first phase, a letter was sent to the EFL/ESL examination boards, asking three broad questions in relation to the standards to which EFL/ESL examination boards followed, and the procedures that they followed to estimate test reliability and to ensure test validity. During the second phase of the study, the two researchers used

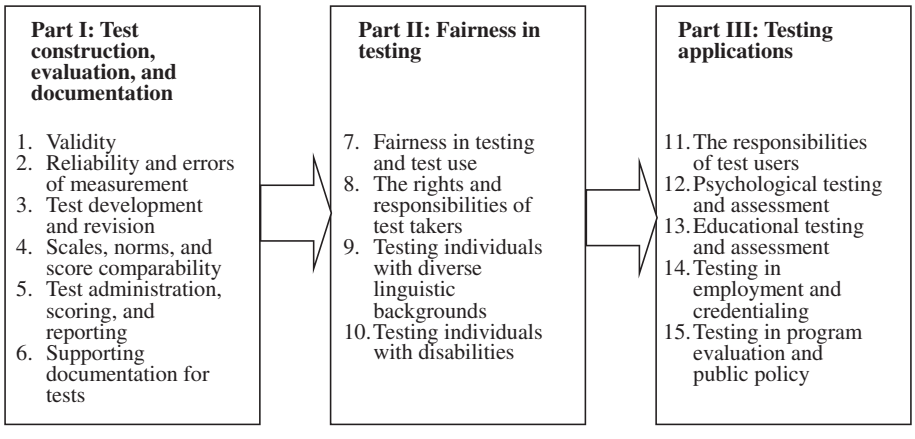

Figure 1 The framework of good testing practice prescribed in the 1999 Joint Standards AERA et al. (1999).

a structured questionnaire to investigate the testing practices of 12 EFL/ESL examination boards, touching upon such areas as test syllabus, test construction, validation, administration, marking, reports and *post hoc* analyses, and test revision. Their study concluded that there was no common agreed set of standards that these EFL/ESL examination boards maintained, and different examination boards did different things, with different degrees of rigor, to monitor the quality of their examinations. The two researchers therefore further concluded that time was ripe for UK examination boards and UK language testers to develop a set of standards all EFL/ESL tests should follow.

In the other study, Alderson (2010) surveyed the testing practices of the examination boards or agents responsible for developing and delivering aviation English language tests. Similar to the 1993 study, this study was also divided into two stages. During the first stage, a 'filter' questionnaire was employed, asking broad questions such as respondents' understanding of the International Civil Aviation Organization (ICAO) scales and a brief section on quality control procedures. During the second stage of the study, a structured follow-up questionnaire was developed on the basis of the *EALTA Guidelines for Good Practice in Language Testing and Assessment* (EALTA, 2006), aiming to survey in detail the testing practices of the test developers. The areas in testing practice under survey included test purpose and specifications, test design and item writing, rating procedures, test administration, and test review. The respondents in this study were 22 persons or organizations from all over the globe who were responsible for developing or delivering aviation English language tests. The study identified worrying problems for aviation English test development and validation, and proposed that standards such as the EALTA Guidelines be adjusted to suit the aviation context.

These two studies are both significant in the sense that they portrayed the testing practices of the examination boards in a certain context. Furthermore, the 1993 study, as the first systematic study of such a nature that appeared in *Language Testing*, helped to inspire the many following discussions of standards and professionalism in the language testing community (see Boyd and Davies, 2002; Davies, 1997, 2004). In contrast, the 2010 study directed our attention to the quality and professionalism in developing aviation English test, a genre of very high-stakes language test. However, neither of these two studies addressed such issues as test washback, test fairness, and test use in the research design, all becoming very important issues nowadays in language testing. Furthermore, neither of these two studies investigated the reasons or rationales behind the inadequacies of test developers' quality control procedures or their failure to comply with relevant standards, and the problems and challenges as reported by these examination boards. Therefore, in this study, we will examine the areas covered by Alderson and Buck (1993) and Alderson (2010). But we will also venture to explore issues such as test washback, impact, and test use. In addition, we will probe the reasons behind the possible inadequacies of quality control procedures, and the challenges facing these test developers in the Chinese context.

Method

Participants

Like in many other testing contexts, data accessibility was inevitably a thorny issue for such a study. Instead of sending letters to the EFL examination boards to get their approval to participate in this study, we decided to focus on some EFL examination boards with which we might have data accessibility. Eventually we selected six EFL examination

boards based in Shanghai and Beijing. The six boards were all well-established, with four of them operating at national level, one at municipal level, and another one at university level. For confidentiality reasons, the six boards were named as Board A to F consistently throughout this study. All the informants who completed our questionnaires and accepted our interviews held a senior position within their respective boards, and were familiar with the routine operations in their boards.

Due to the small sample of the participating examination boards, we didn't intend to generalize the findings we identified in this study, and the report of our study was primarily descriptive. Small as the sample was, we however believed that the findings of this study could, to a certain extent, reflect many of the common issues and concerns in language testing in the Chinese context.

Instruments

Two instruments were developed for this study: a structured questionnaire and a semi-structured interview guide (see Appendix 2 and Appendix 3). The questionnaire was developed on the basis of the framework of good testing practice in *Standards for Educational and Psychological Testing* (AERA et al. 1999, see also Figure 1). The purpose of the questionnaire was to examine the testing practices of the six examination boards in each and every step of test development, administration, and use. It therefore not only subsumed areas such as test development, evaluation, and documentation (Part 1 in Figure 1), but touched on issues such as test fairness (e.g., testing individuals with disability, see Part 2 in the framework), and test use (e.g., using test results to make decisions, see Part 3 in the framework). The questionnaire was structured in such a way as to include the following sections: test purpose and specifications (Section 1), test design and development (Section 2), test administration (Section 3), marking (Section 4), communicating and using test results (Section 5), test analysis and revision (Section 6), and test evaluation (Section 7). Yes/No questions were used in combination with open-ended questions. The purpose of adopting such a format was to examine the practices of the examination boards at the general level, and on the other hand also give informants ample freedom to describe their practices in more detail.

After the survey questionnaire was designed, it was first sent to five colleagues in the language testing community for comment, including three professors who taught language testing courses in the university with practical language testing experience in the Chinese context, and two Ph.D. candidates studying and researching language testing and assessment. The comment from our colleagues proved to be very useful for the improvement of this research instrument. As a result, some new questions were added to the questionnaire such as the practices relating to the identification and prevention of cheating (Q3.1) and test equating (Q6.5, see Appendix 2). Following the revisions, the questionnaire was then piloted with a well-established EFL examination board based in Shanghai, and this board was subsequently included as one of the six participants in this survey. The questionnaire was considered ready for use after the rephrasing of a few questions to improve their clarity to respondents.

Thinking that the informants might be unwilling to answer the open-ended questions in detail and that some issues might need further clarification, we decided to use a follow-up interview with a view to investigating the interested issues in more precision. An interview guide was therefore developed after we finished analyzing the

questionnaire data. It repeated some of the issues in the questionnaire such as item writing and marking, but it was also intended to probe the reasons behind the perceived inadequacies of quality control procedures as well as the challenges facing the examination boards in test development, administration, and use.

Procedures

Before sending the questionnaires to the examination boards, we first of all contacted the six informants to get their consent. To reduce the possible resistance from the informants, we signed an agreement with all informants which stated that their names and the names of examination boards they represented would not appear in our final report, and all data would be used for this research project only. Then the questionnaires were sent to the six examination boards by courier in October, 2010, which, after being completed, were sent back to the researchers by mail within the following month. The interview was conducted in Shanghai and Beijing in December, 2010. Before the interview, the researcher sent the interview guide to all informants. To further alleviate the sense of imposition, the researcher reassured the informants beforehand that each interview would last for a maximum of 40 minutes, and no audio or video-taping would be adopted during the interview. Instead, only notes would be taken. We believed that through taking these measures, we would meet with less resistance from the informants, and they could also report their practices more candidly.

Results [c]

Section 1: Test purpose and specifications

Informants' responses to the Yes/No questions in this section were presented in Table 1, which suggested the dominant uniformity of the testing practices among the six boards in this section. All examination boards had test specifications (Q1.1) which contained a wealth of useful information about the tests they developed and delivered (Q1.2 – 1.7). Also, except Board E who reported 'not sure', the other five boards all conducted needs analysis in their test development (Q1.8).

Section 2: Test design and development

Two criteria were found to be most commonly applied by all the six examination boards in selecting item-writers (Q2.1): teaching experience and understanding of

Table 1 Test purpose and specifications

Questions		A	B	C	D	E	F
Q1.1 Does your board have test specifications?		Y	Y	Y	Y	Y	Y
Does the test specification clearly state/provide: (Q1.2–Q1.7)	The test purpose?	Y	Y	Y	Y	Y	Y
	The language abilities that are tested in each component?	Y	Y	Y	Y	Y	Y
	The targeted test population?	Y	Y	Y	Y	Y	Y
	The sample test papers?	Y	Y	Y	Y	Y	Y
	The sample answers from test takers?	Y	Y	Y	Y	Y	N
	The criteria for evaluation?	Y	Y	Y	Y	Y	Y
Q1.8 Does your board conduct needs analysis?		Y	Y	Y	Y	N/S	Y

Notes: Y=Yes; N=No; N/S=Not Sure.

language testing. However, 'understanding of language testing' was more interpreted by informants as 'having the experience of working as item-writers or coaching students to prepare for important examinations' (in the words of Informant C) than having received a degree or training in language testing. Another common criterion reported by informants was a good command of English, but it seemed that no proof of item-writers' language ability was required. All informants reported that item-writers worked for their boards on part-time basis, and had no fixed periods of service (Q2.2). Despite these similar criteria reported by informants, systematic procedures in selecting and appointing item writers seemed lacking.

Table 2 shows that two boards didn't provide training to item writers (Q2.3) because it was practically very difficult. As Informant C remarked, 'it was virtually infeasible to call a training session since all of our item writers have very tight schedules within their own affiliations.' Two boards reported they didn't provide feedback to item writers (Q2.6), but when asked the reasons why feedback was not provided, their explanations were vague. Though all boards reported that they pretested all test items (Q2.7a), pretesting practice seemed to vary from board to board. For example, Informant C said all their items were pretested, while Board B, E, and F said only objective items were pretested in their boards. The variation became even more noticeable when it came to the sample size used for pretesting (Q2.7b and Q2.7c), ranging from over 1,000 (Board A) to around 20 for every pretest (Board C). Informant F said they randomly selected 7 – 8 students for each pretest. Well aware that the sample size was too small, Informant C and F said they couldn't expand the sample size due to test security concerns.

Soliciting test takers' views about the test at the test development stage had become the prevalent practice for these examination boards while soliciting teachers' views seemed a bit less common (Q2.8). Except Board F, the other five boards all performed statistical analyses of the pretesting data, which, in most cases, included applying the Classical Test Theory (hereafter CTT) to analyzing the facility value and discriminatory index of the items being pretested (Q2.9). Informant F said they went about examining the quality of the pretested items through interacting with the students involved in the pretest. All informants reported that unsatisfactory items were either revised or discarded (Q2.10).

Table 2 Test design and development

Questions		A	B	C	D	E	F
Q2.3 Does your board provide training to item writers?		Y	Y	N	Y	N	Y
Q2.4 Does your board have item writer specifications?		Y	Y	Y	Y	Y	Y
Q2.5 Are there any systematic procedures to ensure the test items meet the requirements set in the test specs?		Y	Y	Y	Y	Y	Y
Q2.6 Do item writers receive feedback about their work?		Y	N	Y	N	Y	Y
Q2.7a Are all the items pretested?		Y	Y	Y	Y	Y	Y
Q2.8 When pretesting, does your board collect data as to:	a. Test takers' views about the level of difficulty?	Y	N	Y	Y	Y	Y
	b. Test takers' views about the appropriateness of test tasks?	Y	N	Y	Y	Y	Y
	c. Teachers' views about the level of difficulty?	Y	Y	Y	N	N	Y
	d. Teachers' views about the appropriateness of test tasks?	Y	Y	Y	N	N	Y

Notes: See Table 1.

Section 3: Test administration

When asked about the measures taken to fight cheating (Q3.1), all informants reported similar measures: test invigilators carefully examining test takers' ID cards and their test registration information. Other measures included using different test papers [d] in different test locations. All informants remarked that they had made strenuous efforts to fight the various kinds of cheating, but it was agreed that they still had a long way to go. Table 3 shows that all boards provided training to test invigilators (Q3.2a) and monitored the process of test administration (Q3.3a), though the length of the training was found to vary greatly from one hour to one day (Q3.3b). Among the six boards, we only received a training pack from Informant A which clearly illustrated the nature of the invigilator training. All six boards had inspectors patrolling the test locations when the test was going on, and these inspectors were required to submit a written report about their monitoring (Q3.3b). Two boards (B and D) reported that they had never provided test accommodation (Q3.4) because, according to the two informants, no individuals had ever applied for such service. The two informants said they were unclear about whether information about the provision of test accommodation was provided to test takers.

Section 4: Marking

Similar criteria were adopted in selecting and appointing markers (Q4.1), including teaching experience and a good command of English. However, proof of English ability seemed not required. Here was a quote from Informant E: 'All markers must be teachers working with universities with over three years of teaching experience or teachers working with secondary schools with the senior professional title.' Informant C and F, however, remarked that the standards sometimes had to be compromised, and English-major postgraduate students were on some occasions hired as markers due to the huge number of test takers. Informant C explained that it was extremely difficult to find adequate number of qualified markers due to the limitations of financial resources.

Table 4 indicates that the variations in the marking practice mainly lay in the areas of double-marking constructed-response items (Q4.4a), and checking for intra- and inter-rater reliabilities (Q4.5 and Q4.6). Only Board C required that every constructed-response item be double-marked while other boards either double-marked part of the items or didn't practice double-marking at all (Q4.4b). Informant D, for example, reported that around 15% of the test papers were double-marked. The reason for failing to double-mark all test papers was attributed to the heavy cost of both time and money. Informant A and F reported that though they didn't double-mark every constructed-response item, rigorous monitoring of the marking process was applied, which became much facilitated after the introduction of the online marking system. When asked why inter- and intra-rater reliabilities were not calculated (Q4.5 and Q4.6), Informant E and F remarked that it was not very necessary since they monitored the marking process

Table 3 Test administration

Questions	A	B	C	D	E	F
Q3.2a Does your broad provide training to test invigilators?	Y	Y	Y	Y	Y	Y
Q3.3a Does your board monitor the process of test admin?	Y	Y	Y	Y	Y	Y
Q3.4 Has your board ever provided test accommodation to test takers with disability?	Y	N	Y	N	Y	Y

Notes: See Table 1.

Table 4 Marking

Questions	A	B	C	D	E	F
Q4.2a Does your board provide training to SET examiners?	Y	Y	Y	Y	Y	Y
Q4.2b Are benchmark scripts used in examiner training?	Y	Y	Y	Y	Y	Y
Q4.3a Does your board convene training for markers of constructed-response components?	Y	N	Y	Y	Y	Y
Q4.4a Are all constructed-response items double-marked?	N	N	Y	N	N	N
Q4.5 Does your board calculate inter-rater reliability?	Y	N	Y	Y	N	N
Q4.6 Does your board calculate intra-rater reliability?	Y	N	Y	Y	N	N
Q4.7 Is the marking process monitored?	Y	Y	Y	Y	Y	Y

Notes: SET = Spoken English Test; Others see Table 1.

(Q4.7). The monitoring, according to the two informants, involved team leaders randomly examining the marking quality of their team members. Though training was provided to markers of constructed-response items (Q4.3a) and benchmark scripts were all used in such training sessions, the length of the training session was found to vary from two hours to one day (Q4.3b).

Section 5: Communicating and using test results

Table 5 indicates that all boards reported composite scores (Q5.1), but three boards also reported component or profile scores (Q5.2). Three boards also provided descriptive information to test takers to help them better interpret their test scores (Q5.3a) though such information, according to informants, was not provided to test takers in the test reports but on their official websites (Q5.3b). In case test takers had doubts over their test scores (Q5.4), Informants A, B, C and E all said that test takers could apply to check their scores, but only Informants A and E said test takers could follow the systematic procedures published on their official websites to apply for such service. Among the six informants, two said that their tests, to the best of their knowledge, were always used for their intended purposes (Q5.5a). Informant A and D, however, both lamented that various uses other than intended had been made of their tests (Q5.5b), such as using their tests to make employment decisions, or even using the test results to determine whether an applicant could be granted the permanent residency, or *hukou*, in a major city. A lot of these uses, according to these two informants, were misuses of their tests, and were not supported with adequate validity evidence.

Section 6: Test analysis and revision

Though all boards performed statistical analyses of their test data after each test administration (Q6.1a, see Table 6), these analyses were mainly used for internal purposes, and were not published or made available to the relevant stakeholders (Q6.2b). The statistical

Table 5 Communicating and using test results

Questions	A	B	C	D	E	F
Q5.1 Does your board report composite test scores?	Y	Y	Y	Y	Y	Y
Q5.2 Does your board report scores on each component?	Y	Y	N	N	Y	N
Q5.3a Does your board provide descriptive information about test scores?	Y	N	N	N	Y	Y
Q5.5a Are your tests used for unintended purposes?	Y	N/S	N	Y	N	Y

Note: See Table 1.

Table 6 Test analysis and revision

Questions	A	B	C	D	E	F
Q6.1a Does your board perform statistical analyses of the test data after each administration?	Y	Y	Y	Y	Y	Y
Q6.2a Are data analyses presented in reports?	Y	Y	Y	Y	Y	Y
Q6.2b Are the reports open to the general public?	N	N	Y	N	N	N
Q6.3a Does your board collect feedback about your tests?	Y	Y	Y	N	Y	Y
Q6.4a Does your board have systematic procedures to revise your tests based on the feedback information?	Y	Y	Y	N	Y	Y
Q6.5a Are different items used in different administration?	Y	Y	Y	Y	Y	Y
Q6.5b Is routine test equivalence research conducted?	Y	Y	N	N	Y	N

Note: See Table 1.

analyses performed on test data (Q6.1b) included, in most cases, using the CTT to analyze facility value and discriminatory index at the item level; other analyses included item-total correlations, reliability coefficients, and in some cases, factor analysis. Informant A and E reported that they had statisticians working within their boards who were responsible for processing and analyzing the test data after each administration.

Most boards collected feedback information about their tests (Q6.3a) primarily through questionnaire surveys (Q6.3b). However, it seemed that some boards did it more flexibly than others. As Informant A reported, 'we collect feedback sometimes through administering questionnaires to test takers, sometimes from the internet, and sometimes from our committee members or markers.' Though all boards reported systematic procedures to revise their tests (Q6.4a), all informants agreed that they refrained from introducing radical revisions or major reforms of their tests for fear of causing a stir among relevant stakeholders (Q6.4b). All boards used different test items in different test administrations (Q6.5a), and the most important reason for doing so was attributed to test security concerns. Also, some informants expressed concerns over the protection of test copyrights since they had heard that their test content was stolen out of the examination rooms (nowadays often with the aid of hi-tech gadgets!), and then used for commercial purposes. Here was a comment from Informant E:

> It is imperative that the government formulate the examination laws so as to protect the copyrights of the test papers, and to bring those who break the laws to justice. Also, it is important to raise the awareness among the general public since many of them do not understand why the test content cannot be disclosed after each test administration.

Though different versions of the test papers were used in different administrations, only three informants said they conducted routine test equating work (Q6.5b), one through test taker anchoring and the other two through test item anchoring (Q6.5c). Other boards either said they didn't do it or they had little research in this regard. The reason was described as the lack of statistical expertise in doing such research.

Section 7: Test evaluation

Two boards had conducted detailed validation studies of their tests (Q7.1a, see Table 7), and the validation research reports had already been published (Q7.1b). Informant F said they had conducted validation studies, but the results were used for internal purposes only. Three boards reported that they had taken measures to ensure that different

Table 7 Test evaluation

Questions	A	B	C	D	E	F
Q7.1a Are there any validation studies of your tests?	Y	N	N	Y	N	Y
Q7.1b Are the validation reports open to the public?	Y	N/A	N/A	Y	N/A	N
Q7.2a Has your board adopted some measures to ensure the equal treatment of different groups of test takers?	Y	Y	N	N	N	Y
Q7.3a Are there any washback studies of your tests?	Y	Y	Y	N	N	Y
Q7.3b Are the washback reports open to the public?	Y	Y	N	N/A	N	Y
Q7.4a Are other any other quality control measures?	N	N	N	N	N	N

Notes: N/A=Not Applicable; Others see Table 1.

groups of test takers were treated equally and fairly in the test process, all at the stage of item writing (Q7.2a), but it seemed that none of these boards adopted systematic fairness review guidelines in the test construction process (Q7.2b). Three boards reported that had already conducted detailed washback studies (Q7.3a), and the research reports had already been published (Q7.3b). Other informants said they were planning to conduct washback studies in the near future.

Discussion and conclusions

The survey of the six examination boards indicates that instead of following an external set of standards to guide their routine operations (e.g., *the ILTA Guidelines for Practice*, ILTA, 2007), all of them followed the quality control procedures developed within their own boards. However, when it comes to the quality control procedures that they follow, three questions merit our attention: how are these quality control procedures developed, are these quality control procedures *per se* valid, and are these quality control procedures strictly adhered to in routine testing operations? In other words, we need to move back to the classical question which is often raised when discussing ethics and good practice in language testing: who guards the guardians themselves (see also Boyd and Davies, 2002)? Though we didn't examine all the documents stipulating these quality control procedures in each examination board, we believe the practices that they reported, to a large extent, reflect the nature and validity of their quality control procedures.

At the more general level, the testing practices of the six examination boards seemed to follow a quite uniform pattern, and appeared to comply with the good testing practice as prescribed in much of the testing literature we reviewed in this paper (e.g., AERA et al. 1999; Alderson et al., 1995; Bachman and Palmer, 1996, 2010; ETS, 2002; Fulcher, 2010) (see informants' responses to the Yes/No questions in Tables 1, 2, 3, 4, 5, 6, 7). In addition, the questionnaire data seemed to demonstrate even better uniformity in Section 1 (test purpose and specifications) and Section 3 (test administration) than the other five sections, suggesting that the practices in these two sections were in better agreement with the good testing practice in the Joint Standards (AERA et al. 1999). However, a closer scrutiny of their operations at the more specific levels revealed quite striking differences in a number of areas, such as the size and representativeness of the pretesting sample, double-marking of constructed-response items, the monitoring of the marking process, and test equating. The findings in this study are largely consistent with what Alderson and Buck (1993) found about the EFL/ESL testing practices in the UK context, and what (Alderson 2010) found about aviation English testing practices, suggesting that different examination boards did quite different things with different

degrees of professional rigour. Though the boards under survey operate at different levels, apparently there are some common standards which they are supposed to follow in their practices. These findings raise important concerns over the validity of the quality control procedures which the examination boards follow, and the measures taken to ensure their full implementation in the testing process. We therefore believe it is essential for the examination boards to carefully review their quality control procedures and the enforcement mechanisms, identify the problematic areas, and work out practicable plans to continuously improve their quality control system in the testing process.

In addition to the above findings, the most important reason behind the inadequacies in the quality control procedures was identified as the various practicality constraints which had prevented examination boards from adequately enforcing their quality control procedures. A common conundrum reported by all informants, for example, was the formidable number of test takers taking their tests every year, an issue which has been repeatedly raised by many testing researchers when discussing language testing in the Chinese context (see e.g., Cheng, 2008; Cheng and Curtis, 2010; Jin, 2010; Li, 1997). Another major challenge reported by the informants was the shortage of financial resources which could be mobilized to employ enough qualified professionals (e. g., item writers, makers, statisticians) working for the examination boards. The reason, according to the informants, was believed to be attributable to the widespread misconception harboured by many people, even including some EFL professionals, that developing an English language test was a simple and easy job which required little or no professional expertise. Therefore, this study suggests that stakeholders shall have better understanding of the nature of language testing, and examination boards shall have access to more financial resources from the relevant educational and examinations authorities. On the other hand, it is important for examination boards to understand that though language testing is all about making compromises (see Alderson, 2011), which compromises they can make needs their careful weighing and planning (Bachman and Palmer, 1996). Test developers need to strike the informed balance between reliability, validity, and practicality (Li, 1997). Furthermore, stakeholders' lack of understanding of language testing was also described by informants as accounting for many of the other problems besetting the examination boards, such as test-oriented teaching and learning, reckless infringement of test copyrights, and the misuses of the test results, to name but a few. Informants believed that if stakeholders were armed with a better understanding of language testing, these issues could probably be much better tackled.

Echoing the call of Yang and Gui (2007) to develop a set of professional standards for language testing in China, this study reiterates the necessity and urgency of such an endeavour. However, the findings of this study suggest that the standards should not be targeted at test developers (the examination boards) alone, as Yang and Gui (ibid.) suggested, but be extended to other stakeholders, such as students, teachers, employers, publishers, etc., and should serve two purposes: 1) raise the awareness of quality and professionalism among test developers, and further improve the quality, fairness, and transparency of their testing practices, and 2) disseminate the basics of language testing and good testing practice to relevant stakeholder groups (see also ALTE, 1994; ILTA, 2007; JCTP, 2004) . We believe that by so doing, the development and implementation of the standards is more likely to bring about the intended and more desirable outcomes.

One limitation with this study is the small sample of EFL examination boards under investigation. To portray a more representative picture of EFL testing in China, it is

apparently necessary to investigate the testing practices of more examination boards operating at different levels. In addition, all the data in this study were self-reported by the informants representing the examination boards. Though various measures were taken to improve the validity of the data, social desirability effect was almost inherent in such research methods (see Brown, 2001; Gorden, 1998). Therefore, to further improve the validity of this study, it is necessary to collect data from more sources, employing different research designs and instruments so as to better triangulate the findings of this study (see Seliger and Shohamy, 1989). Surveys, for example, can be conducted in the future to investigate the views and perceptions of a wide range of stakeholder groups (e. g., students, teachers, parents, publishers) of the testing practices of the EFL examination boards.

Endnotes

[a] We use 'standard' and 'code of practice' interchangeably in this paper, both referring to 'an agreed set of guidelines which should be consulted and, as far as possible, heeded in the construction or the evaluation of the test' (Alderson, Clapham, and Wall 1995, p. 236).

[b] See Appendix 1 for the full spelling of all the acronyms in this paper.

[c] Since the interview was primarily intended to probe in more precision the issues investigated in the questionnaire, the interview data were reported together with the questionnaire data. In addition, since audio or video-taping was not adopted, we refrained from using direct quotes from the informants for fear of inaccuracy. Direct quotes were used only when we were sure about their accuracy.

[d] Here 'different test papers' refers to the different arrangement of the same test content in a test paper.

[e] Only open-ended questions are presented here. Readers are referred to Tables 1, 2, 3, 4, 5, 6, 7 for the Yes/No questions in the questionnaire.

APPENDIX 1: List of abbreviations

AERA: American Educational Research Association
ALTE: Association of Language Testers in Europe
APA: American Psychological Association
EALTA: European Association of Language Testing and Assessment
EFL: English as a Foreign Language
ETS: Educational Testing Service
ILTA: International Language Testing Association
JCTP: Joint Committee on Testing Practices
NCME: National Council on Measurement in Education

APPENDIX 2[e]: Questionnaire

Q2.1: What are the criteria for item writer selection and appointment?
Q2.2: How long are item writers employed?
Q2.7b: How is the pretesting sample decided?
Q2.7c: What is the usual size of pretesting sample?
Q2.9: What statistical analyses are performed of the pretesting data?
Q2.10: What actions are taken if the items are found not to have the required technical quality?

Q3.1: What measures are taken to prevent cheating?
Q3.2b: Who is responsible for test invigilator training, and how long does the training normally last?
Q3.3b: What happens after the monitoring of test administration?
Q4.1: What are the criteria for the selection and appointment of markers?
Q4.3b: How long do the standardization meetings normally last?
Q4.4b: What is the proportion of the test papers that are double-marked?
Q5.3b: Where is the descriptive information available?
Q5.4: What procedures should students follow in case they have doubts about their scores?
Q5.5b: What are the uses other than intended of the tests developed by your board, and how do you look at these uses?
Q6.1b: What statistical analyses are performed of the test data after each administration?
Q6.3b: How and from whom do you collect feedback information?
Q6.4b: How many times have your tests been revised?
Q6.5c: How is test equivalence verified?
Q7.2b: What are the measures taken to ensure different groups of test takers are treated equally and fairly?

APPENDIX 3: Interview guide

1. How do you select or recruit item writers? Please describe the measures taken to ensure the quality of item writing.
2. Please describe how you pre-test the items.
3. Please describe the procedures followed in your board to ensure the standardization of test administration.
4. Please describe how marking is conducted at your board and the measures taken to ensure the quality of marking.
5. How do you view the uses other than intended of the tests that your board develops and delivers?
6. What are the challenges, in your opinion, that face your board in test development, administration, and use?

Competing interests
The authors declare that they have no competing interests

Authors' contributions
JF did the literature review, participated in the design and pilot study of the research instrument, collected the interview data, and drafted the manuscript. YJ worked together with JF on the design and revision of the research instrument, contacted the examination boards, and collected the questionnaire data. All authors read and approved the final manuscript.

Acknowledgements
We would like to thank Shanghai Education Commission for funding this study as a Research Project of Scientific and Technological Innovation (Project No. B3582W), the six EFL examination boards for their participation, and our colleagues in the language testing community for their generous support.
An earlier draft of this paper was presented at the 33rd Language Testing Research Colloquium held at the University of Michigan on June 23-25, 2011.

Author details
[1]College of Foreign Languages and Literatures, Fudan University, 220 Handan Road, Shanghai 200433, China. [2]School of Foreign Languages, Shanghai Jiaotong University, No. 1954, Huashan Road, Shanghai 200030, China.

Received: 17 January 2013 Accepted: 4 March 2013
Published: 8 April 2013

References

AERA, APA, & NCME. (1999). *Standards for educational and psychological testing*. Washington DC: AERA.
Alderson, JC. (2000). *Assessing reading*. Cambridge: Cambridge University Press.
Alderson, JC. (2010). A survey of aviation English tests. *Language Testing, 27*(1), 51–72.
Alderson, JC. (2011). *A lifetime of language testing*. Shanghai: Shanghai Foreign Education Press.
Alderson, JC, & Buck, G. (1993). Standards in testing: a survey of the practice of UK Examination Boards in EFL testing. *Language Testing, 10*(1), 1–26.
Alderson, JC, Clapham, C, & Wall, D. (1995). *Language test construction and evaluation*. Cambridge: Cambridge University Press.
ALTE. (1994). *The ALTE code of practice. Resource document*. Association of Language Testers in Europe. http://www.alte.org/attachments/files/code_practice_eng.pdf. Accessed 20 June 2008.
Bachman, LF. (2004). *Statistical analyses for language assessment*. Cambridge: Cambridge University Press.
Bachman, LF, & Palmer, AS. (1996). *Language testing in practice*. Oxford: Oxford University Press.
Bachman, LF, & Palmer, AS. (2010). *Language assessment in practice: developing language assessments and justifying their use in real world*. Oxford: Oxford University Press.
Boyd, K, & Davies, A. (2002). Doctor's orders for language testers: the origin and purpose of ethical codes. *Language Testing, 19*(3), 296–322.
Brown, JD. (2001). *Using surveys in language studies*. Cambridge: Cambridge University Press.
Buck, G. (2001). *Assessing listening*. Cambridge: Cambridge University Press.
Cheng, L. (2005). *Changing language testing through language testing: a washback study*. Cambridge: Cambridge University Press.
Cheng, L. (2008). The key to success: English language testing in China. *Language Testing, 25*(1), 15–37.
Cheng, L, & Curtis, A. (2010). *English language assessment and the Chinese learner*. New York and London: Routledge, Taylor and Francis Group.
Cheng, L, Watanabe, Y, & Curtis, A. (2004). *Washback in language testing: Research methods and contexts*. London: Lawrence Erlbaum Associates, Publishers.
Davies, A. (1997). (Guest editor) Special issue: ethics in language testing. *Language Testing,* 14.
Davies, A. (2004). (Guest editor) Special issue: ethics in language testing. *Language Assessment Quarterly,* 4.
Douglas, D. (2000). *Assessing English for specific purposes*. Cambridge: Cambridge University Press.
EALTA. (2006). *EALTA: guidelines for good practice in language testing and assessment. Resource document*. European Association of Language Testing and Assessment. http://www.ealta.eu.org/documents/archive/guidelines/English.pdf. Accessed 20 June 2008.
ETS. (2002). *ETS standards for quality and fairness*. Princeton, New Jersey.
Fan, J, & Jin, Y. (2010). Standards in language testing: review, reflection and inspiration. *Foreign Language World, 1*, 82–91.
Fulcher, G. (2010). *Practical language testing*. London: Holder Education.
Fulcher, G, & Davidson, F. (2007). *Language testing and assessment: an advanced resource book*. Oxon: Routledge, Taylor and Francis Group.
Gorden, RL. (1998). *Basic interviewing skills*. Prosper Heights: Waveland Press, Inc.
Gu, X. (2007). *Positive or negative: an empirical study of CET washback*. Chongqing: Chongqing University Press.
Heaton, J. (2000). *Writing English language tests*. Beijing: Foreign Language Teaching and Research Press.
ILTA. (2000). *Code of ethics. Resource document*. International Language Testing Association. http://www.iltaonline.com/images/pdfs/ILTA_Code.pdf. Accessed 20 June 2008.
ILTA. (2007). *The ILTA guidelines for practice. Resource document*. International Language Testing Association. http://www.iltaonline.com/images/pdfs/ILTA_Guidelines.pdf. Accessed 20 June 2008.
JCTP. (2004). *Code of fair testing practices in education*. Washington DC: .
Jin, Y. (2010). The place of language testing and assessment in the professional preparation of foreign language teachers in China. *Language Testing, 27*(4), 555–584.
Kunnan, AJ (Ed.). (2000). *Fairness and validation in language assessment*. Cambridge: Cambridge University Press.
Kunnan, AJ. (2004). Test fairness. In M Milanovic & C Weir (Eds.), *European language testing in a global context: proceedings of the ALTE Barcelona conference*. Cambridge: Cambridge University Press.
Li, X. (1997). *The science and art of language testing*. Changsha: Hunan Education Press.
Luoma, S. (2004). *Assessing speaking*. Cambridge: Cambridge University Press.
McNamara, T, & Roever, C. (2006). *The social dimensions of language testing*. Oxford: Blackwell Publishing.
Murray, JC, Riazi, AM, & Cross, JL. (2012). Test candidates' attitudes and their relationship to demographic and experiential variables: The case of overseas trained teachers in NSW, Australia. *Language Testing, 29*(4), 577–595.
Purpura, J. (2004). *Assessing grammar*. Cambridge: Cambridge University Press.
Qi, L. (2005). Stakeholders' conflicting aims undermine the washback function of a high-stakes test. *Language Testing, 22*(2), 142–173.
Rasti, I. (2009). Iranian candidates' attitudes towards IELTS. *Asian EFL Journal, 11*(3), 110–155.
Read, J. (2000). *Assessing vocabulary*. Cambridge: Cambridge University Press.
Seliger, HW, & Shohamy, E. (1989). *Second language research methods*. Oxford: Oxford University Press.
Shohamy, E. (2001a). *The power of tests: a critical perspective on the uses of language tests*. Essex: Pearson Education Limited.
Shohamy, E. (2001b). Democratic assessment as an alternative. *Language Testing, 18*(4), 373–91.
Weigle, S. (2002). *Assessing writing*. Cambridge: Cambridge University Press.
Yang, H, & Gui, S. (2007). The sociology of language testing. *Modern Foreign Languages, 4*(30), 368–374.

doi:10.1186/2229-0443-3-7
Cite this article as: Fan and Jin: **A survey of English language testing practice in China: the case of six examination boards.** *Language Testing in Asia* 2013 **3**:7.