Speech Recognition Technology in Second Language Assessment: Social Group Preference

Marina Dodigovic

Abstract—**This article follows an attempt to answer the question of how test takers of different social and cultural groups might view speech recognition technology, when applied in second language assessment. The Versant test of English is an application of this technology and is used in a survey-design study to elicit the test taker response. Based on the answers, the author is trying to predict the success of this innovation in light of the diffusion theory. It seems that language teachers of a certain age and residing in South Korea are the most skeptical group in regard to Versant.**

Index Terms—**speech recognition technology, test taker reactions, computer assisted language learning, automated language assessment**

I. Introduction

Is speech recognition technology the kind of innovation which stands a fair chance of being accepted into the modern-day second language assessment? This is a difficult question to answer, since according to Rogers [1], an innovation has to meet several criteria, in order to be "diffused [into the] mainstream" [2, p. 7]. The criteria are not restricted to quantifiable features of the innovation, but are very much dominated by the psychology and mores of the community at hand [2], [3].

This article reviews the controversy of the Versant Test of English in light of Rogers' [1] theory of the diffusion of an innovation. In doing so, it strives to analyse some of the recent criticisms [4], [5] of this speech recognition technology based instrument and assess the likelihood of its diffusion into the mainstream. Mainstream according to Geoghegan [2] refers to three categories of innovation users: 3) early majority (a basically conservative group which accepts an innovation after its value has been established), 4) late majority (a more skeptical group which accepts an innovation when it is quite safe), 5) laggards (a group resistant to change).

What is it that this fully automated test of spoken English does or does not have, and what kind of concerns or paradigms govern its reception by some of the stakeholders? Finally, how likely is it that one day Versant might become widely accepted as valid and reliable evidence of a test taker's English proficiency? These and other questions are pursued in this, both analytical and survey-based study, striving for an increase in understanding of forces at work.

The likelihood of Versant's acceptance into the mainstream is gauged through the test takers' reactions to the characteristics of this test type. Even the test's critics [4], [5] call for the consideration of "the test takers' say in … this" [4]. Consequently, this paper presents a survey in which the test takers' reactions have been solicited immediately following their exposure to Versant.

II. Literature Review

Bachman and Palmer [6] point out that test takers might process the same task in different ways. Thus it can be expected that there would be no uniformity in test takers' perceptions of a test task [7], [8], [4]. Culture and gender of test takers can for instance cause them to experience a variety of emotional states during test-taking [9]. Therefore, it can be expected that different groups or individuals would react quite differently to Versant, some being more accepting of the test than others. This would conform to Rogers' [1] and Geoghegan's [2] findings that values and beliefs of innovation users are crucial to the acceptance of an innovation into the mainstream.

Admittedly, tests can cause "an excessive degree of anxiety [which] can have debilitative effects on the performance of some test-takers" [9, p. 15]. This anxiety is likely to be higher in females [10] and, across genders, in test takers who lack the history of "previous academic success" [9, p. 15]. Thus the level of test takers' comfort or anxiety can be an important predictor of possible test bias toward groups who tend to be more anxious in test situations. Surveying test takers' reactions to a test is a way of gauging the levels of potentially debilitating anxiety.

Connected with the anxiety symptom are other characteristics of candidates and their impact on test performance [11], such as race, social class and background [12], [13]. These and similar issues, investigated often in a test development context [9], can provide insights relevant to construct validity. They also have possible implications for the fairness of the test under development [13], [14]. However, test fairness is not the responsibility of test developers exclusively, but can also be associated with test use [13].

Related to the above is the concept of consequential validity [14], which "focuses on the social consequences of tests" [11]. This kind of investigation is devoted to examining the washback effect of tests on test takers and other stake holders involved [13]. It often resorts to qualitative, predominantly discourse-analysis-based techniques [11]. The washback of high-stakes tests is one of the investigated domains relevant to this study [15], [16]. The use of computer in a test is another dimension of washback, which has received attention, especially as it raises the issue of computer literacy needed to take such a test as an equity issue [17] or considers the effects of

Manuscript received May 17, 2012; revised August 25, 2012.

Marina Dodigovic is with LAU / XJLU (email: mdodigov@gmail.com).

DOI: 10.7763/IJIET.2012.V2.203

interface design [18].

Equity vs. discrimination as inherent to tests and their ethical implications have been much debated [19]. Thus, even test response elicitation techniques can be morally problematic, if they lead to different scores with the same takers [20]. Lynch [19] brings up issues of consent, deception, privacy and confidentiality as crucial to ethical tests. Especially deception is of interest, as it implies, particularly in indirect test measures, such as elicited by Versant, that the taker might be deceived about what is actually being measured. This philosophical question of ethics can also be viewed as relative, i.e. receiving different answers in light of different cultures or even different theoretical approaches [19].

The study at hand explores such issues as the use of computer in a particular language test, equity and fairness of both, computer use and indirect test format, its possible high stakes and the potential washback. It does so by consulting the ultimate stake holder – the test takers themselves. While doing so, it particularly considers the anxiety factor and its relationship to the test taker characteristics.

III. Versant Test of English

Before proceeding to the survey itself, the reader might appreciate a brief summary of Versant's use and function. This fully automated test of primarily spoken English appeared some years ago and was for a while known as the PhonePass or SET 10. Originally, the only way to access it was by phone, which has since been extended to include a much more cost-effective, internet-accessible version. Nowadays, a user anywhere in the world, equipped with an Internet connection, can take this test, either institutionally, using a test pre-paid and administered by a legal entity, such as a school, examination board etc., or can take the test directly from Ordinate, without a mediator. Either way, the test taker is given the test identification code (TIN) to both access the test and subsequently retrieve the test results, which are available in a detailed, user-friendly, competency-based format, only seconds upon test completion.

What a test taker has to do is to respond to six test tasks, mainly by listening and speaking, with the help of on-screen, or in case of phoning in, paper-based instructions. The tasks include: reading random sentences aloud, repeating utterances heard over the speaker, correcting utterances heard over the speaker, speaking short answers to questions into the microphone, retelling short stories heard over the speaker and extended spoken production in answer to opinion questions. Depending on the setup, the test can take up to 30 minutes.

It would be very difficult to cheat on the Versant, as an answer is expected in no less than 20 seconds, thus excluding possibilities of checking external resources. In addition, since each test taker gets a different version of the test, at any one time, it would be extremely difficult for test takers to predict and pre-prepare answers to possible test questions. There are also ways of checking the test taker's identity: either by analysing their recorded voice or by identifying them in person within a test-centre.

This test is being predominantly used for applicant screening in call centre and aviation industries, both of which depend substantially on the language skills of their employees. It has been validated both internally and externally [21]-[24].

IV. Critique of Versant

Comfort and reduced test anxiety are indicated as important positive experiences for test takers [9]. Indeed, anxiety can have an adverse effect on the test takers' performance [9]. With Versant, being able to take the test in the comfort of one's own home, seems like an added benefit, a quality that the test takers would appreciate.

In contrast, a review of Versant [5] brought up concerns, seemingly on behalf of test takers. These claims are that Versant may be difficult to use, that it requires detailed knowledge of operating systems, in addition to having an alienating effect on the test taker, while not being perceived as valid [5]. Although the main concern relates to Versant's authenticity, this aspect has been extensively debated elsewhere [25], [5], and will not be focused on here.

All of the above are legitimate concerns in pursuit of ethical tests. Since Versant is a technology-based test, and technology is known to have introduced both, contaminating variables and anxiety [18] into a testing situation, the issue of user-friendliness of Versant has to be taken particularly seriously. The obvious question to ask would be how representative Chun's [5] claims are of the test takers' perceptions of Versant. This gave rise to the idea to actually identify the key points made by Chun [5] and solicit test takers' judgment on the same. That process is described in the following.

V. Analysis of Critique

The underlying conviction of Chun's [5] review of the Versant test of English is that a test should be authentic. The general concern with authenticity in the language assessment arena, according to Leung and Lewkowicz [8], was spurred by the representatives of the communicative approach to learning and assessment. This "epistemological shift" [26] is known as the sociocultural turn [26], [27] in SLA (Second Language Acquisition), which examines the human cognitive development within the context of social activities it is caused by [28].

This idea is in contrast with the cognitive learning theories, which see learning as an individual and isolated activity [27]. The crux of the debate between the two dialogically engaged approaches has recently been raised around the notion of language learning as separate from that of language use. Thus the strong claim of the sociocultural approach would be that learning, or rather acquisition, equates use, whereas the weak claim would be that some forms of learning are possible apart from use [27].

In language assessment, this dilemma is reflected in the quest for the true nature of language ability [8]. While to the cognitivist, the individual mind is capable of demonstrating competence, the communicationist will question the very existence of such an entity and focus on performance [29], [8]. With the increased interest in the performance per se, authenticity does gain in importance, since it seems crucial to determine how a test taker performs under certain conditions in real life. However, the assessment community does admit that the scores from performance tasks may be less reliable or generalizable than those of e.g. multiple-choice tasks [13], [30], [31].

The Versant Test of English is based on the construct of competence, which in turn is a spin-off from cognitivist approaches to language learning and assessment [21], [22], [25]. This is why it takes a shorter time to complete than most task-based tests, as it is not per se designed to measure performance at a specific task. Given the concerns with the generalizability of performance task scores, it is suggested that rather than having a larger number of raters, one should collect more sizeable samples of task performance for improved reliability [29], [30]. Therefore, high-stakes, task-based English proficiency tests such as TOEFL or IELTS take several hours to complete. In contrast, the Versant, being based on the construct of competence, is short.

It is therefore plausible that Chun [4], [5], coming admittedly from the positions of sociocultural approach to assessment, would be concerned about the authenticity of Versant tasks. Ironically though, the situation in which a test taker takes the test over the phone (or by speaking into the microphone), which to Chun [4], [5] is the ultimate departure from authenticity, is contextually quite appropriate, when the test is used by a prospective flight control or call centre employee. Both of these professions have to be able to pick up the seemingly decontextualised utterances through communication devices, and they do, from time to time, have to read out loud bits and pieces of information to their interlocutors, another task that to Chun [5] appears unmotivated.

VI. Study

A. Study Objectives

In the study at hand, the aim was to examine the acceptability of the Versant to test takers. While sociocultural theory was identified as the theoretical point of departure in the published critique of Versant [5], it did not seem purposeful to try to engage the test taker in the general theoretical discussion of issues in assessment, such as authenticity. Authenticity as a concept in applied linguistics is a value [1], [2] embraced by a specific group of applied linguists. It may not be an issue to test takers at all, especially to those who lack applied linguistic training. While Fulcher [32] argues that the sample of students in his study had a fair understanding of the concept of authenticity or naturalness, in Fulcher's study this concept was introduced in an interview, where the interviewer generally has the chance to explain concepts and check comprehension. The critique of Versant [5] however makes other supporting claims that would be easier for any test taker to understand. Such claims are briefly reviewed here.

Chun [5] thus asserts that Versant "requires a fairly high degree of familiarity with computer technology – that is knowing how to use an operating system on a computer …" [5]. The online instructions are described as "…faceless, unfamiliar and quite possibly unnerving test administrator …" [5]. Chun [5] also doubts that the test taker would be able to adequately interpret his or her score. These characteristics seemed relatively easy to query test takers about.

B. Study Design

Historically, reactions of test takers in language assessment began to be elicited in the 1980's, especially concerning aspects such as construct validity of an instrument and public accountability of test developer/administrator [9]. Closer examination of Chun's work [4], [5] reveals that exactly these two aspects are his principal concern. While Chun's [4], [5] insistence on authenticity translates as concern regarding the construct validity of Versant, his concern regarding the test maker's public accountability is best revealed in the following: "…this test is nothing more than a crass commodity, cynically designed to appeal to the bottom-line concerns…PhonePass objectifies the consumer–[…] the unfortunate test taker subjected to this test" [5].

Test takers are a subgroup of stakeholders in the assessment process [33] that could be particularly affected by the lack of construct validity in a test or the lack of public accountability in the test developer/ administrator. For this very reason, reactions to test format have in fact been investigated before [20], [34], since it holds the key to both validity and ethics, consequently influencing the public accountability of the test maker and potentially causing deception [19].

In this study, an online questionnaire, used because of the ease and speed of administration [34], was intended to confirm or alley the concerns raised by Chun [5]. Since the survey was independent and not to be used in the test development process, information other than that relating to Chun's [5] claims was not of immediate value.

Language testing research according to Huhta et al. [11] has recently shown interest in non-psychometric methods. In line with this, general research literature often suggests combining surveys with interviews [32]; [35]. In this particular case, an unstructured interview with a smaller representative sample of test takers preceding the development of the survey questionnaire would have made sense. However, this was logistically difficult, considering the fact that the population was sampled across countries and professions, with the researcher having immediate access only to a small subgroup of the sample.

C. Study Subjects

Since culture, gender and age can have an influence on the reaction to a test type [34], [9], [33], and can shape the values and beliefs of the test taker as an innovation user [1], it seemed important to query a range of potential test takers in the academia, an area of Versant's application that Chun is particularly concerned about. A total of 127 representatives of test taking population took this test, checked their scores and were asked to complete the questionnaire thereafter. 49 of those were students of non-English-speaking background at English medium universities, while 78 were teachers of English who were not native speakers. Such teachers are frequently required to demonstrate near-native English competence and are therefore as likely a candidate as are university students.

Most of the teachers were located in South Korea (67 out of a total of 68 participants in this country), 6 out of 10 in the

UAE, 3 out of 42 in Qatar, 2 out of 2 in Japan and 2 out of 2 in India. 49 out of the 127 subjects spoke Arabic as their first language, 68 spoke Korean, while Urdu was spoken by 3 and other languages by 7 participants. While 50 were male, 77 were female. 47 of the participants were 41 years of age or older, while the rest were distributed over three categories: 26 – 40 (34 participants), 19 – 25 (21 participants) and 18 years or younger (24 participants).

D. The Questionnaire

Since test taker features may influence the way they perceive a test [9], the first five out of the total of seventeen survey questions were dedicated to demographics, including gender, age, profession, geographic location and the first language of the participants. The format of these questions ranged from binary to multiple-choice, including elements of open-endedness, in particular with the geographic location and the first language of the participant. The remaining twelve five-point Likert scale questions asked for reactions regarding general features of the Versant test, including ease of use, the degree of required computational expertise, the level of test induced nervousness, the absence of a human proctor, the level of anxiety compared to other tests, the emotional response to the computer-generated voice of the test questions, the clarity of online instructions, the strategies used to solve parts of test, the appropriateness of the test given the purpose in question, the accessibility of test scores, and the perceived meaning as well as the accuracy of the test scores. All of these points were formerly addressed in the critique of Versant [5] and relate to important issues of test validity and fairness.

Another option for this survey would have been to use a Likert scale with a different number of points. It is sometimes deemed that a larger number of points can result in a more refined outcome [36]. However, recent research by Dawes [37] suggests that data from 5, 7 and 10 point scales have very similar characteristics. In addition, the choice had to be made between an odd or an even number of points. There seems to be some controversy regarding the issue whether participants tend to choose a neutral solution (neither agree nor disagree) just because it is offered [36]. Adelson and McCoach [36] have demonstrated that an odd point number scale does not contain bias even for primary school students, while Dawes [37] suggests that a 10 point scale does not yield different data from scales with an odd number of points. To simplify matters, the study described in this paper uses a 5- point Likert scale.

The test feature part of the questionnaire is similar to other questionnaires used to elicit test taker reactions toward tests (e.g. [34], [32]). As Brown [34] points out, more reliable information can generally be obtained during the test. However, the Versant does not allow for such interference, leaving only twenty seconds between questions, thus affording insufficient time for the reading and answering of survey questions. While the gender question in the demographics part justifies binary choice, the question about the profession of the participant would not normally call for this format. However, it was clear from the start that test takers would fall into one of two categories, namely English teachers or university students. Hence, the binary choice (i.e. whether English teacher or not) was regarded to be clearer and simpler to answer. While some of the countries as well as languages were anticipated, the participants additionally had the opportunity to add a country or language under "other". This opportunity was however taken up by a relatively small number of participants (4).

E. Results

The majority of the participants found Versant easy to use, without the need for a detailed knowledge of an operating system. Those who were relaxed during the text, led by a small margin over those who did not feel relaxed. A moderate majority of participants preferred being listened to by a machine rather than a human. Compared to other tests, Versant was found to cause slightly less anxiety. There was more certainty about the voice not being irritating. The instructions seemed positively clear. The participants also claimed to be relying more on their grammar knowledge in the sentence mastery part than on guessing. The test was not entirely viewed as appropriate for the purpose for which it was used. Most agreed that the test score was easy to access. Most also found that they understood their test score. Finally, opinions were divided on whether the test score accurately reflected the taker's ability.

TABLE I: CROSS TABULATED RESPONSES (TEACHER/NON-TEACHER, MALE/FEMALE)

Question[1]	Teachers	Non-Teachers		Male	Female
6	3.1	3.88		3.56	3.25
7	3.32	3.54		3.42	3.38
8	2.28	3.34		3.17	2.35
9	2.77	3.34		3.23	2.82
10	2.62	3.34		3.23	2.6
11	2.99	3.4		3.15	3.06
12	3.51	3.4		3.71	3.64
13	3.27	3.6		3.51	3.26
14	2.73	3.29		3.25	2.71
15	3.46	3.85		3.74	3.49
16	3.15	3.67		3.36	3.26
17	2.67	3.34		3.23	3.73

On the average, the test takers did not seem to be overly negative regarding the Versant test, its nature, procedures, purposes or outcomes. However, on questions 8, 9, 10, 11, 14 & 17, there seemed to be a considerable body of "Disagree" and "Strongly disagree statements". In fact, the t-test analysis of cross-tabulated results indicates that there is a significant difference in the entire body of answers between teachers and non-teachers (p = 0.0005), but not between male and female participants (see Table I). In addition, when compared using t-test, age groups tend to fall into two main sections for which there is a significant difference (p = 0.0056): those younger than 25 and those older (Table II). As teased out using t-test, country (Table III) also seemed to play a significant role in differentiating between the results, with the participants in Korea being much more in disagreement with the survey statements than the participants

[1] All questions, including no. 6 – 17, referred to in this table, are listed in the Appendix.

in Qatar (p < 0.0001), UAE (p = 0.0053) or other countries (p < 0.0001).

TABLE II: CROSS TABULATED RESPONSES ACROSS AGE GROUPS

Question[1]	Up to 18	19 - 25	26 - 40	41+
6	3.88	4.05	3.26	2.96
7	3.50	3.71	3.47	3.17
8	3.46	3.45	2.35	2.28
9	3.54	3.05	2.82	2.79
10	3.50	3.10	2.71	2.62
11	3.43	3.50	3.03	2.94
12	3.96	4.11	3.74	3.38
13	3.71	3.70	3.03	3.36
14	3.46	3.10	2.71	2.79
15	3.83	4.05	3.74	3.21
16	3.96	3.43	3.12	3.17
17	3.29	3.48	2.44	2.85

TABLE III: CROSS TABULATED RESPONSES ACROSS COUNTRIES

Question[2]	Qatar	UAE	Korea	Other
6	4.09	2.70	2.99	4.50
7	3.64	3.30	3.24	3.50
8	3.55	3.10	2.04	4.00
9	3.55	2.90	2.60	4.50
10	3.43	3.50	2.41	3.50
11	3.55	3.70	2.73	4.50
12	4.02	4.40	3.34	5.00
13	3.82	3.20	3.09	4.50
14	3.42	3.60	2.46	4.50
15	3.95	4.10	3.25	5.00
16	3.76	3.90	2.96	4.50
17	3.52	2.60	2.51	4.50

F. Discussion

Overall, the average results did not seem to confirm Chun's [5] misgivings about the test takers' reactions to Versant. However, there were noticeable differences in how individuals responded to this test. The results of this survey thus confirm the idea that test takers are not all identical [9]. It seems that age, country and profession have an influence on how the test takers view Versant. However, gender does not appear to be a discriminating factor in this respect.

More specifically, being a language teacher, over 25 and living in Korea seemed to predispose one to being more sceptical regarding the value of this speech recognition based test. This begs the question whether all of the teachers participating in this survey were located in Korea. It is true that the largest number of teachers was located in Korea (67), while a minority (13) were spread across four Asian and Middle Eastern countries. A further 4 participants were non-native speaker English teacher trainees in an MA TESOL program, which is why they declared themselves as students. The attitudes of those 17 participants were overall more positive than the attitudes of the South Korean teachers. This helps us separate the location from the profession to some extent.

Literature on Likert scales reveals that participants do not always select the answer that they would normally consider to be the most appropriate. Instead, especially when they lack motivation or tertiary education, they tend to select the first acceptable answer, depending on the order in which they are presented [38]. Assuming this tendency, the results of this survey seem like an anomaly, for several reasons. The first reason is that the scale starts with the negative, which could have meant that student participants, being younger and less educated might have selected the first acceptable option on the left hand side, which they did not do. However, since most of the student participants spoke Arabic as their first language, in which the directionality of writing is from right to left, this may mean that by them the scales were read starting from the right and settling on the first, although not necessarily the best acceptable answer.

On the other hand, Korean uses multiple scripts, involving vertical as well as horizontal directionality, with the writing in the past evolving from right to left. Thus Korean teachers, especially those over 40 years of age, may have been influenced by the right-to-left directionality as well. Being more educated than the student participants, or simply more patient, due to their maturity, they may have scanned all of the options before settling on one. However, one should allow for the possibility that the teachers in South Korea were simply less motivated to participate in this survey. Since the author of this article was not present when the purpose of both the test and the survey administration was explained to these participants, there is no evidence that they perceived the study as relevant to them.

All of the teachers in South Korea were nationals of that country, mostly South Korean educated as well. Research has previously shown an increased negativity of Korean takers toward ESL tests [16]. These participants seemed much more nervous about taking the Versant test and much less confident in the abilities of the machine to assess their speech than a human rater, compared to any other group. This is consistent with Fulcher's [32] finding that the presence of technology can be anxiety inducing in a speaking test situation. Furthermore, they doubted that the Versant was appropriate for the purpose for which it was used, were not sure how to interpret its results and were less sure of its accuracy in respect of measuring their language ability. All of these responses may point toward the fact that these test takers found the test deceptive [19].

Moreover, the phonics of the Korean language are contrastively very different from English, thus making any speech assessment a potential high-stakes challenge for a native Korean speaker. Consequently, a low speaking test score might lead to a considerable loss of face. Face [39] is a category of great importance to East Asian cultures, signifying the esteem of the community toward an individual. Scoring low on a speech test, because the machine, which is perhaps poorly programmed and possibly not as considerate

[1] All questions, including no. 6 – 17, referred to in this table, are listed in the Appendix.

[2] All questions, including no. 6 – 17, referred to in this table, are listed in the Appendix.

as a human rater, would constitute a logical reason for having negative sentiments toward Versant.

In contrast, the reason for the acceptance of the Versant by the test takers in the Arabian Gulf might be the oral traditions of the Middle East, as well as the emphasis on English as a medium of everyday communication in the Gulf countries [40]. Both of these advantages might have desensitised the issue of speaking for a test in the Arabian Gulf context. Rather than rejecting the machine as a potentially inconsiderate rater, there might have been a tacit acceptance of it as an objective rater. Thus, culture seems to be an important catalyst in test takers' perceptions of Versant.

A concern here might be perhaps the overrepresentation of the teachers in the sample. In response to such concern, one might point out that teachers are a significant stakeholder group, whose acceptance of any test is crucial, since they might be the ones who recommend tests to school administrations and regional authorities. They also write test reviews, thus influencing a much wider public than their local educational systems. For this reason, it was very important to find out how they relate to Versant. Overall, it can be said that language teachers seem less convinced of the immediate value of Versant than language students.

G. Limitations

This study is not without its limitations. One of them is not pairing the survey up with interview or introspection data. Another one is the considerable overlap of categories teacher with the Korean location and language as well as with the age group of over 25 in the sample. This happened because the sample was not randomised, since randomisation was not practicable, as it was not easy to find volunteers for this study. Language students and teachers had genuine motivation to participate in this study and were only accessible to the researcher in a limited number of locations. Also, two questions contained the negative particle, a detail which might have been confusing or introducing bias. However, this was necessary in order to make all the statements affirmative, so that mutually comparable answers could be elicited. These two sentences, being based on Chun's [5] original statements, contained adjectives, such as "irritating ", which gave them a negative meaning. Attempting to choose an appropriate antonym for such expressions would have had the ramifications equivalent to those of semantic differential scales, discussed in a previous section of this article.

VII. Conclusion

The paper started with the intention of determining a trend in the attitudes of users toward speech recognition technology in second language assessment, as exemplified in Ordinate's Versant Test of English. Based on Rogers' [1] insights about the acceptance of an innovation, this study has sought to explore to what extent such a test may or may not be on its way to being accepted into the mainstream. Two major stakeholder groups were surveyed in this effort, university student test takers and language teachers, with significantly differing results. While the students seemed more ready to accept Versant, teachers, especially those over 25, living in Korea, seemed to resist this innovation.

While one might think that test takers would be important stakeholders, experience shows that teachers, and the organisations they are affiliated with, are the actual test users [13]. Whereas the majority of surveyed teachers lived in Korea, thus causing the results to be less generaliseable, it was clear that the teachers' attitudes differed from those of students at a statistically significant level. Therefore, it would seem that speech recognition technology, especially in combination with cognitivist approach to language testing, may not have yet reached the point of entering the mainstream.

I. References

[1] E. Rogers, *Diffusion of Innovations*. London: Macmillan, 1983.

[2] W. Geoghegan, "Instructional technology and the mainstream: The risk of success," in *The Future compatible campus*, D. Oblinger and C. Rush, Ed. Bolton: Anker Publishing, 1998.

[3] M. Dodigovic, "Speech recognition technology in language testing," *6th CamTESOL Conference on 'English Language Teaching One World: World Englishes,' Phnom Penh, Cambodia,* February, 2010.

[4] C. W. Chun, "Comments on 'Evaluation of the usefulness of the Versant for English test: A response': The author responds," *Language Assessment Quarterly*, vol. 5, no. 2, pp. 168–172, 2008.

[5] C. W. Chun, "Commentary: An analysis of a language test for employment: The authenticity of the PhonePass test," *Language Assessment Quarterly*, vol. 3, no. 3, pp. 295–306, 2006.

[6] L. F. Bachman and A. S. Palmer, *Language Testing in Practic*e, Oxford: Oxford University Press, 1996.

[7] L. F. Bachman, "Some reflections on task-based language performance assessment," *Language Testing*, vol. 19, no. 4, pp. 453–476, 2002.

[8] C. Leung and J. Lewkowicz, "Expanding horizons and unresolved conundrums: language testing and assessment," *TESOL Quarterly*, vol. 40, no. 1, pp. 211 – 234, 2006.

[9] J. Bradshaw, "Test takers' reactions to a placement test," *Language testing*, vol. 13, no. 7, pp. 13–30, 1990.

[10] H. D. Madsen, "Determining the debilitative effects of test anxiety," *Language learning*, pp. 32, no. 1, pp. 133–143, pp. 1982.

[11] A. Huhta, P. Kalaja and A. Pitkanen-Huhta, "Discursive construction of a high-stakes test: the many faces of a test-taker." *Language testing*, vol. 23, no. 3, pp. 326–350, 2006.

[12] A. J. Kunnan, "Modelling relationships among some test-taker characteristcs and performance on EFL tests: an approach to construct validation," *Language testing*, vol. 11, no. 3, pp. 225–250, 1994.

[13] T.McNamara and C. Roever, *Language testing: The social dimension*. Blackwell: Malden, MA, 2006.

[14] S. Messick, "Validity and washback in language testing," *Language testing*, vol. 13, no. 3, pp. 241 – 256, 1996.

[15] J. C. Alderson and L. Hamp-Lyons, "TOEFL Preparation Courses: a study of washback," *Language testing*, vol. 13, no. 3, pp. 280–297, 1996.

[16] I. C. Choi, "The impact of EFL testing on EFL education in Korea," *Language testing*, vol. 25, no. 1, pp. 39–62, 2008.

[17] K.van den Branden, V. Depauw, and S. Gysen, "A computerized task-based test of second language Dutch for vocational training purposes," *Language testing*, vol. 19, no 4, pp. 438–452, 2002.

[18] G. Fulcher, "Interface design in computer-based language testing," *Language testing*, vol. 20, no. 4, pp. 384–409, 2003.

[19] B. K. Lynch, "In search of the ethical test," *Language testing*, vol. 14, no. 3, pp. 315–327, 1997.

[20] E.Shohamy, "Testing methods, testing consequences: are they ethical, are they fair?," *Language testing*, vol. 14, no. 3, pp. 340–349, 1997.

[21] Ordinate, *SET-10: Test description – Validation summary*, enlo Park: Harcourt, 2005.

[22] Ordinate, *Versant for English – Technical manual*, Menlo Park: Harcourt, 2007.

[23] Ordinate, *A comparison of three internationally recognized tests of spoken English*, Palo Alto: Author, 2008.

[24] M. Dodigovic, *Artificial Intelligence in Second Language Learning: Raising Error Awareness*, Clevedon: Multilingual Matters, 2005.

[25] R. Downey, H. Farhady, R. Present-Thomas, M. Suzuki and A.Van Moere, "Evaluation of the usefulness of the Versant for English test: A response," *Language Assessment Quarterly*, vol. 5, no. 2, pp. 160–167, 2008.

[26] K. E. Johnson, "The sociocultural turn and its challenges for second language teacher education," *TESOL Quarterly*, vol. 40, no. 1, 235–257, pp. 2006.

[27] J.Zuengler and E. R. Miller, "Cognitive and sociocultural perspectives: two parallel SLA worlds?," *TESOL Quarterly*, vol. 40, no. 1, pp. 35–58, 2006.

[28] L. S. Vygotsky, *Educational Psychology*, Boca Raton: St. Lucie Press, 1997.

[29] A. S. Canagarajha, "TESOL at 40: what are the issues?," *TESOL Quarterly*, vol. 40, no. 1, pp. 9–34, 2006.

[30] M. K. Enright, "Research issues in high-stakes communicative language testing: reflections on TOEFL's new directions," *TESOL Quarterly*, vol. 38, no. 1, pp. 147–151, 2004.

[31] A. Davies, "Three heresies of language testing research," *Language Testing*, vol. 20, no. 4, pp. 355–368, 2003.

[32] G. Fulcher, "Testing tasks: issues in task design and the group oral," *Language testing*, vol. 13, no. 1, pp. 23–51, 1996.

[33] P. Rea-Dickins, "So why do we need a relationship with stakeholders in language testing? A view from the UK," *Language testing*, vol. 14, no. 3, pp. 304–314, 1997.

[34] A. Brown, "The role of test-taker feedback in the test development process: test takers' reaction to a tape-mediated test of proficiency in spoken Japanese," *Language testing*, vol. 10, no. 3, pp. 277–301, 1993.

[35] J. McDonough, and S. McDonough, *Research Methods for English Language Teachers*, London: Arnold. 1997.

[36] J. L Adelson and D. B. McCoach, "Measuring the mathematical attitudes of elementary students: the effects of a 4-point or 5-point Likert-type scale," *Educational and psychological measurement*, vol. 70, no 5, pp. 796–807, 2010.

[37] J. Dawes, "Do data characteristics change according to the number of scale points used? An experiment using 5-point, 7-point and 10-point scales," *International Journal of Market Research*, vol. 50, no. 1, pp. 61–77, 2008.

[38] L. J. Weng and C. P. Cheng, "Effects of response order on Likert-type scales," *Educational and psychological measurement*, vol. 60, no. 6, pp. 908–924, 2000.

[39] R. Scollon and S. W. Scollon, *Intercultural Communication: A Discourse Approach*, Oxford: Blackwell, 1995.

[40] A. S. M. Al-Issa, "The implication of implementing a 'flexible' syllabus for ESL policy in the Sultanate of Oman," *RELC Journal*, vol. 38, no. 1, pp. 199–215, 2007.

Marina Dodigovic has a PhD in Linguistics, with a thesis in Computer Assisted Language Learning from the University of Bremen, in addition to an MA in English and a BA in Language Education. Her teaching career at tertiary level spans almost two decades and includes English as a second language, writing, linguistics, literature as well as English teacher training.

She is an associate professor at Xi'an Jiaotong-Liverpool University, who has taught English and trained ESL teachers in Europe, Australia and the Middle East. In addition to directing ESL and writing programs, she has conducted research in applied linguistics and second language acquisition. Her book, entitled *Artificial Intelligence in Second Language Learning*, focuses on educational technology. She has published refereed research articles in journals such as *CALL* and *Language Awareness*, to name just a few.

A winner of several international scholar awards and a number of research grants, Dr Dodigovic has a keen interest in second language acquisition. She is a member of TESOL, TESOL Arabia and Association for Language Awareness

,