An Overview of Models of Speaking Performance and Its Implications for the Development of Procedural Framework for Diagnostic Speaking Tests

Zhongbao Zhao[1]

[1] Foreign Languages College, Zhejiang Gongshang University, Hangzhou, China

Correspondence: Zhongbao Zhao, Foreign Languages College, Zhejiang Gongshang University, Hangzhou, China, No.18, Xuezheng Str., Foreign Languages College, Zhejiang Gongshang University, Xiasha University Town, Hangzhou, China. E-mail: Michaelzhao998@hotmail.com

Received: January 4, 2013 Accepted: January 9, 2013 Online Published: January 27, 2013

doi:10.5539/ies.v6n3p66 URL: http://dx.doi.org/10.5539/ies.v6n3p66

The research is financed by National Research Center for Foreign Language Education in China (Project ZGWYJYJJ2012A25)

Abstract

This paper aims at developing a procedural framework for the development and validation of diagnostic speaking tests. The researcher reviews the current available models of speaking performance, analyzes the distinctive features and then points out the implications for the development of a procedural framework for diagnostic speaking tests. On basis of the overview and the experience gained from the development of Diagnostic College English Speaking Test, The researcher elaborates on the process of the development of the procedural framework. The framework is composed of four major phases: phase 1 needs analysis, phase 2 test design, phase 3 test piloting, administration and validation, and phase 4 test impact. Each phase is further divided into several steps, each having concrete aims and objectives. Literatures relevant to each component of the procedural framework are also recommended by the researcher to help teachers to develop their own diagnostic tests. The paper concludes with suggestions for further research on the development of computerized diagnostic assessment systems.

Keywords: models of speaking performance, diagnostic testing, diagnostic speaking test, framework for development of diagnostic test

1. Introduction

With the flourishing of researches in the field of second language speaking assessment, several models depicting second language performance and speaking performance in particular have been established (e.g., Fulcher, 2003; McNamara,1996; Milanovic & Saville, 1996; Skehan, 1998). These models describe the relationship between the construct being measured, the tasks used to operationalize the test construct and the assessment of the performances that are used to make inferences about test-takers' language ability. This paper will give an overview of four models of speaking performance, point out the similarities and differences between them and highlight their contributions to the development of diagnostic speaking tests. Based on the overview and the experience gained from the development of Diagnostic College English Speaking Test, the researcher elaborates on the process of the development of the procedural framework. Literatures relevant to each component of the procedural framework are also recommended by the researcher to help teachers to develop their own diagnostic tests.

The specific questions to be addressed by the study are as follows:

1) What are the common features and distinctive characteristics of currently available models of speaking performance?

2) What are the implications of the overview of models of speaking performance for the development of the procedural framework?

3) What is the procedural framework suggested for the development and validation of diagnostic speaking tests?

2. An Overview of Models of Speaking Performance

2.1 Milanovic and Saville's Model

Milanovic and Saville (1996) provide a useful description of the variables that interact in performance testing and suggest a conceptual framework for carrying out different aspects of research. The major factors of this framework include: specifications and construct, the test-taker, the examiner, the assessment criteria, the task, and the interaction between these elements (see Figure 1). This framework highlights the factors that must be considered when designing a test from which particular inferences are to be drawn about performances. It is argued that all the factors illustrated in the model may pose potential threats to the reliability and validity of the test to be designed (O'Sullivan *et al.*, 2002).

Three phases of a speaking test are illustrated in this model. The first phase is the development of a speaking test, in which test developers have the greatest responsibility for the reliability and validity of a test. The component of construct and specifications in Milanovic and Saville's model can be regarded as an operationalization of Alderson's validation specification.

The second phase is the administration of a speaking test, in which candidates' speech samples are elicited by test tasks and evaluated by the examiners under the examination condition. Factors affecting candidates' performance and the interactions between them are displayed in the model. It shows that candidates' performances are influenced by their knowledge and ability, the examination conditions, the tasks and the assessment criteria. However, the personal characteristics of test takers are not included in this model. Besides, some attributes of the examiners such as age, gender and the like that may affect candidates' performance are also neglected.

The third phase is the marking of candidates' performance. Examiners rate the candidates' performance in accordance with the assessment criteria provided under the assessment condition. The examiners' marking is mainly affected by their knowledge and ability, the task, the assessment criteria, and the assessment condition and training. It is argued that training can improve the examiners' marking consistency.

O'Sullivan *et al (2002)* consider this framework as one of the earliest and most comprehensive one in elaborating variables involved in a performance test.

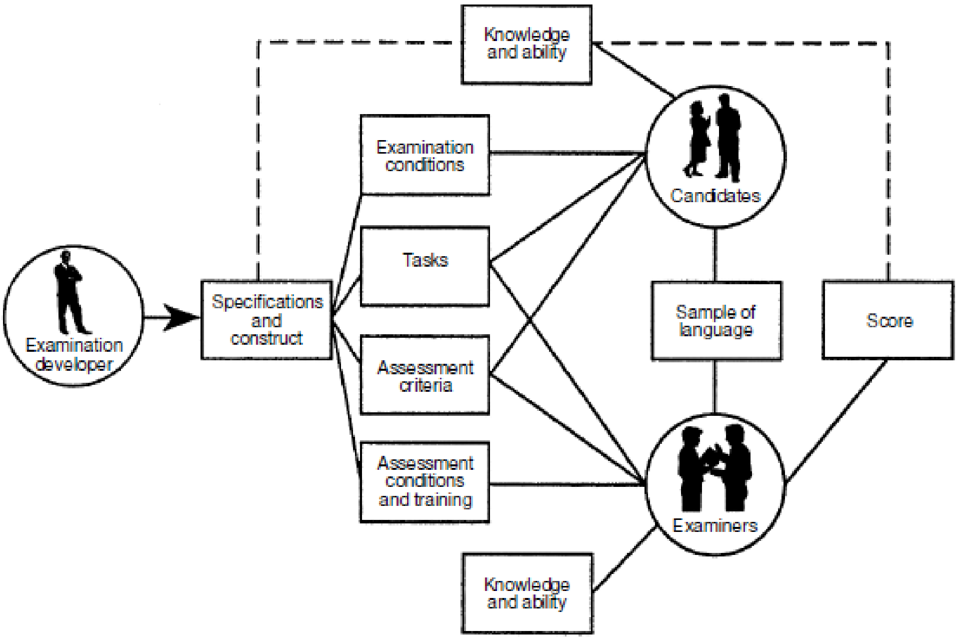


Figure 1. A conceptual framework for performance testing (Milanovic & Saville, 1996: 16)

2.2 McNamara's Model

McNamara (1996) puts forward a model to illustrate the interactional nature of performance assessment with a focus on the rating process. It is argued that the communicative language ability (CLA) model proposed by Bachman (1990) is the underling linguistic theory of McNamara's model.

This model describes how the interlocutor elicits candidate's performance and how the rater rates the candidate's performance (see Figure 2). It places performance in a central position. The arrows indicate that performance is influenced by several factors, including the tasks, which drive the performance and the raters who judge the performance using rating scales and criteria. The final score can therefore only be partly seen as a direct index of performance. The performance is also influenced by other contextual factors like, for example, the test taking conditions. The interactions between candidate and task, candidate and interlocutor, and candidate and rater are displayed in the model. It is argued that the interpretation of test scores should take these different types of interactions into consideration.

In addition, this model also includes two processes of a speaking test. One is the candidate's test-taking process; the other is the rater's rating process. The former shows how the candidate interacts with the interlocutor to finish the task provided, while the latter reveals how the rater marks the performance of the candidate with reference to the scales and criteria. These two processes are of crucial importance to a speaking test in terms of reliability and validity.

Compared with Milanovic and Saville's (1996) model, McNamara's model is more concise and much easier to follow. There are some changes in the terminology. And some new concepts like rater, interlocutor and rating are introduced in the model.

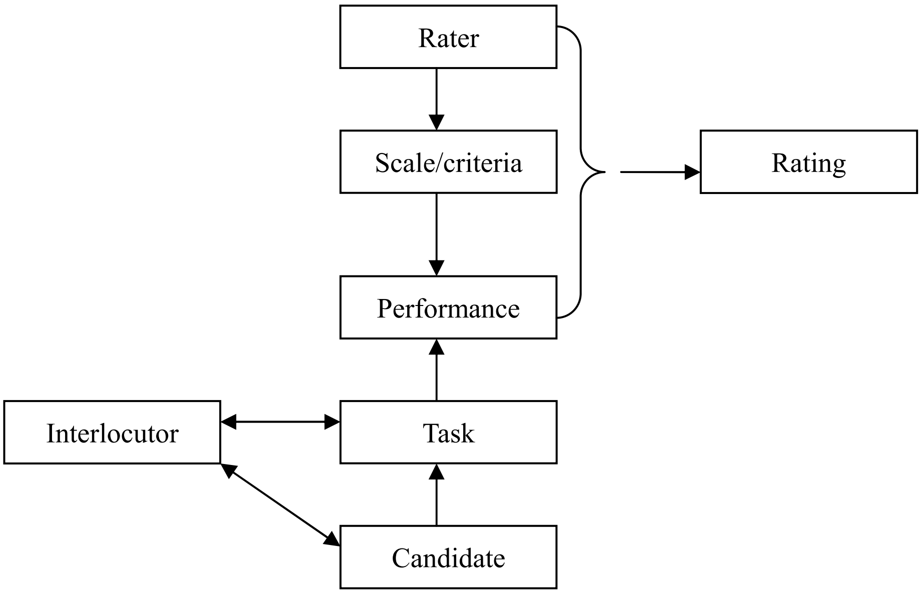


Figure 2. Proficiency and its relationship to performance (McNamara, 1996: 86)

2.3 Skehan's Model

Skehan (1998) proposes a model of oral test performance with an attempt to describe a number of additional variables for the purpose of exploring the complexity of the speaking event more comprehensively (see Figure 3). This model refines McNamara's (1996) model in two ways. Firstly, Skehan argues that tasks need to be analyzed further to account for task characteristics and task implementation conditions. Secondly, McNamara's model does not account for what Skehan calls the dual-coding capacities of the learner. Skehan (1998: 171) argues that "second language learners' abilities require not simply an assessment of competences, but also an assessment of ability for use". Fulcher (2003) comments that one distinctive feature of Skehan's model is that it depicts three factors mainly affecting test scores. These factors are stated as follows: "the interactive conditions of the performance, the abilities of the test taker, and the task (as described by conditions or characteristics) used to elicit the performance" (Fulcher, 2003: 113).

O'Sullivan et al. (2002) points out that the testing of spoken language is exclusively performance-based, thus research might be expected to focus on factors that systematically affect that performance, with an additional focus on other factors that affect test outcomes and uses. The arrows in Figure 3 indicate that task is influenced by several factors, including task qualities and task conditions, which will determine the difficulty of the task. However, characteristics of test-takers that may affect the test performance are not discussed in detail in this model.

The description of task qualities and task conditions in the model makes it much easier for language testers to develop and compare tasks. Studies have proved that task difficulty can be changed by manipulating the task conditions and task qualities (e.g, McNamara, 2002; Norris, 2002).

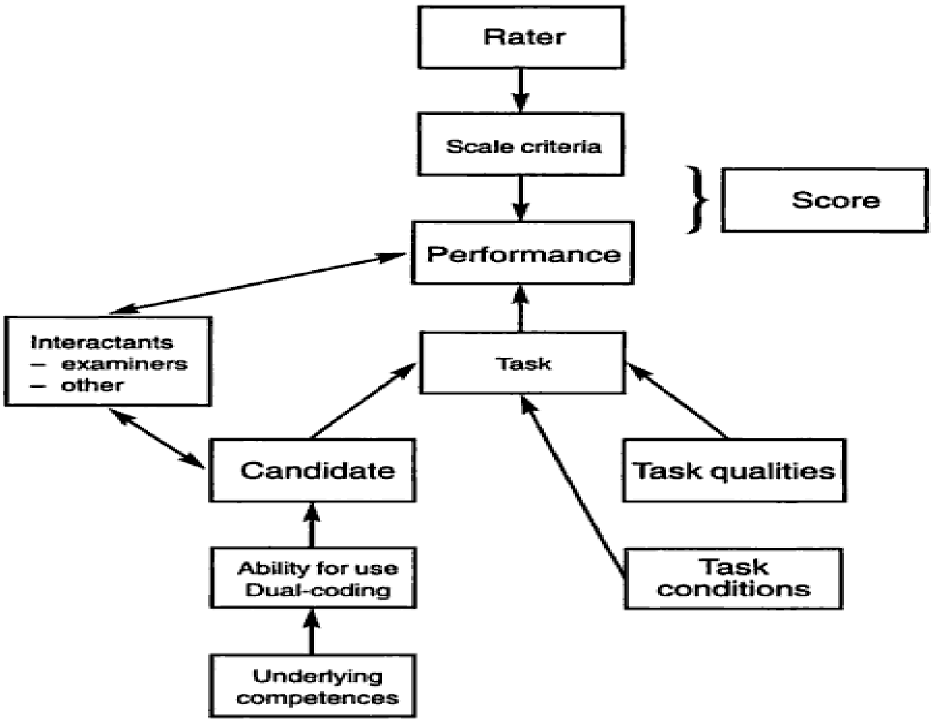


Figure 3. Model of oral test performance (Skehan, 1998: 172)

2.4 Fulcher's Model

Fulcher (2003) revises Skehan's model and puts forward an expanded model of speaking test performance for the purpose of expanding the understanding of the role of the construct, task and scale in the meaning of the score (see Figure 4). Fulcher's model places construct definition at the heart of rating-scale and band-descriptor design, the understanding of what constructs are being assessed through the performance of a test taker, and the inferences that are drawn from scores. This model shows the effect of the nature of rating scale and the scoring philosophy on test score and its meaning. In the case of the raters, Fulcher acknowledges that rater training and rater characteristics play a role in the scoring process. He also indicates that there is an interaction between the rating scale and a test taker's performance which results in the score and any inferences that are made about the test taker. Fulcher further acknowledges the importance of context in test performance by including local performance conditions. Like Skehan (1998), Fulcher also includes aspects that influence the task in his model. Among these are the task orientation, goals, and topics, as well as any context-specific task characteristics or conditions. Finally, Fulcher's model shows a number of variables that influence the test taker. These include any individual differences between candidates (like personality), their ability for real-time processing and any task-specific knowledge or skills they might possess. Comparatively speaking, Fulcher's model is much more comprehensive and exhaustive than the other three models.

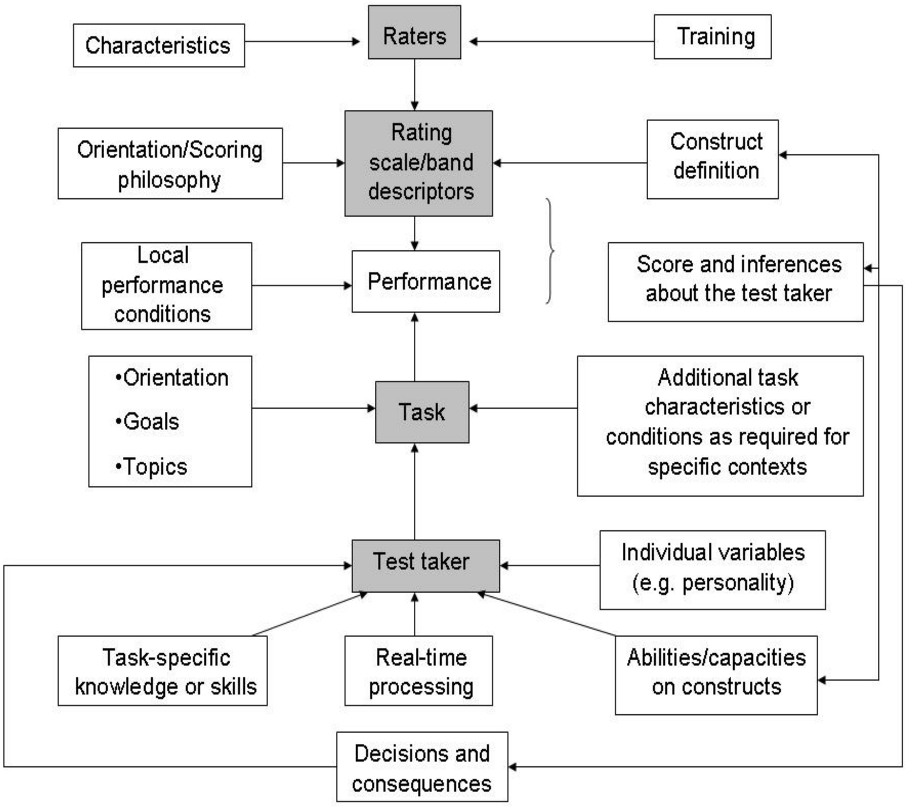

Characteristics
Raters
Training
Orientation/Scoring philosophy
Rating scale/band descriptors
Construct definition
Local performance conditions
Performance
Score and inferences about the test taker
•Orientation
•Goals
•Topics
Task
Additional task characteristics or conditions as required for specific contexts
Test taker
Individual variables (e.g. personality)
Task-specific knowledge or skills
Real-time processing
Abilities/capacities on constructs
Decisions and consequences

Figure 4. An expanded model of speaking test performance (Fulcher, 2003: 115)

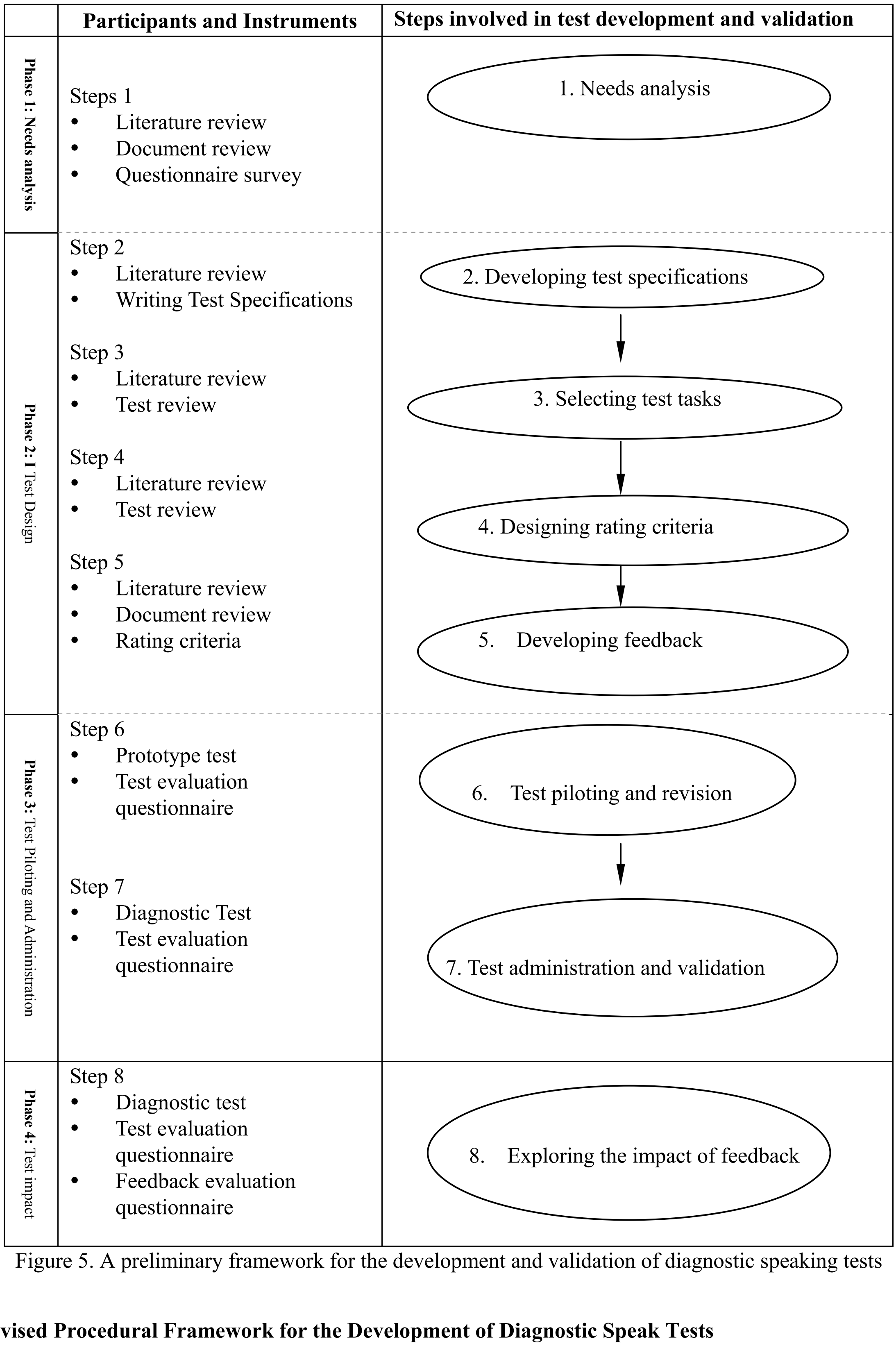
From the above review of the models of language performance, it can be seen that linguists' understanding of speaking test performance has undergone changes in the past two decades. The possible impetuses that may facilitate the expansion of these models are posited by the researcher as follows:

- The development of theories in the field of linguistics, language testing, language teaching and second language acquisition provides important implications for the understanding of the speaking construct.
- Research of the nature of spoken language enables language testers to define the construct more precisely.
- The refinement of experimental design and statistical analysis instruments makes it possible for language testers to assess the interactions between variables affecting test-takers' performance.

2.5 A Preliminary Procedural Framework for the Development of Diagnostic Speaking Tests

Reviews on models of speaking performance indicate that a speaking test is mostly composed of two essential cycles: rating of test takers' performance and evaluation of the test. For diagnostic tests, it is argued that "the impact on learning and teaching" should be treated as an important aspect in the validation of the test.

In light of the models of speaking performance reviewed and taking into consideration the unique features of diagnostic tests, the researcher proposed a preliminary framework for the purpose of facilitating the development and validation of diagnostic speaking tests (see Figure 5). This framework includes the unique features of diagnostic testing such as the provision of feedback and study of the impact of feedback on learning and teaching, etc. The framework comprises four major phases: phase 1 needs analysis, phase 2 test design, phase 3 test piloting, administration and validation, and phase 4 test impact. Each phase is further divided into several steps. In addition, the participants and instruments that may be employed in each phase are also described.



	Participants and Instruments	Steps involved in test development and validation
Phase 1: Needs analysis	Steps 1 • Literature review • Document review • Questionnaire survey	1. Needs analysis
Phase 2: I Test Design	Step 2 • Literature review • Writing Test Specifications Step 3 • Literature review • Test review Step 4 • Literature review • Test review Step 5 • Literature review • Document review • Rating criteria	2. Developing test specifications ↓ 3. Selecting test tasks ↓ 4. Designing rating criteria ↓ 5. Developing feedback
Phase 3: Test Piloting and Administration	Step 6 • Prototype test • Test evaluation questionnaire Step 7 • Diagnostic Test • Test evaluation questionnaire	6. Test piloting and revision ↓ 7. Test administration and validation
Phase 4: Test impact	Step 8 • Diagnostic test • Test evaluation questionnaire • Feedback evaluation questionnaire	8. Exploring the impact of feedback

Figure 5. A preliminary framework for the development and validation of diagnostic speaking tests

3. Revised Procedural Framework for the Development of Diagnostic Speak Tests

Based on the empirical study of the development and validation of the Diagnostic College English Speaking Test (DCEST)(Zhao,2011), the researcher revised the preliminary framework and established a procedural framework of diagnostic speaking test development and validation with a view to providing guidance to college English teachers who need to develop diagnostic speaking tests tailored to their teaching contexts. The revised framework is composed of four major phases: phase 1 needs analysis, phase 2 test design, phase 3 test piloting, administration and validation, and phase 4 test impact. Each phase is further divided into several steps, each having concrete aims and objectives.

Compared with the preliminary framework, this procedural framework, specifically designed for the development and validation of diagnostic language tests, is more systematic and practical because it describes the sequence of data collection and methods for collecting each type of data for each step at each phase of the test development and validation cycle. By following the procedure, teachers can design and implement diagnostic language tests in a phase-by-phase and step-by-step manner. One of the unique features of this framework is that it describes the method of writing detailed feedback descriptors. The provision of detailed feedback report is different from the traditional score report in that the feedback report provides a profile score with detailed information on the strengths and weaknesses of students' language ability. And the most distinguishing feature of this framework is that it provides a method of data triangulation at phase 4 for the purpose of validating the diagnostic tests with a focus on the consequential validity.

The revised procedural framework is displayed in the form of a flowchart (see Figure 6). The left column of the flowchart describes the participants of and the instruments used at each phase, the middle column illustrates the steps to be followed in each phase of the procedure, and the right column highlights the research focuses of each phase.

3.1 Phase 1: Needs Analysis

Due to the diagnostic nature of the test to be developed, a detailed analysis of the teaching and learning needs in the educational context concerned should be conducted. Questionnaire surveys and structured interviews can be employed at this stage for data collection. A student self-assessment checklist may also help collect useful supplementary information on students' learning difficulties. Work in Phase 1 is hoped to lay the foundation for defining the construct of the diagnostic test and formulating the test's specifications.

In addition, the CLA model (Bachman, 1990; Bachman & Palmer, 1996), speech production models (Bygate, 1987; Douglas, 1997; Levelt, 1989), and frameworks of speaking construct (Fulcher, 2003) could provided useful references for the defining of the test construct.

3.2 Phase 2: Test Design and Operationalization

Phase 2 focuses on the operationalization of the construct of the diagnostic test identified in Phase 1, which is crucial to ensure test reliability and validity. What follows is a detailed description of the four steps involved in the operationalization of the test's construct.

Step 1 Developing test specifications

Test specifications are the blueprint for test development. They state clearly the test purpose, the definition of the test construct, types of tasks to be used, time allotment for each test tasks, instructions for test takers, etc. Our experience with the DCEST's test specifications shows that Alderson *et al.* (1995) framework and Luoma (2004) modular specifications are very good examples to be followed in the development of the specifications for a diagnostic English speaking test.

Step 2 Selecting appropriate test tasks

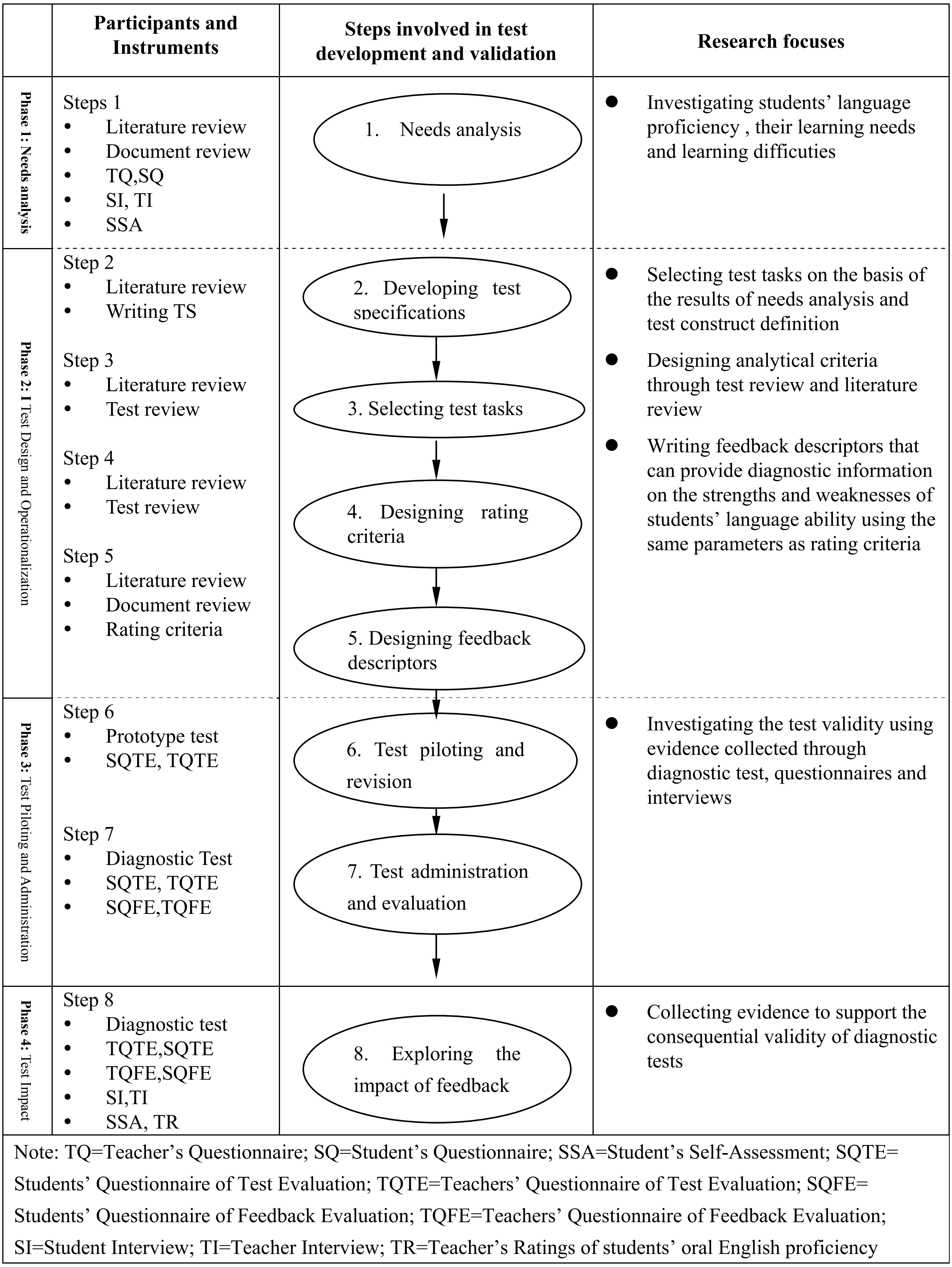
Test tasks are the most direct operationalization of a test's construct. Appropriate types of test tasks should be selected on the basis of the test construct and test specifications developed in Step 1. Our experience in the selection of test tasks for the DCEST reveals that studies on speaking tasks (Luoma, 2004; Underhill, 1987; Weir, 1990, 1993) offer good references for test task selection.

Step 3 Designing rating criteria

Rating criteria are also the operationalization of a test's construct (Fulcher, 2003). It is argued that an analytic scale is more efficient in providing diagnostic information (Cohen, 1994). The type of rating criteria should be designed in accordance with the test construct defined in Phase 1 and test tasks selected in the previous step. An analysis of the rating criteria and scales used by the tests reviewed in Phase 1 can be a valuable reference. The present study found that a broad range of rating criteria should be used to maximize the diagnostic function of a diagnostic English speaking test. Teachers are encouraged to design their own rating criteria based on their definition of the test construct and their choices of tasks.

Step 4 Designing feedback descriptors

Feedback is defined by Ramaprasad (1983: 4) as "information about the gap between the actual level and the reference level of a system parameter which is used to alter the gap in some way." It is argued that feedback is an integral part of diagnostic tests. The feedback report of a diagnostic English speaking test must provide useful information on the strengths and weaknesses of students' oral English ability.



	Participants and Instruments	Steps involved in test development and validation	Research focuses
Phase 1: Needs analysis	Steps 1 • Literature review • Document review • TQ,SQ • SI, TI • SSA	1. Needs analysis	● Investigating students' language proficiency , their learning needs and learning difficuties
Phase 2: I Test Design and Operationalization	Step 2 • Literature review • Writing TS Step 3 • Literature review • Test review Step 4 • Literature review • Test review Step 5 • Literature review • Document review • Rating criteria	2. Developing test specifications 3. Selecting test tasks 4. Designing rating criteria 5. Designing feedback descriptors	● Selecting test tasks on the basis of the results of needs analysis and test construct definition ● Designing analytical criteria through test review and literature review ● Writing feedback descriptors that can provide diagnostic information on the strengths and weaknesses of students' language ability using the same parameters as rating criteria
Phase 3: Test Piloting and Administration	Step 6 • Prototype test • SQTE, TQTE Step 7 • Diagnostic Test • SQTE, TQTE • SQFE,TQFE	6. Test piloting and revision 7. Test administration and evaluation	● Investigating the test validity using evidence collected through diagnostic test, questionnaires and interviews
Phase 4: Test Impact	Step 8 • Diagnostic test • TQTE,SQTE • TQFE,SQFE • SI,TI • SSA, TR	8. Exploring the impact of feedback	● Collecting evidence to support the consequential validity of diagnostic tests

Note: TQ=Teacher's Questionnaire; SQ=Student's Questionnaire; SSA=Student's Self-Assessment; SQTE= Students' Questionnaire of Test Evaluation; TQTE=Teachers' Questionnaire of Test Evaluation; SQFE= Students' Questionnaire of Feedback Evaluation; TQFE=Teachers' Questionnaire of Feedback Evaluation; SI=Student Interview; TI=Teacher Interview; TR=Teacher's Ratings of students' oral English proficiency

Figure 6. A revised procedural framework for the development and validation of diagnostic tests

As for the development of a feedback report, Luoma (2004) suggests the use of rating checklists to develop more structured feedback mechanisms for speaking assessment. Therefore, it is recommended that the feedback be designed using the same parameters as rating criteria. The feedback of the DCEST was designed in such way, that is, the 10 rating criteria of the DCEST were used as the parameters or dimensions to develop feedback descriptors. The results indicated that the feedback report of the 10 parameters was considered as quite accurate and useful.

3.3 Phase 3: Test Piloting, Administration and Validation

This phase mainly involves the revision, administration and validation of the diagnostic language test under development in real educational settings. There are two major steps in this phase: Step 1 test piloting and revision, and Step 2 test administration and validation.

Step 1 Test piloting and revision

The aim of a pilot test is to improve the design of the test and the quality of test items, and to examine the practicality of the test. During the pilot testing, for example, attention needs to be paid to students' understanding of the test instructions, their perception of the test difficulty, their evaluation of task familiarity and test time allotment, and their comment on the usefulness of test feedback, etc. Samples for a pilot test should include a sufficient number of students at different levels of language proficiency. Through piloting the test, teachers can also get themselves familiar with the rating process and provide useful feedback on the rating procedure, rating criteria and the rating scale.

Step 2 Test administration and validation

At the stage of test administration, students should be provided with a profile score and a detailed feedback report. As part of the test validation at the a posteriori stage, the questionnaires of test evaluation and feedback evaluation can be designed to collect information about students' and teachers' comments on the validity of the test, with focuses on the effectiveness of each test task and the usefulness of the feedback provided. Information gathered after the first administration of the test guides further test revisions and the refinement of rating criteria and feedback descriptors.

3.4 Phase 4: Test Impact

The provision of diagnostic feedback is the unique feature of diagnostic tests. It is of great importance for teachers and students to evaluate the effects of the diagnostic feedback on their teaching and learning in a real educational context. As a type of evidence of consequential validity, the investigation of the impact of feedback on teaching and learning should be conducted over a period of time. The triangulation method should be used for data collection and analysis. Types of data to be collected include diagnostic test scores, students' self assessment, and teacher's ratings of students' English proficiency. In addition, structured interviews with students and teachers should also be conducted to obtain more information on the impact of feedback on English teaching and learning.

4. Conclusion

The review of the models of speaking performance in this paper reveals that they all share the same essential elements, namely, the test-taker, the examiner, the assessment criteria, and the task, but with some variations in the terminology. In addition, these models can serve as a general framework for language testers to formulate explicit hypotheses about the relationship between test takers' performance, rater behavior and test scores, which may have the following implications for the development of the procedural framework for the development of diagnostic speaking tests:

1) They shed some light on the process of speaking test performance and evaluation.

2) They illustrate the factors that may influence test-takers' performance and the interactions between them.

3) They provide guidelines for the defining of the test's construct, the construction of test specifications, and the design of test tasks and rating scales.

4) They can be used as a framework for test validation.

From the practical perspective, the procedural framework established in this study can serve as guidelines for college English teachers to develop diagnostic language tests with reference to the curricular goals and learning objectives for students at the tertiary level in China. This framework provides a systematic procedure for the development and validation of a diagnostic test.

With the advancement of information technology, future research can focus on development of computerized diagnostic assessment systems and in the process the procedural framework can be refined. The computerized assessment system should contain all information about test specifications, skill specifications along with the diagnostic feedback of each type of test tasks. The specifications should provide a comprehensive description about the skills and the test tasks aligned with the learning objectives. This system will provide teachers with a more effective way to develop and use their own diagnostic tests.

References

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language Test Construction and Evaluation.* Cambridge: Cambridge University Press.

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

Bachman, L. F., & Palmer, A. S. (1996). *Language Testing in Practice: Designing and Developing Useful Language Tests*. Oxford: Oxford University Press.

Bygate, M. (1987). *Speaking*. Oxford: Oxford University Press.

Cohen, A. (1994). *Assessing Language Ability in the Classroom* (2nd ed.). Boston: Heinle and Heinle.

Douglas, D. (1997). *Testing Speaking Ability in Academic Contexts: Theoretical Considerations*. TOEFL Monograph Series Ms. 8. Princeton, NJ: Educational Testing Service.

Fulcher, G. (2003). *Testing Second Language Speaking*. London: Longman/Pearson Education.

Levelt, W. (1989). *Speaking: From Intention to Articulation*. Cambridge, MA: MIT Press.

Luoma, S. (2004). *Assessing Speaking*. Cambridge: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511733017

McNamara, T. (2002). Discourse and Assessment. *Annual Review of Applied Linguistics*, *22,* 221–242. http://dx.doi.org/10.1017/S0267190502000120

McNamara, T. F. (1996). *Measuring Second Language Performance*. London: Longman.

Milanovic, M., & Saville, N. (1996). Introduction. Performance Testing, Cognition and Assessment. *Studies in Language Testing, 3*, 1-17. Cambridge: University of Cambridge Local Examinations Syndicate.

Norris, J. M. (2002). Interpretations, Intended Uses and Designs in Task-based Language Assessment. *Language Testing*, *19*(4), 337-346. http://dx.doi.org/10.1191/0265532202lt234ed

O'Sullivan, B., Saville, N., & Weir, C. (2002). Using Observation Checklists to Validate Speaking-test Tasks. *Language Testing, 19*(1), 33-56. http://dx.doi.org/10.1191/0265532202lt219oa

Ramprased, A. (1983). On the Definition of Feedback. *Behavioral Science*, *28,* 4-13. http://dx.doi.org/10.1002/bs.3830280103

Skehan, P. (1996). A Framework for the Implementation of Task Based Instruction. *Applied Linguistics*, *17,* 38-62. http://dx.doi.org/10.1093/applin/17.1.38

Skehan, P. (1998). *A Cognitive Approach to Language Learning*. Oxford: Oxford University Press.

Underhill, N. (1987). *Testing Spoken Language: A Handbook of Oral Testing Techniques*. Cambridge: Cambridge University Press.

Weir, C. (1990). *Communicative Language Testing*. Exeter: Exeter University Press.

Weir, C. (1993). *Understanding and Developing Language Tests*. Hertfordshire: Prentice Hall International Ltd.

Zhao, Z. B. (2011). *Development and Validation of the Diagnostic College English Speaking Test* (Unpublished doctoral dissertation, Shanghai Jiao Tong University).