Using summative assessments for formative purposes:

Design of new score report for the Public English Test System (PETS) in China

Zuhua Zhao
National Education Examinations Authority, China

This paper reports on assessment theory and practice in designing new score report for the Public English Test System (PETS) in China. Originally established as a summative assessment framework, PETS reports only aggregated scores to test takers and provides no performance feedback to support their learning. Recently, the idea that summative assessments of language ability can be adapted to be used formatively to enhance language learning has been advanced and recognized in the field of language assessment and also in the PETS management group. One of the PETS implementations around this idea is to improve its score reporting to include more candidate performance information. This paper reports major considerations in the designing process as well as the final structure and format of the new score report, which provides comprehensive information on test takers' performance (skill achievements and sub-skill achievements) and describes their overall language ability levels. Test takers can view reports online, and those who have taken the tests more than once (on the same level or different levels) can view their degrees of progress from the reports. Several drawbacks in the design will also be discussed in this paper.

Keywords: summative assessment; formative assessment; PETS; score report

1. Background

The Public English Test System (PETS) in China is designed and developed by the National Education Examinations Authority (NEEA) with the help of the University of Cambridge Local Examinations Syndicate (UCLES, now known as Cambridge Assessment), and formally introduced to the public in the late 1990s. The system has been developed in response to the growing need and effort throughout the country to improve communication in English and the increasing interest in assessing the ability more effectively. Its materials and assessment criteria provide an adequate focus on communicative language ability, and it is hoped that the system will encourage more communicative language teaching and learning and lead to an increase in the standard of English across the country.

The framework of tests incorporates 5 distinct but coherent levels of assessed proficiency, ranging roughly from the level of English expected at Junior High School (after 3 years' English study) to the level of English graduates planning to study and/or work abroad. Levels of proficiency are positioned along a continuum by using

Rasch model.

There is a recognizable similarity in test form, content and focus across different levels of proficiency. The test at each of the 5 levels is composed of two separate test components. One is referred to as the Written Test and consists of sections assessing listening comprehension, use of English, reading comprehension and writing containing within a single test booklet; the other component is an Oral Test designed to assess candidates' speaking ability. Candidates need to pass both the written and the oral component in order to receive a certificate.

Candidates can take tests at levels 1, 2, 3, and 5 twice a year, and level 4 once a year. In 2008, the number of candidates at all levels totaled 1.09 million, a 10 percent increase compared with that in 2007. Most of the candidates take the tests for summative purposes, especially at level 2 and 3. In a few provinces, level 2 is being employed as a supplement to or a substitute for the English test of National Matriculation Examinations. And nationwide, level 3 functions as a compulsory test for self-taught learners to get diploma. Candidates from these two levels account for 78 percent of the total population.

2. Considerations on using PETS for formative purposes

Why should a large-scale summative test be used for formative purposes?

Recently a number of researchers, particularly those working in the Asia region (Lee, 2007; Davidson, 2008) have argued that summative assessments can also be used for formative purposes, i.e., to improve learning and teaching. To be formative, summative assessments must be undertaken while students are still learning (and teachers are still teaching); more importantly, constructive feedback is required and should be provided to ensure student involvement, understanding and action. While these researchers are largely involved in small-scale or low-stake assessments, large-scale summative tests on the other hand are also feeling the call to be formative and pedagogically useful. As Kunnan summarizes, one of the main challenges facing large-scale language assessments around the world today is that "testing consumers have called for more descriptive test information that allows for meaningful interpretations and fair use of test results in order to improve instructional design and guide students' learning." (Kunnan, 2008: 149) Traditional large-scale language testing endeavors to quantify an individual test-taker's language ability by providing only a single letter or number grade, notwithstanding the score aggregation process is often done in a complicated manner, as illustrated in the case of UCLES' First Certificate in English (Alderson et al, 1995: 151), and in PETS practice as well. The use of aggregated test scores as an overall measure of language proficiency is more convenient for decision-making than for educational remediation; students cannot understand the meaning of such scores, from which they cannot see their strengths and weaknesses in learning.

Originally established as a summative assessment framework, PETS provides very brief scoring information to test takers. For the oral test, a single score is provided;

for the written test, sub-section scores will be weighted, equated, aggregated and finally distributed to candidates, but scores for each sub-section are not given. Candidates receive their scores in two ways depending on whether they pass or fail the test: those getting a fail are noticed on paper, and those getting a pass can view their scores from the websites of provincial/regional testing councils. This manner of score reporting has been put into practice since PETS started in 1999, a reasonable and practical way for a large-scale assessment, especially for a test system with test-taker population of over 1 million a year.

No doubt that this old method is gradually incompatible with the emerging needs of candidates, it is also deviating from one of its initial purposes bit by bit. PETS was originally built to reform existing English tests in China and develop more criterion-referenced tests placing greater emphasis on communicative language skills in English. Given the role of examinations in the social and educational context of China, it was supposed that this test system over time would improve teaching of English, leading to higher student achievement. Ten years passed, however, the goal of integrating this system within a coherent and cohesive national framework has not been fully accomplished due to educational policy changes. In this situation, PETS is willing to renovate itself to better support English learners around China, by encouraging the use of PETS for formative purposes.

How can a large-scale summative test be used for formative purposes?

For a large-scale summative test to be used for formative purposes, first and foremost, constructive feedback should be given to candidates to help them to identify what needs to be learnt next, indicating that the feedback must contain diagnostic information (Fulcher and Davidson, 2007: 29). This is not usually found in large-scale summative tests, often due to “lack of motivation and researching capabilities of testing agencies” (Kunnan, 2008: 150). While PETS has generated the necessary motivation for providing more information to test takers on a routine basis, the pressing questions are what can be provided and what still cannot in present administrative and researching situation. The questions has been thoroughly discussed and explored in the feasibility study. One thing to be certain, besides the aggregated score, sub-section scores can also be equated and given to candidates to reveal their strengths and weaknesses. Another aspect of information is detailed performances of candidates on sub-skills in all sections except writing and speaking, which can be unfolded through a summary of item testing focuses and achievements of candidates. Above all, insomuch as the proficiency scale of PETS is established on Rasch model, a candidate’s language ability level can be worked out from one test and compared through several tests of the same level or different levels. These three facets of diagnostic information shall combine to draw a full picture of students’ language performances and language ability levels as well as illuminate the direction of future learning for them. Indeed, PETS is devoted to presenting a different outlook on the nature of feedback that summative assessments provide for formative purposes.

In classroom, teachers are deeply involved in traditional formative assessment

and care much about its outcome because the primary purpose of the assessment is to improve teaching and support learning. As regards PETS, however, the assessment has not been integrated into the national curriculum, and teachers may care little about students' performances in PETS simply because those have nothing to do with their accountability. Students usually take tests without teachers' intervention, and those self-taught language learners may have no teachers at all. Thus, the principal users of new score report for PETS should be test takers themselves; in other words, test takers may have to read the reports on their own and learn independently from the reports about their performances in the tests and their overall English ability levels. Nevertheless, it is still recommended that test takers share the diagnostic information on the reports with their teachers or parents so as to obtain a better support, though it is a free choice on their part.

If here we are saying PETS is about to play not only the role of the assessor but also that of the teacher, then this teacher is trying to interact with each test taker in the most neutral way. That is to say, he makes rather persistent, objective and comprehensive diagnosis of learners' language abilities. He leaves concrete advice in a fixed form; he points out students' strengths and weaknesses in a quantitative manner, and he even marks out clearly the distance that a learner is away from an upper level of language proficiency. Such practice can be seen in terms of Vygotsky's (1978) notion of the zone of proximal development, or "that space between what the individual can accomplish independently and what he or she can do with assistance". Furthermore, this formative assessment may also have a dynamic impact on test takers, meaning that they can observe their learning progresses through a series of score reports in case they attend tests of the same level or different levels twice or more. In this way, assessment and learning can form an advancing cycle similarly managed in the context of language classroom, although the interaction here is definitely long-term and depends to a large extent on how often a candidate takes the tests.

Development principles underpinning the project

The essential development principles which underpin the new score report are further identified as follows:

- *that its principal users are test takers themselves*
  The chief purpose of this new score report is to facilitate learning instead of proving ability, so potential decision-makers like personnel officers and admission tutors are not included in the scope of score users. In fact, the real world situation is: decision-makers only want to know the candidate's proficiency in broad terms and may not accept a complex score report (Alderson et al, 1995: 153).
- *that it should emphasize test-takers' levels of language ability*
  Test-takers' performances are presented by referencing to the criterion of the test. Test takers can view their strengths and weaknesses in skills and the gaps between their proficiency levels and the immediate upper/lower level of the test

system. With these, learners can position themselves better in the ability continuum and set appropriate goals for themselves.

- *that it should be clear and coherent to users*
  Descriptions on the report should consist of terminologies of language testing or linguistics as few as possible in order to make common users understand its meaning easily. Tables and graphs are preferred to long descriptions. Different parts of the report should be coherent and combine to present a unitary picture of the test taker's language ability.

3. Content and format of new score report

The sample report for PETS Level 2 in the appendix shows how the information is displayed.

- **A: Candidate personal information.**
  It includes a candidate's name, candidate number, ID number as well as his/her total score of the Written Test.
- **B: User guide.**
  It outlines the main contents of the report. Moreover, it stresses that each test taker can observe his/her strengths and weaknesses in English from the report, and set a new goal for his/her future learning.
- **C: PETS level descriptions.**
  It introduces the overall proficiency scale and likely candidature for each PETS level in relation to educational/occupational background. PETS level 3, for example, aims to assess the candidates who wish to obtain a diploma or certificate in the context of self-taught learning or private university education. It also aims to assess the candidates who wish to work as secretaries, assistant managers, research assistants, laboratory technicians, etc.
- **D: Sub-skill achievement.**
  Firstly, the assessed sub-skills in each of the four sections of the Written Test are named in the tables. The Listening Comprehension section, for example, is characterized by four main sub-skills: (1) listening for gist; (2) listening for specific information; (3) understanding through inference; and (4) understanding attitudes, opinions and intentions. Secondly, the number of items to assess each sub-skill is listed, accompanied by the number of items that the test taker answers correctly. As shown in the sample report, there are four items being used to assess candidates' ability to infer in listening comprehension, and the candidate got half of them right. More attention needs to be taken to the Writing section, which is comprised of two parts in PETS Level 2: proofreading and essay writing. Since no detailed sub-skill information is being set to these two parts in the item writing phase, and the essay part is being rated holistically instead of analytically, only total scores and obtained scores will be provided.
- **E: Skill achievement.**
  The histogram below shows cut-off scores for each of the four sections of the Written Test, and the one above shows the scores obtained by a candidate in each

section. By comparing the corresponding figures or the lengths of bars, the candidate can find out his/her strengths or weaknesses in skills. As in the sample report, the candidate performed fairly well in the Use of English section but poorly in the Writing section. To be noted, the reported scores are not raw scores but scaled scores calculated after weighting and equating procedures.

- **F: Person ability.**
  It shows a vertical axis on which the scale of language competence is marked. Each level is subdivided into five intervals, and three immediate levels are shown to a certain candidate, i.e., the level he/she took and its immediate upper and lower levels. The pass/fail boundaries for each of the three levels are prominently marked on the axis. The candidate's language ability level, which is estimated from his/her reported score through Rasch modeling, is shown along the axis. Judging from the chart in the sample report, the candidate's communicative language ability is slightly higher than level 2's pass/fail boundary, which determines that his/her performance is adequate for level 2 but still requires efforts in order to reach level 3. To be noted, even if the candidate's ability is indicated higher than the level 3 pass/fail boundary, he/she would not receive a level 3 certificate because he/she did not take the test.

4. Discussions

Despite its renovation as a promising formative measurement tool, the new score report for PETS still has drawbacks. The most obvious one is that such a report is less diagnostic than what is commonly thought to be, or it is still quantitative. This is due to the fact that qualitative evaluation like pertinent comments and elaborated suggestions teachers constantly make in the classroom assessment environment cannot be generated automatically in a large-scale assessment situation. Although researchers have taken the initiative in integrating traditional descriptive suggestions into small diagnostic assessment systems (Alderson, 2005), large-scale summative assessments often consider it less impersonal and are unwilling to invite more inconveniencies. In doing this project, we strongly felt caught in a dilemma. On the one hand, suggestions may be too general to be useful to students; on the other hand, trivial as they are, evaluations may run the risk of unreliability. In other words, students may benefit little when they are simply recommended to read more books in English or keep listening to English radio programs; on the other side, if we conclude that a student needs to improve his/her sub-skills of listening for specific information because he/she gets only one out of four such items correct, our judgment may be too absolute because the number of items is not enough for us to make it. Considering this complicated situation, we would rather leave this area blank.

Another problem is that the new score report targets at the written test takers only. In current PETS practice, the written and oral test components are administered separately; test takers can choose to attend either the Written Test or the Oral Test, although both tests are required to be passed to obtain a certain level certificate. In this sense, it would be impractical to develop a unified score report to include both

components for now. Since the Written Test contains much more contents than the Oral Test, from which much more performance information can be analyzed and presented to test takers, it should be reasonable to start from the Written Test score report; nevertheless, without covering oral skills, the picture of communicative language ability is incomplete. Moreover, the feedback on the writing section is quite brief in that only a single score is provided. This is because the rating of essays is done in a holistic rather than analytic manner; in this sense, test-takers' performances cannot be analyzed according to different facets of the rating scale such as accuracy, cohesion, coherence, etc.. For a large-scale language assessment like PETS, this option is made out of practical concerns, but it is beyond doubt a flaw in using the tests for formative purposes. This situation is expected to change when the rating efficiency and reliability is improved through internet-based scoring.

Finally, students' understanding of the new score report and their further actions are critical in learning, but unlike teachers in classroom, PETS is quite passive and cannot check or supervise the learners. Though the information on the report is succinct, students may possibly encounter problems in understanding them. Moreover, self-taught learners may have nobody to turn to, or they just have no strong motivation to seek help from others. So, further notifications should accompany the new score report to emphasize the necessity of integrating the score report into learning. The testing agency should provide consultancy whenever a help is sought from test takers. In writing this paper, a questionnaire is being designed and administered online to gather score report reading and using information in order to provide better support for language learners.

References

Alderson, J. Charles. *Diagnosing Foreign Language Proficiency: The Interface between Learning and Assessment*. London and New York: Continuum, 2005.

Alderson, J. Charles, Caroline Clapham, and Dianne Wall. *Language Test Construction and Evaluation*. Cambridge: Cambridge University Press, 1995.

Davidson, Chris. "Using Summative Assessments for Formative Purposes: The Ultimate Justification for Learners and Teachers." *30th LTRC*. Hangzhou, China, 2008.

Fulcher, Glenn, and Fred Davidson. *Language Testing and Assessment*. London and New York: Routledge, 2007.

Kunnan, Antony John. "Large Scale Language Assessments." *Language Testing and Assessment*. Eds. Shohamy, Elana and Nancy H. Hornberger. 2 ed. Vol. 7. Encyclopedia of Language Education. New York: Springer, 2008.

Lee, Icy. "Feedback in Hong Kong Secondary Writing Classrooms: Assessment for Learning or Assessment of Learning?" *Assessing Writing* 12 3 (2007): 180-98.

Vygotsky, L. S. *Mind in Society*. Cambridge, MA: Harvard University Press, 1978.

Appendix: Sample score report for PETS

全国英语等级考试

Public English Test System

PETS 全国英语等级考试

笔试成绩报告单

第二级

姓名：王 鑫

得分： 63

A

考试时间：2006 年 9 月 9 日

准考证号：xxxxxxxxxxxxxxxxxxx

身份证件号：110110199010101010

说明

B

本报告单提供你本次考试 各部分得分、以及你在 PETS 量表中所处的位置。

通过本报告，你可以了解你在英语学习上有哪些优势，还存在哪些问题，可以在下一步的学习中有所针对。同时你还能知道你的英语水平离 PETS 本级别及较高级别的合格线有多大距离，从而为自己制定更合理的学习目标。

报告单编号：1111111111

全国英语等级考试

Public English Test System (PETS)

全国英语等级考试（Public English Test System，简称 PETS）是教育部考试中心设计并负责的全国性英语水平考试，是面向全体公民的、多级别的权威英语测试体系。PETS 全面考查考生的语言交际能力，分为笔试和口试两个相对独立的考查部分，共含有 5 个级别。

C

PETS-1 是初始级，通过该级考试的考生，其英语基本符合请如出租车司机、宾馆行李员、门卫、交通管，以及同层次其他工作在对外交往中的基本需要。（PETS-1 下设一个附属级 PETS-1(B)，通过该级考试的考生能够掌握最基本的英语日常用语。）

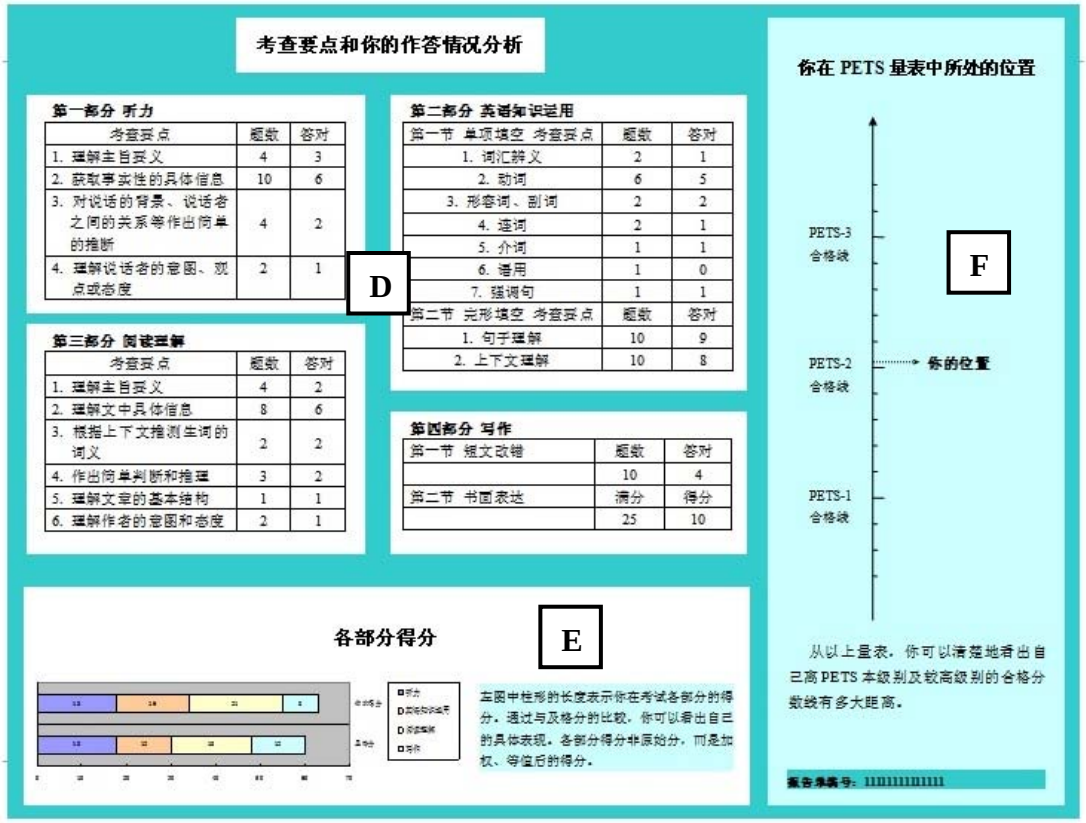
PETS-2 是中下级，通过该级考试的考生，其英语水平基本满足进入高等院校继续学习的要求，同时也基本符合请如宾馆前台服务员、一般银行职员、涉外企业一般员工，以及同层次其他工作在对外交往中的基本需要。

PETS-3 是中间级，通过该级考试的考生，其英语达到高等教育自学考试非英语专业本科毕业水平或符合普通高校非英语专业本科毕业的要求，基本符合请如企事业单位行政秘书、经理助理、一般管理人员或科技工作者、外企职员，以及同层次其他工作在对外交往中的基本需要。

PETS-4 是中上级，通过该级考试的考生，其英语水平基本满足攻读高等院校非英语专业硕士研究生的需要，基本符合一般专业技术人员或研究人员、现代企业经理等工作对英语的基本要求。

PETS-5 是最高级，通过该级考试的考生，其英语水平基本满足在国外攻读非英语专业硕士研究生或从事学术研究工作的需要，也能满足他们在国内外从事专业和管理工作的基本语言需要。

教育部考试中心 NATIONAL EDUCATION EXAMINATIONS AUTHORITY



考查要点和你的作答情况分析

第一部分 听力

考查要点	题数	答对
1. 理解主旨要义	4	3
2. 获取事实性的具体信息	10	6
3. 对说话的背景、说话者之间的关系等作出简单的推断	4	2
4. 理解说话者的意图、观点或态度	2	1

D

第二部分 英语知识运用

第一节 单项填空 考查要点	题数	答对
1. 词汇辨义	2	1
2. 动词	6	5
3. 形容词、副词	2	2
4. 连词	2	1
5. 介词	1	1
6. 语用	1	0
7. 强调句	1	1
第二节 完形填空 考查要点	题数	答对
1. 句子理解	10	9
2. 上下文理解	10	8

第三部分 阅读理解

考查要点	题数	答对
1. 理解主旨要义	4	2
2. 理解文中具体信息	8	6
3. 根据上下文推测生词的词义	2	2
4. 作出简单判断和推理	3	2
5. 理解文章的基本结构	1	1
6. 理解作者的意图和态度	2	1

第四部分 写作

第一节 短文改错	题数	答对
	10	4
第二节 书面表达	满分	得分
	25	10

各部分得分

E

左图中柱形的长度表示你在考试各部分的得分。通过与及格分的比较，你可以看出自己的具体表现。各部分得分非原始分，而是加权、等值后的得分。

你在 PETS 量表中所处的位置

F

从以上量表，你可以清楚地看出自己离 PETS 本级别及较高级别的合格分数线有多大距离。

报告单编号：1111111111