ASIAN EFL JOURNAL
The EFL Professional's Written Forum

Title

Validating a Simulated Test of CET 4

Author

Yang Miao

Shantou University Medical College, China

Abstract

A simulated test of CET4 (College English Test, Band 4) was validated to check if it served the specific purposes of predicting and diagnosing. The study data came from a CET 4 simulated test sat by a class of sophomores who were to take a CET 4 test one month later. Based on Messick's framework of validation, the test's content coverage and representativeness were checked, and correlation analyses including inter-consistency reliability, item correlation, factor analysis and item analysis were computed. The analysis results show that the test is of modest reliability and validity, with the most serious problem in the reading section, which had too many misfit items and failed to effectively test the candidates' discourse reading ability. The contextual difficulties or inadequacy of efforts in other aspects of the validation framework implied a very unsatisfactory situation of simulated test practice in the Chinese context. It is stated that for a simulated test to effectively fulfill its purposes of predicting and diagnosing, trial tests and post hoc analysis are essential, and empirical investigations into the process of test taking, the effects of coaching and practice and the motivation problems should be advocated, and effective remedial support should be provided afterwards to ensure the positive washback of such a test.

Keywords*:* Messick's validation framework； Reliability; Validity; Test practice

Introduction

A simulated test paper of CET 4 (College English Test, Band 4) is evaluated in terms of its reliability and validity. The discussion is guided by the specific purposes of a simulated test and based on Messick's framework of validation. Possible approaches to validating the test paper are suggested and statistical measurements are conducted to get the related data. The introduction section begins with a brief introduction to CET and the simulated test, and moves to the explanation of Messick's framework of validation.

CET 4 and the simulated test

Put into practice in 1987, College English Test Band 4 & Band 6 (hereafter CET4 & CET6) is a national standardized English examination sponsored by the Higher Education Department of the Ministry of Education in China and administered by the National College English Testing Committee. It is a criterion-related norm-referenced test (Yang & Weir, 1998). The test criteria are based on the College English Syllabus which was designed in 1985 by the Education Department to guide the English teaching at university level. According to the Syllabus, the designing of CET should strike a balance between linguistic knowledge and linguistic competence, between accuracy and fluency, between semantic level and discourse level, and between conceptual abilities and expressive abilities (College English Syllabus, 1985). It is maintained that reliability and validity are important indexes of test quality in a standardized examination and increasing the test validity is the pivot of modern language testing research (Yang & Weir, 1998). In order to ensure scientific, objective, unified and standardized testing, the design of CET strictly follows the procedures of questions setting, initial examining, predicting, item analyzing, further examining, test composing, testing, scoring, statistic analyzing and bank building.

To check the validity of CET, the National College English Test Committee conducted a 3-year project (from 1995 to 1998) with the British Council, in which the construct validity, content validity, concurrent validity and face validity of CET were studied through comparison tests and large-scale surveys. It is concluded that CET is of high reliability (0.90) and validity (92% of the teacher subjects believe CET reflects students' actual English proficiency levels, 86% think the test contents are reasonable) (For more details of the project results, please see Yang & Weir, 1998).

Candidates of CET are undergraduates and postgraduates who have gone through with a general English course based on the College English Syllabus. Formerly, this test is composed of five components: listening, reading, vocabulary & structure, cloze and writing. Except for writing, all test items are in objective multiple-choice format. Since 1996, new test tasks such as compound dictation (a combination of partial dictation and dicto-comp), short answers to questions and English-Chinese translation have been adapted to measure students' pragmatic English competence.

In the last two decades, CET has developed to be one of the most important English exams in China. In 2005, as many as 11 million students participated in CET. Its results are regarded as authoritative evidence of English proficiency level and a pass in CET is one of the criteria for graduation in many institutions. To help students achieve higher scores in CET, simulated tests become common practice. The candidates usually have several simulated tests before CET. But contrary to the ubiquitous practice of simulated tests, its validation is seldom questioned and studied. In most cases, the test papers are ready-made, taken from CET preparation books published by different presses. As a result, the quality of simulated test papers is not guaranteed. And most of the time, the test designers give little or even no explanation of how the test is designed. The claim that the test papers follow the CET test specifications seems to be self-evident. And more often than not, users of the test papers (the English teachers) just use the papers, score the results and arrange another test without statistic treatment and analysis. Contrary to the little effort in test validation is the great amount of time, energy and resources spent in preparing and managing the simulated tests, indicating the significance of powerfully validating CET simulated test in such highly test-oriented context as China.

Messick's framework of validation

The key to understanding Messick's framework is the concept of unitary validity. A conventional view of validity identifies different types of validity, i.e. face validity, content validity, criterion-related validity and construct validity (Hughs, 1989). But according to Messick, such a view is inadequate (Bachman, 1990; Wood, 2001). He distinguishes a number of complementary facets of validity within a unified theory of validity, in which the social nature of assessment (values and consequences of score use) is a key feature and construct validity is essential in each aspect (Bachman, 1990; Messick, 1996; McNamara, 2001; Chapelle, Jamieson & Hegelheimer, 2003). In this framework, six distinguishable aspects of validation are identified to provide 'an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores' (Messick, 1989, p13 cited in Bachman, 1990, p236).

Judgmental/logical analyses: involving the discovery of content relevance and representativeness which demonstrates that a test is relevant to and covers a given area of content or ability

Correlation analyses: involving the quantitative analyses to gather evidence in support of the test

scores and its interpretation, such as inter-consistency reliability, item correlation, multitrait-multimethod design, factor analysis and item analysis.

Analyses of process: involving the qualitative analyses to investigate the processes of test taking themselves employing approaches such as protocol analysis (concurrent or retrospective verbal reports), computer modeling, analysis of response time, analysis of reasons given by test takers for choosing a particular answer and analysis of systematic errors

Analyses of group difference and change over time: involving cross-sectional and longitudinal studies to examine the extent to which score properties and interpretations generalize to and across population group, settings and tasks.

Manipulation of tests and test conditions: involving getting the empirical knowledge about the effects of test intervention such as instruction and coaching that alter test scores in theoretically predicted ways

Test consequences: involving the evaluation of value and intended or unintended consequences of score interpretation, which concerns issues that are associated with bias in scoring and interpretation, with unfairness in test use, and with positive or negative washback effects on teaching and learning.

(Bachman, 1990; Messick, 1996)

In view of the complexity of the validation process, the suggestion that a test's use or purpose should serve as a guide to validation is accepted (Worthen, Borg & White, 1993; Read & Chapelle, 2001). But this does not mean that the validation of low-stake tests is not as essential as that of high-stake tests such as entrance tests or selection tests. Many studies of low-stake tests, such as Chapelle, Jamieson & Hegelheimer's (2003) study of a web-based ESL test, Guerrero's (2000) study of a Spanish proficiency exam and Wall, Clapham & Alderson's (1994) and Fulcher's (1997) evaluation of placement tests, also conduct rigorous validation. As one of the low-stake tests, the validation of a simulated test is underlined by its specific purposes, i.e. to locate candidates' proficiency, to predict possible pass rate and to diagnose existing problems in teaching and learning so that remedial support will be provided. In light of these purposes and within Messick's framework, the validation of a simulated test paper **ideally** includes

a. the analyses of test content in terms of the test (in this case CET) specifications;

b. the quantitative analyses of test scores and interpretation;

c. the qualitative analyses of test processes to identify individuals' test strategies;

d .the examination of possible discrepancies between different groups of candidates or between different times of test taking among the same group of candidates;

e. the investigation of the effects of examination practice, and

f. the consideration of both positive and negative washback such as remedial help and motivation problems.

But simulated tests are relatively little mentioned by research literature despites the fact that they are popular and important test practice in many EFL (English as a Foreign Language) context, which means that this study is only a trial one in this aspect. In later sections, possible way of validation will be discussed, but unfortunately most of the analyses are just armchair strategies and deserve further research.

Methods

Subjects

The subjects of this study are a class of sophomores in a medical college in Southeast China who are to sit for CET4 one month later. Of the 43 students in this class, 5 fail to finish the writing component. So finally only 38 complete sets of data are collected and analyzed.

Procedures

The simulated test was organized as scheduled, using paper taken from a collection of CET simulated test papers without adaptation (Luo, 2003). The test tasks are specified in Table 1. After the answer sheets were collected, the MC items were scored by optical mark reader (OMR). For Part V (writing), all essays were double marked by two independent raters using the CET rating criteria, which is holistic and based on scales ranging from 2 to 14 points. The average of the two raters' scores is taken as the candidates' final scores for writing. Statistical analyses were conducted using the Statistical Package for Social Sciences (12.0) and Gitest 3+ System, a software designed by GuangDong University of Foreign Studies in China for item analysis.

Table 1 Composition of the simulated test paper

Components	N0.of Items	Points	Weighting	Time (minutes)
Part I: Listening				
Section A: Short Conversations	10	10	10	10

Section B: Compound Dictation	1	10	10	10
Part II: Reading Comprehension	20	20	40	35
Part III: Vocabulary & Structure	30	30	15	20
Part IV: Cloze	20	20	10	20
Part V: Writing	1	15	15	30
Total	80+2	105	100	125

Analysis

The test paper was reviewed and checked against the test specifications to see its content coverage and representativeness. To evaluate the reliability of this test, Conbach alpha was calculated to examine the overall reliability. Subtest and inter-rater reliability coefficients were also provided. Product-moment correlation coefficients were computed between subtests as well as between subtests and the total score and factor analysis was conducted, both of which provided validity evidence. In order to gather details of the test items, item analysis was done to find out the difficulty and discrimination of each item and identify misfit items for further discussion.

Results and discussion

Judgmental/logical Analyses

Because this test paper is to simulate CET 4, it is essential that it cover and represent the content and language abilities that are designated in the CET4 test specifications. Checked against the CET 4 test specifications provided by Yang & Weir (1998, pp. 198-200), this test paper appears to basically cover and represent the most important sub-skills of listening and reading comprehension that are identified in the specifications. As for the test points in vocabulary & structure, they are also representative to a large extend despites the fact that it is impossible to cover all of them in a single sample test. As the specifications require, some difficult grammatical points for Chinese learners are tested, such as verb forms including tenses and voices, non-finite verbs, adjective clause, noun clause and subjunctive mood. The only problem is the proportion of vocabulary items to grammar items. According to the test specifications, only 40% of this subtest should be devoted to vocabulary items while the other 60% to grammar items (Yang & Weir, 1998). But a close analysis of this subtest shows that half of the 30 items test vocabulary. Examples of vocabulary item and grammar item from this test paper are given as follows. To ensure the accuracy of this analysis

result, the researcher consulted the English teacher who chose and scored the papers and agreement between them was made. If adjustment is made to include more grammar items and less vocabulary items in this part, the overall content validity can be improved.

Example of vocabulary item:

47. I can't _____ him from his brother. They look very much alike.

A) keep B) separate C) distinguish D) prevent

Example of grammar item:

36. If this university _____ such a good reputation, I would not have come here.

A) didn't have B) doesn't have C) hasn't had D) hadn't had

Correlation Analyses

Inter-consistence Reliability

The reliability (Cronbach alpha) of the whole test paper is 0.80. The subtest reliability coefficients (for MC items only) range from 0.04 to 0.75 (Table 2), among which the reliability coefficient of reading component is the lowest (0.04) and the MC listening component is not satisfactory (0.46). To check the reliability of the writing scores, Pearson product moment correlation analysis of the two sets of writing scores provided by two independent raters shows that two sets of scores are significantly correlated (correlation coefficient = 0.854) at the 0.01 level (2-tailed), and inter-rater reliability coefficient is 0.91, which prove the writing scores to be highly reliable. Both raters are experienced English teachers familiar with the marking system of CET. Every year when the simulated tests take place, they grade hundreds of essays of this kind. So it is reasonably unsurprising that the two raters achieve high consistency.

Basically, the whole test paper is reliable and is modestly adequate for low-stake tests. As for the subtests, listening (MC) is of low reliability and reading is extremely unreliable. More detailed discussion of these two parts is provided in the following sections.

Table 2. Item Analysis for MC Items

	Mean	SD	P	R	Misfit Items
PIA	8.95	1.23	0.89	0.46	1, 2, 3

PII	11.11	2.12	0.56	0.04	13,14,18,21,23,24,26,27,29
PIII	19.37	4.46	0.65	0.75	32,33,35,40,43,44
PIV	8.71	3.67	0.44	0.73	72

P=item facility; R=reliability

PIA=Listening Comprehension (MC); PII=Reading Comprehension; PIII=Vocabulary & Structure; PIV=Cloze.

Item Analysis

Item analyses done with Gitest identified 19 misfit items, i.e. 23.8% of the 80 MC items, which is far from satisfactory because only 5% is allowed for a reliable and valid test (LI, 1997). Of the 19 misfit items, 3 belong to listening, 9 to reading, 6 to vocabulary & structure and 1 to cloze (Table 2). The reading component is the most problematic with 9 out of 20 items misfit (45%), which is consistent with the reliability test result (reliability coefficient =0.04). The listening component is too easy (index of item facility = 0.89) with all 10 items falling on the easy and very easy scales. This explains a relatively low reliability (0.46).

Several kinds of problems are discovered with the misfit items. First, some items are too easy for the intended population so they show small discrimination figures and contribute little to the differentiation of different levels of candidates. They may be good ones for candidates of lower abilities. So retaining them for other tests or replacing them with more difficult ones are possible solutions. The second kind of problems lies in the given keys. Some items have more than one key. This problem is typical of reading comprehension items and is particularly warned of by Li (1997). Usually, reading comprehension questions are more controversial than other kinds, especially when higher levels of understanding, such as understanding implied meanings and making inferences, are concerned (ibid). As serious as the double-key problem is the wrong-key problem. To get rid of these misfit items, the extra keys should be changed to distracters and the wrong keys should be rewritten. The final problem lies in the distracters. Some distracters are so strong that they unreasonably attract more candidates than allowed. These distracters should be carefully examined and rewritten.

Item Correlation

The correlations between the subsets and the total test score are all significant at the 0.01 level,

suggesting that every one of them reasonably contributes to the measurement of the whole test (Table 3). Since the subtests are intended to test different aspects of language, they are not expected to correlate very highly with one another. The intercorrelation coefficients are supposed to fall in between 0.3 and 0.7 (Yang & Weir, 1998). But some coefficients fall out of this scope, which indicates that some items intercorrelations are not satisfactory. Listening (MC) fails to significantly correlate with the other items except reading, and cloze only significantly correlates with dictation. Moreover, reading does not sufficiently correlate with dictation, cloze and writing.

Table 3 Correlation Matrix

	PIA	PIB	PII	PIII	PIV	PV	TOTAL
PIA	1						
PIB	0.197	1					
PII	0.358*	0.018	1				
PIII	0.29	0.431**	0.392*	1			
PIV	0.311	0.408*	0.125	0.148	1		
PV	0.246	0.445**	0.147	0.509**	0.16	1	
TOTAL	0.572**	0.532**	0.737**	0.734**	0.481**	0.595**	1

* Correlation is significant at the 0.05 level (2-tailed).

** Correlation is significant at the 0.01 level (2-tailed).

PIA=Listening Comprehension (MC); PIB=Compound Dictation; PII=Reading Comprehension;

PIII=Vocabulary & Structure; PIV=Cloze; PV=Writing.

According to Oller (1979), low correlations between different tests or measures are sometimes too simply taken to mean that they are measuring different skills. For example, the intercorrelation between listening (MC) and compound dictation is low, but they are both intended to test listening skills. Possible reasons for low intercorrelation may be found in Oller's explanation:

> A low correlation may result from the fact that one of the tests is too easy or too hard for the population tested. It may mean that one of the tests is unreliable. Or that both of them are unreliable or a low correlation may result from the fact that one or both tests do not measure what they are supposed to measure (i.e., are not valid), or merely that one of them (or both) has (or have) a low degree of validity. (Oller, 1979, p56)

Item analyses show that the listening (MC) component is too easy (index of item facility=0.89, see

Table 2). This may explain the low intercorrelation between this part and the others. Meanwhile, the results that the reading component is unreliable (reliability coefficient=0.04) may also be the reason why reading fails to sufficiently correlate with dictation, cloze and writing. At this point, two assumptions about validity are concerned: One is whether the test scores accurately reflect the trait they are intended to measure; the other is whether the differences in the scores obtained by various students represent different degrees of possession of that trait (Worthen, Borg & White, 1993). If these two assumptions are confirmed, the inferences or interpretations drawn from the test scores are accurate, or we can say that the test is valid. Since the listening (MC) component is too easy for the students, it fails to represent the differences of their ability though it may reflect the trait it is intended to measure. In this sense, the listening (MC) component in this test paper is of low validity.

As for reading, with 45% of misfit items in this part (Table 2), its discrimination ability is rather low, so it also fails to represent the differences of students' ability. Moreover, it does not accurately measure what it is designed to measure. According to Yang & Weir (1998), the reading comprehension abilities CET is expected to test should include three levels of processing: syntactical level, discourse level and inference level; and items involved in these three levels should be well-proportioned. A close analysis of the reading items reveals that students can obtain correct answers for 11 items of the total 20 (55%) by only using syntactical processing, contrary to 25% and 35% of the cases in which discourse and inference levels of reading processing are needed (Table 4). This reduces the degree of validity of this part because it does not accurately measure the reading comprehension abilities that are expected of CET.

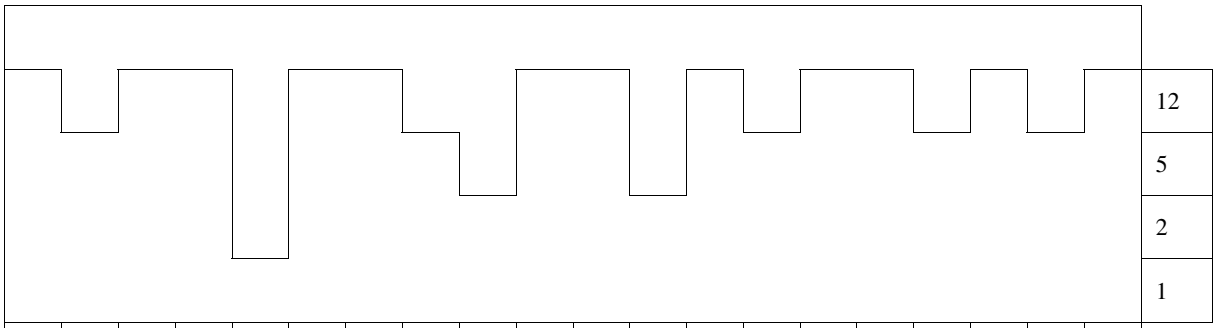
Table 4 Analysis of the Levels of Reading Processing

Level of reading processing	Item No.	%
Syntactical Level	11,12,13,14,16,19,20,21,23,24,25	55
Discourse Level	18, 22,26,27,28	25
Inference Level	15,17,26,27,28,29,30	35

Normally, the correlation between reading and cloze is expected. Many studies of cloze tests (e.g. Bachman, 1981; Hanania & Shikhani, 1986) show that cloze tests can be reliable and valid measures of second language proficiency. In the studies of Streiff (1977, cited in Oller, 1979) and

Hofman (1974, cited in Oller, 1979), cloze tests are even used as measures of reading proficiency. To confirm that it is the problematic reading component that leads to the low intercorrelation between reading and cloze, the validity of the cloze component deserves closer examination. Li (1997) proposes a method of analyzing different levels of test points in a cloze test, in which the levels of test points are identified as word, phrase, sentence and discourse. Accordingly, three categories of test point factors are recognized: grammar, collocation and meaning. According to Li (1997), the higher the level of test points is, the higher the degree of validity the cloze test achieves. Following her method, the cloze subtest is analyzed and, as shown in Figure 1, 12 items (60%) require discourse level of processing, which is in accord with the CET test specifications that the cloze component is aimed to test the candidates' comprehensive language abilities and should include substantial items that involve discourse comprehension (Yang & Weir, 1998). So the validity of the cloze component is of a high degree. Combined with its reliability index (0.73), it can be safely claimed that the reasons of low intercorrelation between cloze and reading do not lie in the cloze's part.

Figure 1 Analysis of Test Point Levels of Cloze



Item	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80
Factor	M	C	M	M	G	M	M	G	C	M	M	C	M	G	M	M	G	M	G	M

D=discourse; S=sentence; P=phrase; W=word; M=meaning; C=collocation; G=grammar

Factor Analysis

The correlation matrix in Table 3 was then subjected to factor analysis. As a result, 2 factors with Eigen values larger than 1 were extracted, contributing to 59.436% of the variance explained (Table 5). The loadings of each subtest on the 2 factors are shown in Table 6, where dictation, vocabulary & structure, cloze and writing are found to contribute to factor 1, and Listening (MC), reading and vocabulary & structure to factor 2.

Table 5 Factors with Eigen values larger than 1

Component	Eigenvalue	% of Variance	Cumulative %
1	2.432	40.529	40.529
2	1.134	18.907	59.436

Table 6 Factor Analysis

	Component	
	1	2
PIA	.257	.672*
PIB	.880*	-.033
PII	-.052	.890*
PIII	.560*	.532*
PIV	.571*	.114
PV	.700*	.229

Extraction Method: Principal Component Analysis.

Rotation Method: Varimax with Kaiser Normalization.

The above analyses of reading and cloze also help to explain the results of factor analyses. As shown in Table 6, the strongest loader on factor 2 is reading (0.890), followed by listening (MC) (0.672) and vocabulary & structure (0.532). The commonalities between listening (MC), reading and vocabulary & structure help to explain that they are testing more or less the same trait. The listening (MC) part is made up of 10 short conversations, in which syntactical level of listening processing is mostly involved to achieve successful understanding. And as discussed above, the reading part fails to include an adequate portion of items that require discourse level of processing, so it is testing more syntactical ability than discourse ability. Finally, vocabulary & structure is designed with an obvious aim to test the usage of words, phrases, collocations and grammatical structures (Yang & Weir, 1998). In this way, factor 2 can be interpreted as a specific factor that accounts for lower level of language ability such as words, grammar and sentence structures.

Meanwhile, the first factor extracted receives its highest loading from dictation (0.880), the second highest from writing (0.700), followed by two modest but still significant ones: cloze (0.571)

and vocabulary & structure (0.560). This factor can be explained as a global one. Of the four strongest loaders on this factor, three (dictation, cloze and writing) are integrative and all invoke discourse processing skills. This is consistent with the results of some studies of second language proficiency that 'the global factor seems to be best measured by tests that are highly integrative in nature---especially discourse oriented tasks' (Oller & Khan, 1981, p14). The factor analyses results confirm the weak global factor hypothesis that 'there exists a general factor accounting for a large portion of the variance in all valid measures of language proficiency' (pp. 5-6) and at the same time, this general factor is complemented by 'various specific factors' (p. 16). Further support comes from Bachman and Palmer's (1981) construct validation study in which one general factor and two specific ones are found. Similarly, in their validation study of CET, Yang & Weir (1998) identify the most heavily loaded factor as 'general language ability'.

According to the quantitative analysis discussed above, the simulated test paper is modestly reliable and valid. But at the same time many problems are exposed. The listening subtest is too easy and not satisfactorily reliable. The reading subtest is extremely unreliable and not highly valid with too few items testing the discourse level of processing and too many misfit items. If pilot tests and post hoc analysis can be conducted, the too-easy questions and the misfit items can be singled out and revised before the simulated test is launched so that its reliability and validity can be guaranteed. And with higher reliability and validity, the test will be better judgment of the candidates' present language levels and better prediction of their performance in the real CET. It will then help teachers to identify students' weaknesses in language learning and take remedial actions.

Unfortunately, these practices are seldom found in the Chinese context. The reasons are two-fold. For one thing, people involved in the construction and administration of an English test do not have adequate knowledge of reliable and valid language testing to take these practices into consideration. For another, launching pilot tests and conducting post hoc analysis consumes a lot of time and energy. So in most cases the only chance of using statistical techniques is to count the pass rate, which is adequate enough to cater for the administrators.

Other aspects of validation analyses

The other aspects of validation analyses in Messick's framework tend to investigate the factors of

different individuals, tasks, settings, and test conditions, as well as the influences of the test as a social act. Though they are not performed in the present study because of contextual difficulties, the possibility and significance of these analyses deserve thoughtful attention and systematical research in further studies of simulated tests.

Analysis of process

In Yang & Weir's (1998) validation study of CET, retrospective verbal reports are utilized to explore the reading strategies employed by candidates in different score bands. Yang & Weir find that contributory reading strategies are most often employed by candidates of higher scores, whereas non-contributory reading strategies are more frequently used by those of lower scores but fail to help in choosing correct answers, implying that contributory reading strategies deserve more training and learning. The time constraints of this study do not allow such an analysis but its significance is obvious. If it can be conducted, different test strategies can be identified and those contributing to better test performance will be encouraged among students. In this way, a simulated test can greatly help to guide the post test language teaching and learning.

Analyses of Group Difference and Change over Time

If a reliable and valid simulated test paper is used by both sophomores (to sit for CET 4 very soon) and freshmen (to sit for CET 4 at least one year later), discrepancies of language abilities can be discovered. Combined with the analysis of test process, a great deal of information will be drawn to guide curriculum planning for the freshmen group. Furthermore, checking the test against some external criteria, concurrent or predictive, can provide further proofs of reliability and validity. In the case of a simulated test, this analysis involves the seeking of other forms of assessment, such as teacher assessment or results of classroom tests. But disappointingly, these forms of assessment seldom exist in China's highly test-oriented context. Teachers here seldom assess their students in other ways than CET-format standardized tests. Furthermore, the confirmation of predictive validity is possible after the results of CET are reported. The correlation study between the simulated test and CET sat by the same group of students will show how well the scores of the simulated test predict students' performance in a real CET. Regrettably this study is not considered in this context either. And again the present study fails to include it due to the time constraints. But the necessity of doing it is suggested here for later researchers.

Manipulation of Tests and Test Conditions

Test preparation practice or coaching is ubiquitous in China. Its emphasis on test familiarization and anxiety reduction may improve validity but the testwiseness strategies that are encouraged in coaching correspondingly lower validity (Messick, 1996). Hamp-Lyones (1998) uses the case of TOFEL preparation as a general example of the problems of practice in this area. But the discussions of coaching for CET are few and far between. Whether test preparation is ethical or not is a big concern and deserves more empirical work.

Test Consequences

Among the issues concerned with test consequences, the washback issue is of special importance to a simulated test. As Messick (1996) points out, less valid tests could precipitate bad educational practices (negative washback) while more valid tests could facilitate good educational practices (positive washback). A simulated test of high reliability and validity will accurately reflect candidates' present level and existing weaknesses and positively result in proper remedial support in teaching. To this end, post hoc analysis of test items and follow-up revision are essential. Meanwhile, as essential is the investigation of teaching/learning context and persons (teachers/students) responses. Although formal survey or interviews are not conducted, inquiries into the reasons why some students (5 out of 43 in this case) give up writing do reveal some problems of de-motivation. The most important reason is they feel it useless writing the composition since they have done badly in the previous parts (their total scores except writing range from 30 to 41). The more simulated tests they take, the more frustrated they feel. If no remedial support is supplied by teachers after the tests, simulated tests are repetitions of frustration and failures for less proficient students. The suggestion of the present study is that the investigation of test consequences should be conducted and remedial actions should be subsequently taken in a context where simulated tests regularly take place.

Conclusion

The validation of this simulated test paper is discussed in Messick's framework from six aspects. Detailed analyses are mainly conducted in the first two aspects. Generally speaking, this simulated test paper is of modest reliability and validity. The most serious problem lies in the reading

component, in which as high as 45% of the items are misfit and deserve revision. Moreover, it fails to test candidates' discourse reading ability effectively. A secondary problem lies in the listening component which is too easy for the intended candidates. These problems decrease the test's reliability and validity and hinder it from best fulfilling its function of predicting candidates' performance in future exams and diagnosing the current problems in language learning. Failure to conduct analyses in the other aspects due to contextual difficulties indicates a very unsatisfactory situation of simulated test practice in the Chinese context: the practices to study the test process, carry out comparative, concurrent or predictive tests, investigate test preparation practice and wash back effects are seldom seen, but they are of great significance for a simulated test to serve as remedial action in language teaching instead of preparation practice that leads to misjudgment and demotivation. For such a large-scale exam as CET, simulation tests costs great time, energy and resources. Its validation deserves thoughtful consideration and research.

The implications of this study are several. First, simulated test papers should undergo careful examination and only those proved to be reliable and valid can be kept for further use. Trial tests and post hoc analyses are essential. Second, empirical investigations into the process of test taking, the effects of coaching and practice and the motivation problems should be advocated. Furthermore, the establishment of various assessments in daily teaching activities is important to triangulate test results. Last but not least, effective remedial support can become one part of normal teaching and lead to positive washback. Only when the validation study is done in this way, can a simulated test be improved and reused, serving its purposes of diagnosis and prediction. Although the findings of this study might not contribute anything new to the language testing theories, by using the simulated test's purposes as a guide to validation, it indicates that if a simulated test is energetically validated, it can serve as a tool to improve language teaching and learning in addition to its function of assessment.

References

Bachman, L. F. (1990). *Fundamental considerations in language testing.* Oxford: Oxford University Press.

Bachman, L. & Palmer, A. (1981). Basic concerns in language test validation. In J. Read (Ed.),

Directions in language testing (pp. 41-71). Singapore: Singapore University Press.

Chapelle, C. A., Jamieson, J. & Hegelheimer, V. (2003). Validation of a web-based ESL test. *Language Testing*, *20*(4), 409-439.

College English Syllabus, Shanghai: Shanghai Foreign Language Education Press, 1985.

Fulcher, G. (1997). An English language placement test: issues in reliability and validity. *Language Testing, 14*(2), 113-138.

Guerrero, M.D. (2000). The unified validity of the four skills exam: applying Messick's framework. *Language Testing*, *17*(4), 397-421.

Hamp-Lyones, L. (1998). Ethical test preparation practice: The case of the TOEFL. *TESOL Quarterly*, *32*(2), 329-337.

Hanania, E. & Shikhani, M. (1986). Interrelationships among three tests of language proficiency: standardized ESL, cloze and writing. *TESOL Quarterly*, *20*(1), 97-109.

Hughes, A. (1989). *Testing for language teachers.* Cambridge: Cambridge University Press.

Li, Xiaoju (1997). *The science and art of language testing.* Hunan: Hunan Education Press.

Luo, L. (2003). *Tsinghua version guidebooks to CET 4: simulated test papers.* Beijing: Tsinghua University Press.

Messick, S. (1996). Validity and washback in language testing. *Language Testing*, *13*(3), 241-256.

McNamara, T. (2001). Language assessment as social practice: challenges for research. *Language Testing*, *18*(4), 333-349.

Oller, J. (1979) *Language tests at school: A pragmatic approach.* London: Longman Ltd.

Oller, J. W. & Khan, F. (1981). Is there a global factor of language proficiency? In J. Read (Ed.), *Directions in language testing* (pp. 3-40). Singapore: Singapore University Press.

Read, J. & Chapelle, C. A. (2001). A framework for second language vocabulary assessment. *Language Testing*, *18*(1), 1-32.

Wa1l, D., Clapham, C. & Alderson, J.C. (1994). Evaluating a placement test. *Language Testing*, *11*(3), 321-344.

Worthen, B. R., Borg, W. R. & White, K. R. (1993). *Measurement & evaluation in the schools*. London: Longman.

Wood, R. (2001). *Assessment and testing: A survey of research.* Beijing: Foreign Language Teaching and Research Press.

Yang, H. & Weir, C. (1998). *Validation study of the national college English test.* Shanghai: Shanghai Foreign Language Education Press.