Zhou, Z. (2016). Speaking section in English speaking and writing test - ESWT [Review]. In J. Levis, H. Le, I. Lucic, E. Simpson, & S. Vo (Eds). *Proceedings of the 7th Pronunciation in Second Language Learning and Teaching Conference*, ISSN 2380-9566, Dallas, TX, October 2015 (pp. 305-309). Ames, IA: Iowa State University.

SOFTWARE REVIEW

Speaking Section in English Speaking and Writing Test (ESWT)

Ziwei Zhou, Iowa State University

INTRODUCTION

The English Speaking and Writing Test (ESWT) is a web-based, free of charge, and low-stake testing program, developed in Pai Chai University in Daejeon, South Korea to provide appropriate assessment of local university students' speaking and writing proficiency (Kim, 2011). Its development was originally motivated by the various advantages computer-assisted language testing (CALT) offers such as assessing large number of test takers and the ability to track students' improvement. Another motivation derived from the observation that the expensive test fees in many popular commercial testing products caused financial burdens for students (Kim, 2011). Therefore, a test that is sensitive to local context is needed to assess students speaking and writing proficiency on the one hand, and provide constructive feedback to track and facilitate students' English language development on the other (Kim, 2011).

SPEAKING TASKS OF THE ESWT

The speaking consists of 4 tasks that prompt test takers to produce extended responses with open-ended questions. In the first task, test-takers are asked to introduce themselves in 45 seconds, with 10 seconds preparation time. When they are ready to respond, they need to hit the RECORD button. When they finish recording, they hit the STOP button. They are asked to speak at least 30 seconds. While they are responding, they can see their recording volume as well as remaining time on the screen. In the second task, test-takers need to narrate a story based on a set of six pictures. They have 30 seconds preparation time and 60 seconds response time. The pictures involve common topics for university students. The third task provides test-takers with visuals such as tables, bar graphs, etc. and requires them to describe the visuals. In the last task, test-takers are prompted to give their opinions on familiar topics that are closely associated with their personal life (e.g. "What is your major? How will you contribute to the society with your major? Why?") or some contentious issues that are common to the them (e.g. political issues between North and South Korea). They have 30 seconds to plan their answers and 60 seconds for answering. They are also required to speak at least 30 seconds.

VALIDATION

The development of the ESWT adopted Chapelle, Jamieson, and Hegelheimer's (2003) suggestion to collect validity argument evidence during the entire process of test development. It integrated Davidson and Lynch's (2002) test-spec approach, ADDIE model (i.e. Analyze, Design, Development, Implementation, and Evaluation), and Bachman and Palmer's (1996) Test Usefulness framework (Kim, 2011).

According to Kim (2011), the constructs of the speaking section, including fluency, functional competence, pronunciation, grammar, vocabulary and expressions, and coherence, were based on ACTFL proficiency guidelines in speaking. Tasks were designed on the basis of the specified constructs in the scoring rubric. Additionally, correspondence in terms of topics, situations, sources, preparation and response time, and answering methods between the test tasks and tasks in the TLU domain were attended to so as to strengthen context validity (Kim 2006b; Kim, 2011). The main test was implemented via multimedia authoring tool and administered by FTP-based management system. Finally, the test was evaluated by statistical analysis of test scores as well as feedback from students.

EVALUATION

Construct Validity

The seminal work of Cronbach and Meehle (1955) defined construct as "some postulated attributes of people assumed to be reflected in test performance" (p. 178). Therefore, the construct model concerns with indirect measures of abilities or attributes of human behavior. In his adoption of the Usefulness Analysis Table (Bachman & Palmer, 1996) to evaluate the EWST, Kim (2011) established construct validity based on task design and difficulty as well as interface design. He pointed out that this quality of usefulness was supported by empirical evidence showing that students perceived test tasks as acceptable and test interface as satisfying. In defining the constructs, Kim (2011) only made brief reference to scale descriptors in ACTFL, TSE, and TWE. Beyond this, no statements were made about test takers' certain attribute as reflected in their test performance. The construct definition in the test development and validation process remained largely absent. The fundamental rationale of the author's position in conceptualizing the constructs was also unstated. This lack of explicit attendance to construct definition may partly explain the lack of reference to any construct theory in Kim (2011).

The sole reference to ACTFL, TSE, and TWE may be problematic because the particular contexts where the ESWT was used may entail different requirements for university students in Korea than the U.S. where these proficiency rubrics and rating scales were developed. Even though the test was used for low-stake purposes, solidifying the theoretical rationales and beliefs by referencing to "theory of the construct" and "construct theories" is also needed since it forms the basis for test specification as well as the hypothesized relations in the nomological network and test constructs and other constructs (Messick, 1989; Chapelle et al., 2003).

Content Validity

According to Luoma (2004), comprehensive content coverage in relation to the definition of test purpose should be the major validity concern of speaking assessment. Such relation can be carefully investigated by delineating task features between the test and non-test situations. In the ESWT, Kim (2011) did not seem to provide sufficient evidence to show that the test tasks were representative of the TLU domain. Though evidence from university teachers' opinions and students' perceptions are crucial, domain description

and modeling has not been reported. Task representativeness should be carefully addressed by systematic and exhaustive attempts to map features between test and non-test situations with reference to refined framework of task characteristics, interactions, responses, and evaluations of tasks. Granted, even though the task feature correspondence is carefully drawn, some task in the TLU domains are impossible to be captured in CALT settings. At the very least, prompts should contain sufficient contextual cues to engage test takers' *discourse domain* (Douglas, 2000) in interacting with the test tasks. Otherwise, there may be no basis to infer that the intended constructs are appropriately and adequately elicited by the test tasks.

Interactiveness and Reliability

Kim (2011) used Bachman and Palmer's Test Usefulness framework to evaluate the development and validation results of EWST (see Figure 1 below).

Quality	Positive Attribute	Negative Attribute
Con-struct Validity	* The test was developed to elicit only variation by the intended English speaking and writing ability using the spec-driven testing framework. * Task difficulty is acceptable with mean and standard deviation scores of speaking (4.00(.74)) and writing (3.89(.72)). * Test takers' perceptions of task difficulty are acceptable with mean and standard deviation scores of speaking (3.71(.69)) and writing (3.82(.74)). * Test takers are satisfied with the test interface design in general.	* Test takers were a little dis-satisfied with the task topics and situations. * Test takers felt that preparation time and writing response time were a little short. * For item #2 and #9, test takers perceived task difficulty contrary to the real test scores.
Relia-bility	* Test topics and situations were designed to elicit only variation by the intended constructs that were extracted based on construct theories. * Task difficulty in terms of test scores and perceptions was acceptable.	* To elicit precise constructs, constructed-response tasks should have been included.
Authen-ticity	* The test topics and situations were chosen from appropriate TLU domain. * Test takers felt that tasks were highly relevant to their real life. * Test takers are highly familiar with the Web-based test tool.	* The TLU domain is too wide. * The tasks were not developed to take care of all the English language proficiency of the test takers. * Three test takers failed to respond to the first task.
Interac-tiveness	None	* There is no interlocutor so that it is impossible to measure the quality.
Impact	* The test offers testing experience to the participants. * The test may achieve positive washback-effects so that they improve their English speaking and writing ability.	* Some students may feel pressure due to the test preparation.
Practi-cality	* The test was developed by the department faculty with low cost and can be administered at general computer labs without extra cost.	* Rating fees and time are inevitable.

Figure 1. The ESWT usefulness analysis table.

In the ESWT Speaking, quality of interaction was impossible to measure due to the lack of interlocutor (Kim, 2011). Without the interactiveness component, it becomes impossible to investigate the extent to which certain elements in the theoretical construct model is activated during test takers' engagement with the test tasks. In addition, the basis of investigating strategic competence through analyzing test taking processes is utterly discarded. Moreover, even though the test is functioning in the local university setting regularly (J, Kim, personal communication, November 12, 2015), no evidence of performance consistencies is documented. Also, no empirical study results in terms of

rating agreement or reliability coefficient between test and re-test are reported. Kim (2011) established his reliability evidence on the basis of construct relevancy and task difficulty. However, since explicit construct definition and model is lacking, the claim that only relevant and intended constructs are elicited as a function of the choice of test topics and manipulations of test situations seems unconvincing.

CONCLUSION

The ESWT is an up-to-date example of CALT product meeting local demands. Its development and evaluation process incorporated multiple frameworks across disciplines including language assessment, curriculum and instruction, business management, and CALT. The test is authentic since test tasks were chosen from appropriate TLU domain (Kim, 2011). Moreover, the ESWT gives due attention to generating positive impact. Test results are used as materials for the weekly coaching sessions, which facilitates students' self-directed learning (J, Kim, personal communication, November 18, 2015). Finally, the ESWT fulfills its original motivation of development by providing free testing experiences and easy access for all users. Recently, the test developer is collaborating with software engineers to upgrade the program system and implement the test on mobile devices (J, Kim, personal communication, January 20, 2016). For future development, EWST should pay closer attention to construct definition and domain description in order to consolidate and explicate the theoretical rationales and beliefs about the speaking abilities in academic contexts. Such efforts will also help to pinpoint task features at a more refined level, which provides the basis to elicit relevant constructs through test takers' own engagement in the test tasks. Future development and evaluation can also investigate test takers' performance consistencies and rater behaviors across time and settings in order to enhance reliability.

REFERENCES

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice.* Oxford: Oxford University Press.

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

Davidson, F., & Lynch, B. (2002). *Testcraft: A teacher's guide to writing and using language test specifications.* New Haven, CT and London: Yale University Press.

Douglas, D. (2000). *Assessing language for specific purposes.* Cambridge, UK: Cambridge University Press.

Kim, J. T. (2006a). *The effectiveness of test-takers' participation in development of an innovative web-based speaking test for international teaching assistant in American colleges.* Unpublished dissertation, University of Illinois, Urbana-Champaign.

Kim, J. T. (2006b). Context validity of the Speaking Proficiency English Assessment Kit (SPEAK). *Korean Journal of the Applied Linguistics, 22*(2), 137-158.

Kim, J. T. (2011). The validation process in developing a web-based English speaking and writing test. *Multimedia-Assisted Language Learning, 14*(2), 181-209.

Luoma, S. (2004). *Assessing speaking*. Cambridge, UK: Cambridge University Press.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13-103). New York: Macmillan.