Language Testing in Asia

RESEARCH Open Access

Test-taker perception of and test performance on computer-delivered speaking tests: the mediational role of test-taking motivation

Yujia Zhou* and Asako Yoshitomi

* Correspondence: zhou_yujia@tufs.ac.jp
Tokyo University of Foreign Studies, 3-11-1 Asahi-cho, Fuchu-shi, Tokyo 183-8534, Japan

Abstract

Background: Research on the test-taker perception of assessments has been conducted under the assumption that negative test-taker perception may influence test performance by decreasing test-taking motivation. This assumption, however, has not been verified in the field of language testing. Building on expectancy-value theory, this study explored the relationships between test-taker perception, test-taking motivation, and test performance in the context of a computer-delivered speaking test.

Methods: Sixty-four Japanese university students took the TOEIC Speaking test and completed a questionnaire that included statements about their test perception, test-taking motivation, and self-perceived test performance. Five students participated in follow-up interviews.

Results: Questionnaire results showed that students regarded the TOEIC Speaking test positively in terms of test validity but showed reservations about computer delivery, and that they felt sufficiently motivated during the test. Interview results revealed various reasons for their reservations about computer delivery and factors that distracted them during the test. According to correlation analysis, the effects of test-taker perception and test-taking motivation seemed to be minimal on test performance, and participants' perception of computer delivery was directly related to test-taking effort, but their perception of test validity seemed to be related to test-taking effort only indirectly through the mediation of perceived test importance.

Conclusion: Our findings not only provide empirical evidence for the relationship between test-taker perception and test performance but also highlight the importance of considering test-taker reactions in developing tests.

Keywords: Test-taker perception, Test-taking motivation, Computer-delivered speaking test, TOEIC Speaking test, Expectancy-value theory

Introduction

In the past two decades, emphasis on fostering communicative ability has called for more efficient testing of speaking skill, which has led to an increase in the use of computers to deliver speaking tests. However, notwithstanding the advantages of computer-delivered tests (e.g., reduced costs, increased standardization, and immediate feedback to test-takers), they pose potential threats to test validity, particularly regarding what test-takers think of talking to a computer during the test. Since oral communication usually involves interaction between people, computer-delivered speaking tests seemingly lack validity in terms of authenticity. If this causes negative perception among test-takers, the face validity of the tests would be threatened. Moreover, their motivation to perform well on test tasks may decrease, which in turn, may lead to poorer test performance (e.g., Alderson, Clapham, & Wall, 1995; Bachman & Palmer, 2010). As such, if test-taker perception and test-taking motivation are systematically related to test performance, test scores could be biased estimates of test-taker proficiency, which would confound the interpretation of the test scores (Kane, 2006; Messick, 1989).

Despite the importance of these issues, little is known about the relationships between test-taker perception, test-taking motivation, and test performance on computer-delivered speaking tests. To narrow this gap, our study set out to provide empirical evidence that helps address the concerns over score interpretation of such tests in the context of the Test of English for International Communication (TOEIC) Speaking test, with a focus on Japanese learners of English as a foreign language.

Our study is contextualized in Japan, where the reform of Japanese university entrance examinations has necessitated more efficient testing of students' speaking skill. The National Center Test for University Admissions, one of the major domestically available tests used for selecting university applicants, has only assessed English reading and listening skills thus far. To address the imbalance of the English skills assessed, the Japanese Ministry of Education, Culture, Sports, Science, and Technology has announced a new policy that encourages universities to use externally produced, four-skill English tests for university entrance purposes from 2021 (Ministry of Education, Culture, Sports, Science, & Technology, 2017). A total of eight local and international English tests has been approved by the National Center of University Entrance Examinations (National Center of University Entrance Examinations, 2018), and five of them include a computer-delivered speaking component.

Among these, one test that is designed to measure the ability to communicate in spoken English in the context of daily life as well as the global workplace is the TOEIC Speaking test (Educational Testing Service, 2016). The test was introduced in 2006 by Educational Testing Service as part of TOEIC Speaking and Writing (TOEIC SW) test to address test users' concerns that some test-takers lacked English speaking and writing skills despite their high scores in the TOEIC Listening and Reading (TOEIC LR) test (Educational Testing Service 2019). Since the inception of the TOEIC SW, the number of people taking the test has radically increased from 1200 in 2006 to 38,000 in 2017, and 380 corporations have used the test in Japan (Institude for International Business Communication, 2018).

Under this circumstance, some Japanese universities have started to offer students opportunities to take the TOEIC Speaking test, aiming to raise their motivation to improve speaking skill and help them become familiar with the test. However, although it is true that many companies use TOEIC scores as one of the criteria for selecting prospective employees, these are usually limited to TOEIC LR scores. As a result, the TOEIC Speaking test has not yet gained prominent recognition among students. It can therefore be considered a low-stakes test at the present stage for college students, which gives rise to concerns over a lower level of motivation when they take the test. Moreover, studies to date have examined various aspects of the validity of the TOEIC

Speaking test (e.g., Liao & Qu, 2010), but to our knowledge, no information regarding test-taker perception of this test is yet available.

Test-taker perception of computer-delivered speaking tests and test performance

Recently, test-taker perception has drawn increasing attention in the field of language testing. Although no agreement has been reached regarding terminology (Murray, Riazi, & Cross, 2012), test-taker perception has been widely recognized as an important aspect of test attitude and been studied from various perspectives, including test validity (e.g., Zhou, 2012), test demand (Sato & Ikeda, 2015; Xie, 2011), and test difficulty (e.g., Elder, Iwashita, & McNamara, 2002).

Research on test-taker perception of speaking tests has focused on the measurement and comparison of perception of test validity between different test types. Overall, results have revealed mixed views of technology-based speaking tests. Test-takers generally react positively to the construct, content, and predictive validity of technology-based speaking tests (Fan, 2014; Kiddle & Kormos, 2011; Qian, 2009; Zhou, 2012). In contrast, they seem to be more conservative toward computer-delivered speaking tests, compared to computer tests of other skills (Fan & Ji, 2014; Stricker & Attali, 2010) and face-to-face tests (Brooks & Swain, 2015; Fan, 2014; Kiddle & Kormos, 2011; Qian, 2009), citing lack of interaction in the computer-delivered test as the main reason for their negative perception.

Being a potential source of negative test-taker perception, what test-takers think of using a computer to deliver a speaking test can negatively influence speaking test performance. For instance, test-takers in Brooks and Swain (2015) commented on how lack of interaction (particularly lack of feedback) and time constraints in the Test of English as a Foreign Language Internet-based Test (TOEFL iBT) affected their performance. Despite the evidence from interview studies, few studies have used questionnaires to explore the relationship between test-taker perception and speaking test performance, and the results of previous studies did not confirm the relationship between the two variables (Fan & Ji, 2014; Stricker & Attali, 2010). These results, however, should be interpreted with caution, in that Stricker and Attali (2010) elicited test-taker perception of the TOEFL iBT as a whole test, rather than specifically of the speaking section, and Fan and Ji (2014) used the total score of the Fudan English Test, which was a four-skills test. Therefore, there is a clear need to continue this line of research to gain a better understanding of how the lack of interaction may affect performance on computer-delivered speaking tests.

Test-taking motivation and test performance

Test-taking motivation is closely related to test-taker perception but is underexplored in the field of language testing. Broadly, motivation refers to "the process whereby goal-directed activity is instigated and sustained" (Pintrich & Schunk, 2002, p. 5), whereas test-taking motivation is task-specific and defined as "the willingness to engage in working on test items and to invest effort and persistence in this undertaking" (Baumert & Demmrich, 2001, p. 441). Language testing research has been primarily concerned with goal orientation (e.g., Cheng et al., 2014; Fan, 2014; Xie, 2011), possibly because these studies have examined high-stakes tests, where the level of test-taking motivation was not a particular concern.

In contrast, understanding test-taking motivation is important for low-stakes tests whose results have few, if any, consequences for test-takers. The nonconsequential nature of low-stakes tests can undermine test-taking motivation, artificially deflating performance and thus jeopardizing the validity of test-based inferences (e.g., Cole & Osterlind, 2008). It is thus essential to report test-taking motivation to determine the extent to which end-users can make valid inferences and sound decisions from the results of low-stakes tests (American Educational Research Association, American Psychological Association, & National Council for Measurement in Education, 2004).

One theory of motivation that has been applied to low-stakes testing situations is expectancy-value theory[1] (Atkinson, 1957; Wigfield & Eccles, 2000). According to the theory, test-taking motivation in low-stakes testing is multidimensional, comprising two interrelated components: perceived test importance and test-taking effort (e.g., Wise & DeMars, 2005), and the extent to which test-takers perceive the test as important influences the effort they put into it (e.g., Pintrich, 1999). These proposals have been supported by empirical evidence from many other fields (e.g., Zilberberg, Finneya, Marsha, & Anderson, 2014), and test-taking effort has consistently been found to be a predictor of test performance (e.g., Zilberberg et al., 2014).

Despite theoretical and empirical support for the relationship between test-taking motivation and test performance, to our knowledge, only one study (Zhao & Cheng, 2010) has applied the expectancy-value theory to exploring test-taking motivation during language tests. Although the study found the test-taking motivation to be a significant predictor of language test performance, without differentiating perceived test importance and test-taking effort, it remains unclear how these two components of test-taking motivation are interrelated in affecting language test performance.

Mediational role of test-taking motivation between test-taker perception and test performance

Whether test-taking motivation plays a mediational role between test-taker perception and language test performance is an important, yet unresolved issue. Test-taker perception research in language testing has been conducted under the assumption that, if test-takers hold negative perception, they may expend less effort and their scores may not accurately reflect their ability (Alderson et al., 1995; Bachman & Palmer, 2010). However, to date, no empirical evidence exists to support this assumption, as previous research has only separately explored test-taker perception (Stricker & Attali, 2010; Xie, 2011; Zhou, 2012) and test-taking motivation (e.g., Cheng et al., 2014).

More recently, researchers in other fields have sought to determine how test-taking motivation mediates test-taker perception and test performance by considering the multidimensional nature of test-taking motivation (e.g., Zilberberg et al., 2014). It is proposed that negative test-taker perception may not only directly lead to decreasing test-taking effort but also impact perceived test importance, which in turn negatively affects test-taking effort, resulting in suboptimal performance (e.g., Swing, 2001). However, whether these proposals apply to a language-testing situation remains uninvestigated.

With the above as the background, this study aimed to provide empirical evidence regarding test-taker perception of and test-taking motivation during the TOEIC Speaking test, as well as the mediational role of test-taking motivation between test-taker perception and test performance. To explore the relationships between the examined variables, we controlled for the effects of two potential moderator variables. One was test-takers' general English ability, given that test-takers may do better on an English test not because of their favorable reactions but because of their higher English ability (Xie, 2011). The other variable was test-takers' self-perceived test performance, which may influence their test reactions because of self-serving bias.[2] Specifically, our study addressed the following research questions:

RQ1: How do Japanese university students perceive the TOEIC Speaking test in terms of test validity and computer delivery? What are the possible reasons underlying their perceptions?

RQ2: To what extent do Japanese university students report feeling motivated during the TOEIC Speaking test? What are the possible factors underlying their test-taking motivation?

RQ3: To what extent are test-taker perception, test-taking motivation, and performance on the TOEIC Speaking test related?

Methods

Participants

Sixty-four undergraduate students at a key national Japanese university in Tokyo (one of the top universities with foreign language majors) were recruited through posters placed around campus. Participants were a convenience sample, consisting of 20.3% men ($n = 13$) and 79.7% women ($n = 51$), with ages ranging from 18 to 22 years ($M = 22.1$, $SD = 5.58$). There were 15 freshmen (23.4%), 16 sophomores (25.0%), five juniors (7.8%), and 28 seniors (43.8%). They were from 20 language majors (including English), but the largest groups studied French (12.5%), German (12.5%), Arabic (10.9%), and Russian (7.8%). Among the participants, five students (four females and one male) were randomly recruited via email for follow-up interviews and were paid 1000 yen for participating.

All participants had received at least seven years of high school and university-level English instruction at the time of the study, and the majority of the participants were required to take two English courses designed to improve their receptive and productive skills. Based on their TOEIC speaking test scores ($M = 134.6$; $SD = 19.0$), the speaking proficiency levels of most participants were B1 (72.2%) and A2 (16.7%), according to the Common European Framework of Reference for Languages (Tannenbaum & Wylie, 2008). Their TOEIC LR test scores ($M = 767.2$; $SD = 115.1$) showed that their receptive skills mainly belonged to B1 (51.8%) and B2 (37.0%) levels.

Instruments

TOEIC Speaking test

The TOEIC Speaking Institutional Program (IP) test was used in this study. The test is computer-delivered and consists of 11 tasks categorized into six types. Test-takers (1) read aloud short passages, (2) describe a photograph, (3) respond to three questions based on personal experience, (4) answer three questions based on a written schedule of events, (5) listen to a telephone message and propose a solution that addresses the question raised in the message, and (6) express their opinion on a topic. The test takes approximately 20 minutes. Responses to questions require test-takers to respond immediately; for the other tasks, they are given 15–45 seconds to prepare a response. Response time allowed for each task ranges from 15 to 60 seconds. Test responses on different tasks are scored by certified raters on a scale of 0–3 or 0–5 with different combinations of rating categories (e.g., pronunciation, intonation and stress, grammar, vocabulary, cohesion, the relevance of content, and completeness of content). The sum of the ratings is converted to a score of 0–200, which was used to indicate participants" speaking test performance in this study.

TOEIC LR test

The TOEIC LR test is a paper-and-pencil, multiple-choice assessment. The test has two sections (listening and reading) with 100 questions for each. It takes two hours in total (45 minutes for the listening section and 75 minutes for the reading section). This study used the TOEIC LR IP test, which provides two section scores for listening and reading (ranging from 0 to 495) and a total score (ranging from 0 to 990). In this study, the total score was used as the indicator of participants' general English ability.

Questionnaire

The questionnaire consisted of 19 items about test-taker perception of the TOEIC Speaking test, test-taking motivation, self-perceived performance on the test, and demographic information (including age, gender, academic year, and English test experiences). The questionnaire was designed in English and translated into Japanese.

Test-taker perception Test-taker perception was assessed with seven items on various aspects of test validity of the TOEIC Speaking test. In order to explore specifically the potential effect of test-taker perception of computer delivery, the test-taker perception scale was divided into two subscales of perceived test validity and perceived computer delivery. The subscale of perceived test validity was adapted from Fan (2014) and Stricker and Attali (2010). It had three items about perceptions of the test's construct validity (i.e., "The abilities measured in the TOEIC Speaking test were essential to oral communication"), content validity (i.e., "The TOEIC Speaking test is a good way to learn how well people can speak English in daily life"), and predictive validity (i.e., "People who receive high scores on the TOEIC Speaking test have a high speaking ability").

The subscale of perceived computer delivery was designed based on Stricker and Attali (2010) and Zhou (2012). It included two items about test fairness related to computer delivery (i.e., "It is all right to give the TOEIC Speaking test to people without much computer experience" and "It is all right to give the TOEIC Speaking test to people who do not like computers"), one item about perceived authenticity (i.e., "It is not unnatural to talk to a computer during the speaking test"), and one item about personal experience (i.e., "Taking the speaking test on a computer was a pleasant experience"). All items were rated by participants on a five-point scale ranging from 1 (*strongly disagree*) to 5 (*strongly agree*). The reliability estimates (Cronbach's alpha) of perceived test validity and perceived computer delivery subscales were .75 and .61, respectively.

Test-taking motivation Test-taking motivation was measured with eight items adopted from the Student Opinion Scale[3] (Sundre & Moore, 2002), which consisted of two subscales: perceived test importance and test-taking effort. To fit the TOEIC Speaking test, item wording in the original scale was modified slightly with "the test" replaced with "the TOEIC Speaking test." The perceived test importance subscale included four items that measured the degree to which test-takers viewed the TOEIC Speaking test as important (e.g., "Doing well on the TOEIC Speaking test was important to me"). The subscale for test-taking effort had four items that assessed the degree to which test-takers put forth effort on the test (e.g., "I engaged in good effort through the TOEIC Speaking test"). For a complete list of the specific items refer to Table 2. Test-takers responded to items using a five-point scale (1 = *strongly disagree* and 5 = *strongly agree*). The reliability estimates (Cronbach's alpha) of perceived test importance and test-taking effort subscales were .70 and .83, respectively.

Self-perceived test performance Four items adopted from the literature (Sanchez, Truxillo, & Bauer, 2000; Stricker & Attali, 2010; Zhou, 2012) were used to measure test-taker perception of their performance on the TOEIC Speaking test (e.g., "I believe that I did well on the TOEIC Speaking test"). Test-takers evaluated their performance on a five-point scale ranging from 1 (*strongly disagree*) to 5 (*strongly agree*). The reliability estimate (Cronbach's alpha) of the scale was .85.

Data collection procedures

Participants took the TOEIC Speaking IP test in February 2015 on campus and responded to the questionnaire immediately after completing the test.[4] Written directions were provided to assure them that their responses to the questionnaire would be confidential and would not affect their academic grades. A week later, they took the TOEIC LR IP test, as required by the university. Official scores of both tests were obtained directly from the university.

Semi-structured interviews were conducted in Japanese with each of the five participants in November 2015.[5] To remind interviewees of test content, a sample test accessible online was shown at the beginning of the interview. Interview questions were mainly related to questionnaire items about how they perceived the TOEIC Speaking test, how important they considered the test, and their degree of test-taking effort. The interviewees were also asked to describe their motivation to take the test and the influence the test-taking experience had on their English learning. All interviews, which lasted 40 min on average, were recorded.

Data analysis

Prior to data analysis, all questionnaire items were inspected for data entry accuracy and missing values. Cases were deleted if more than 10% of the items were not completed, which resulted in ten cases being removed; the remaining 54 cases had no missing data and were used for subsequent analysis.

To examine test-taker perception (RQ1) and test-taking motivation (RQ2), descriptive statistics at the item and subscale levels were computed. Given that the mean score did not give a clear indication of the distribution of responses, frequency statistics for each item were also calculated by merging "strongly agree" and "agree" categories into one overall "agree" category and by combining "strongly disagree" and "disagree" categories into a single "disagree" category. Chi-square tests were run to determine whether response frequency was statistically different between categories.

To explore relationships between test-taker perception, test-taking motivation, and speaking test performance (RQ3), descriptive statistics and Spearman rank-order correlations between all variables were calculated. If the TOEIC LR score was significantly correlated to the TOEIC speaking score, semi-partial correlations were computed to control for the effect of the TOEIC LR score. Semi-partial correlations were also calculated to remove the effect of self-perceived test performance on variables of test reactions where significant correlations were found. All analyses were performed using Statistical Package for Social Sciences 22.0. All significance tests were two-tailed, and the alpha level was set to .05, if not otherwise reported.

Results

In this section, findings are reported along with specific excerpts from the interview.

Test-taker perception of the TOEIC Speaking test

Table 1 presents descriptive and frequency statistics for questionnaire items on test-taker perceptions of test validity and computer delivery of the TOEIC Speaking test.

The mean score at the subscale level (*M* = 3.85 on a five-point scale) showed that participants held a moderately positive attitude toward the validity of the TOEIC Speaking test. This trend was confirmed by Chi-square tests, which indicated that significantly more participants chose "agree" over "neutral" or "disagree" for each item. However, participants showed slightly different degrees of agreement with different aspects of test validity. The mean scores on construct validity and predictive validity were slightly higher than that on content validity. Indeed, an overwhelming majority of participants (above 80%) tended to agree that abilities measured by the test were essential to oral communication and that

Table 1 Descriptive and frequency statistics for questionnaire items on test-taker perception

			Frequency (%)		
Item	*M*	*SD*	Disagree	Neutral	Agree
Perceived test validity					
1. The abilities measured in the TOEIC Speaking test were essential to oral communication.[a]	3.91	0.65	2 (3.7)	8 (4.8)	44 (81.5)
2. People who receive high scores on the TOEIC Speaking test have a high speaking ability.[a]	4.07	0.80	3 (5.6)	6 (11.1)	45 (83.3)
3. The TOEIC Speaking test is a good way to learn how well people can speak English in daily life.[a]	3.57	0.92	7 (13.0)	15 (27.8)	32 (59.3)
Total	3.85	0.64			
Perceived computer delivery					
1. It is all right to give the TOEIC Speaking test to people without much computer experience.	2.91	1.17	23 (42.6)	11 (20.4)	20 (37.0)
2. It is all right to give the TOEIC Speaking test to people who do not like computers.	3.09	1.17	20 (37.0)	11 (20.4)	23 (42.6)
3. Taking the speaking test on computer was a pleasant experience.	3.37	0.98	10 (18.5)	20 (37.0)	24 (44.4)
4. It is not unnatural to talk to a computer during the speaking test.	3.26	1.12	15 (27.8)	14 (25.9)	25 (46.3)
Total	3.16	0.76			

[a] Means that Chi-square test was significant

people receiving high scores had a higher ability. In contrast, only 59.3% of participants considered the test a good indicator of their English-speaking ability in daily life.[6]

In comparison, the mean score for the subscale of perceived computer delivery revealed a more neutral tendency ($M = 3.16$), and the mean score for each item was around the scale's middle point. However, there was a larger variation in responses to individual items as confirmed by the non-significant Chi-square test results. Compared with positive views (37.0%), a slightly higher percentage of participants expressed reservations (42.6%) about giving the test to those without much computer experience, particularly older people who are unaccustomed to using computers, as illustrated in excerpt 1 below:

1. People who are not used to using computers, especially those who take the speaking test on a computer for the first time, will probably feel confused. Older people without much exposure to computers may feel very distracted during the test.

While 44.4% of participants felt that taking the test on a computer was a pleasant experience (excerpt 2), the rest either disagreed (18.5%) or felt neutral (37.0%):

2. It was interesting to take the test on the computer because it was my first time taking this kind of test. It was new for me to record my voice on the computer and to experience various types of tasks that seem unique, as I did not expect much besides reading paragraphs aloud.

In answering whether they thought it natural to talk to a computer during the speaking test, 46.3% of participants responded positively, but others felt negative (27.8%) or neutral (25.9%). They cited lack of feedback ($n = 3$), wearing headphones ($n = 1$), time pressure ($n = 1$), and being unable to ask questions ($n = 1$) as the main reasons they felt the testing situation was unnatural (excerpts 3 to 6, respectively):

3. Without an interviewer, I could not get any reactions, such as nodding or feedback, which intensified my anxiety.
4. It felt somewhat strange to wear headphones during the speaking test. I felt uneasy about whether I could answer the questions well with headphones, whether my response was properly recorded, and whether my pronunciation could be understood well.
5. I felt a lot of pressure to answer the questions quickly within the time limit. If it were an interview, the interviewer would wait for my response, so I could talk more comfortably at my own pace.
6. I was asked about what kind of TV program I like to watch. But actually, I do not watch TV because there is no television in my house. I felt very frustrated when I could not explain my situation and ask the computer to change the question. It is not flexible, which makes it different from a conversation with a person.

Test-taking motivation during the TOEIC Speaking test

Table 2 summarizes the descriptive and frequency statistics for items on perceived test importance and test-taking effort, which indicated that participants, on average, were sufficiently motivated.

Table 2 Descriptive and frequency statistics for questionnaire items on test-taking motivation

			Frequency (%)		
Item	*M*	*SD*	Disagree	Neutral	Agree
Perceived test importance					
1. Doing well on the TOEIC Speaking test was important to me.[a]	3.91	0.83	3 (5.6%)	9 (16.7%)	42 (77.7%)
2. I am concerned about the scores I will receive on the TOEIC Speaking test.[a]	4.41	0.60	0	3 (5.6%)	51 (94.4%)
3. The TOEIC Speaking test was an important test to me.[a]	3.72	0.81	4 (7.4%)	15 (27.8%)	35 (64.8%)
4. I would like to know how well I did on the TOEIC Speaking test.[a]	4.56	0.50	0	0	54 (100%)
Total	4.15	0.51			
Test-taking effort					
1. I engaged in good effort through the TOEIC Speaking test.[a]	4.19	0.70	2 (3.7%)	3 (5.6%)	49 (90.7%)
2. I gave the TOEIC Speaking test my full attention while completing it.[a]	3.85	0.76	4 (7.4%)	8 (14.8%)	42 (77.8%)
3. I gave my best effort on the TOEIC Speaking test.[a]	4.11	0.72	1 (1.9%)	8 (14.8%)	45 (83.3%)
4. While taking the TOEIC Speaking test, I was able to persevere in completing the tasks.[a]	3.93	0.91	3 (5.6%)	15 (27.8%)	36 (66.6%)
Total	4.02	0.63			

[a] Means that Chi-square test was significant

The subscale mean score for perceived test importance ($M = 4.15$) showed that participants were quite positive regarding the importance of the TOEIC Speaking test. Most participants (above 60%) chose to agree with items 1 and 3, suggesting that it was either rather important or very important for them to do well on the test. The reasons they cited were related to job hunting ($n = 1$) (excerpt 7) and competitive personality ($n = 1$) (excerpt 8):

7. Being able to speak English well will be one of my advantages when job hunting, so it is important for me to get good results to prove it.
8. Although test results will not affect my school record, being a competitive person, I usually want to get good results on all tests, especially when it comes to English.

Participants who did not regard the test results as important cited being nervous on tests ($n = 1$), taking the TOEIC Speaking test for the first time ($n = 1$), and needing to have their speaking skill assessed objectively ($n = 1$) (excerpts 9 to 11, respectively):

9. I always get very nervous on tests, so as long as I can speak well in daily conversations, it does not matter much to me whether I did well on the speaking test.
10. This is my first time taking the TOEIC Speaking test, and I did not know the content of the test very well, so it will not upset me very much if I did badly on the test.
11. I would be happy if I did well on the test, but it is more important to have my English-speaking skill assessed objectively than to get a good result.

In addition, nearly all participants agreed that they would like to know how well they did on the test (items 2 and 4). However, participants in different academic years seemed to have different concerns about their English-speaking skill (excerpts 12 to 14):

12. I had stayed in the USA for a year when I was a high school student. Now, I have to spend a lot of time learning my major language, and I feel my spoken English is not as good as when I came back from the USA, so I am eager to know what my spoken English level is now. [a first-year non-English major]
13. I took the interview test administered by the university the other day, and the test result was not as good as I had expected. So, I wanted to know whether I could do a better job on the TOEIC Speaking test. [a second-year non-English major]
14. I do not need to take many English classes now, so I am worried that my English-speaking skill is getting worse. I want to know my present English-speaking level and then decide whether to take more classes to improve it. [a third-year non-English major]

As shown in Table 2, participants, on average, reported putting forth a good effort on the TOEIC Speaking test ($M = 4.02$). Most participants agreed that they engaged in a good effort (90.7%) or gave their best effort (83.3%), whereas a slightly lower percentage of participants reported paying full attention during the test (77.8%) or toward completing tasks (66.6%). Regarding distractions, interviewees mentioned testing environment ($n = 2$) (excerpt 15) and confusions about how to use extra time ($n = 2$) (excerpt 16):

15. Even with headphones, I could hear others' voices and got distracted by them, especially during the read-aloud task. I could have concentrated more in a quieter environment.
16. When there was extra time for responses, I was not sure about whether I should say as much as possible or just keep silent, waiting for the computer to finish the recording.

Correlations between test-taker perception, test-taking motivation, and test performance

Table 3 provides descriptive statistics and Spearman rank-order correlation coefficients for all variables. There were statistically significant correlations between test-taker perception and test-taking motivation sub-dimensions: Test-taking effort was weakly correlated to perceived computer delivery, $r_s = .30$, $p < .05$, and moderately correlated to perceived test importance, $r_s = .44$, $p < .01$; perceived test validity had a weak correlation with perceived test importance, $r_s = .31$, $p < .05$. The rule of thumb for interpreting the relationship in a correlation was based on Guilford (1956): $r_s < .20$: very weak correlation; $.20 < r_s < .40$: weak correlation; $.40 < r_s < .70$: moderate correlation; $.70 < r_s < .90$: strong correlation; $r_s > .90$: very strong correlation.

Contrary to our expectation, self-perceived test performance was not significantly correlated to the sub-dimensions of either test-taker perception or test-taking motivation; this means that participants' perception of their test performance did not affect their perceived test validity, perceived computer delivery, perceived test importance, or test-taking effort. As expected, there was statistically significant correlation between the TOEIC LR score and the TOEIC speaking score, $r_s = .59$, $p < .01$, suggesting that the TOEIC LR score may confound the relationships of the TOEIC speaking score with other variables.

Table 3 Summary of mean, standard deviations, and intercorrelations for scores on the examined variables

Variable	*M*	*SD*	1	2	3	4	5	6	7
1. TOEIC speaking score	134.63	19.00	–						
2. Perceived test validity	3.85	.64	− .07	–					
3. Perceived computer delivery	3.16	.76	− .03	− .01	–				
4. Perceived test importance	4.15	.51	.14	.31*	.04	–			
5. Test-taking effort	4.02	.63	.12	.07	.30*	.44**	–		
6. TOEIC LR score	767.22	115.06	.59**	.12	− .08	.11	− .04	–	
7. Self-perceived test performance	2.10	.68	.39**	− .23	.04	.07	.20	.05	–

**p* < .05; ***p* < .01

Given these results, semi-partial correlations were calculated between the TOEIC speaking score and subscale scores of test-taker perception and test-taking motivation, with the effect of the TOEIC LR score on the TOEIC speaking score held constant. As shown in Table 4, there was no statistically significant correlation between the TOEIC speaking score and any of other variables. Speaking test performance was negatively related to perceived test validity and positively related to test-taking effort, but neither of the semi-partial correlation coefficients was statistically significant.

Table 4 Semi-partial correlations between TOEIC Speaking score and other variables

Variable	TOEIC speaking score	*p*
TOEIC speaking score	–	–
Perceived test validity	− .23	.09
Perceived computer delivery	.01	.94
Perceived test importance	.10	.46
Test-taking effort	.27	.06

Discussion

In the following sections, the three research questions are discussed in the light of the results obtained.

How do Japanese university students perceive the TOEIC Speaking test in terms of test validity and computer delivery? What are the possible reasons underlying their perceptions? (RQ1)

Test-taker perception was explored from the perspectives of both test validity and computer delivery. Participants were found to have moderately positive perceptions of the construct validity, predictive validity, and content validity of the TOEIC Speaking test. Our results were consistent with those of previous studies conducted on computer-delivered speaking tests (Fan, 2014; Kiddle & Kormos, 2011; Zhou, 2012). In contrast, participants indicated more neutral and conservative attitudes toward computer delivery. Participants also commented that they found the lack of interaction undesirable, mainly due to wearing headphones, time pressure, being unable to ask questions, and receiving no feedback from the computer. As with previous qualitative studies (Brooks & Swain, 2015; Fan, 2014; Qian, 2009), our findings provide much needed empirical evidence that a lack of interaction is a crucial aspect of test-takers' unfavorable perceptions of computer-delivered speaking tests.

However, it was also found that perception of test validity was unrelated to perception of computer delivery. One explanation for this result is that the two kinds of test-taker perception are distinct. Even if test-takers find using computers to deliver the TOEIC Speaking test unfair or unnatural, they think highly of the test in terms of test validity. During interviews, some participants commented that they regarded the test as valid because of its reputation. Others seemed to consider the testing situation separately from natural conversations, as two interviewees mentioned that they considered it a viable compromise to use a computer to deliver a speaking test, given the potential issues of procedural standardization and practicality of administration.

Another viable explanation is that the two kinds of test-taker perception are related, but their relationship was undetected due to the unsatisfactory reliability of the perceived computer delivery scale ($\alpha = .61$). Note that the scale only included four items that were designed to measure several different aspects of perception related to computer delivery. Considering the fact that attitudes can consist of beliefs, opinions, and emotions (Murray et al., 2012), the number of items may have been insufficient to provide a broad enough representation of the construct to achieve a satisfactory level of internal consistency.

To what extent do Japanese university students report feeling motivated during the TOEIC Speaking test? What are the possible factors underlying their test-taking motivation? (RQ2)

The participants, on the whole, were found to be sufficiently motivated: They rated the TOEIC Speaking test highly in terms of importance and reported feeling motivated to do their best during the test. These results are contrary to our concern that, as a low-stakes test, the TOEIC Speaking test might be associated with lower levels of test-taking motivation. Rather, the reported test-taking motivation levels were similar to those reported for the College English Test, a high-stakes English test in China (Zhao & Cheng, 2010).

These unexpected results may be largely attributed to the fact that our participants were not "real" test-takers, as they volunteered and were exempted from test charges. Accordingly, one may consider this context a poor proxy for low-stakes testing situations. Our interview data showed that, although the TOEIC Speaking test results did not affect students' academic records, they valued good test performance for other reasons. Intrinsic reasons they mentioned included knowing and improving their spoken English level or having competitive personality. They also cited extrinsic reasons to succeed in the test. Since they studied foreign languages at an esteemed Japanese university, students considered themselves proficient at English, as compared to their peers in other universities. As a result, even if the TOEIC Speaking test were a low-stakes test, it may have been regarded as important by these students as a good opportunity to confirm their speaking abilities through objective assessment (cf. excerpt 11).

To what extent are test-taker perception, test-taking motivation, and performance on the TOEIC Speaking test related? (RQ3)

A key finding of this study was that speaking test performance was not significantly related to either test-taker perception or test-taking motivation, suggesting that these two variables did not account for significant proportions of variance in the TOEIC speaking score.

Although test-taking effort, an important indicator of test-taking motivation, was found to be positively correlated to the TOEIC speaking score, the effect was not statistically significant, which did not support findings from previous studies (e.g., Zilberberg et al., 2014). One of the possible explanations for the discrepancy lies in the different types of test items used. Tests in most previous studies consisted of multiple-choice items, whereas most tasks in the TOEIC Speaking test required test-takers to formulate responses.[7] In open-ended responses, length of responses may not accurately reflect test-taking effort. Our participants may not have used all the time available when they responded: They may have stopped to think about what to say during a response or had nothing more to add after giving a response. If they had interpreted such behaviors as not giving their best effort or not persevering in completing the tasks, their replies to questionnaire items may have inaccurately reflected their actual test-taking effort and hence its relationship with test performance. To better understand the issue, future studies would benefit from asking participants how they interpreted the questionnaire items about their engagement with test tasks.

The lack of relationship between test-taking effort and test performance may also be due to the restricted range of English proficiency among the participants. The levels of the participants' English-speaking proficiency were relatively high, compared to those of average Japanese university students.[8] Given that restrictions of this range can impact reliability results (Sackett, Laczo, & Arvey, 2002), the correlation could have been attenuated.[9] Therefore, if a sample with a wider range of speaking proficiency were examined, the correlation coefficient which was close to being statistically significant ($p = .06$) may have reached statistical significance.

Our results provided supportive evidence for the relationship between test-taker perception and test-taking motivation: The two variables were significantly correlated, but in different ways regarding their sub-dimensions. Perceived test validity seemed to be indirectly related to test-taking effort through the mediation of perceived test importance. In other words, those who perceive the test as more valid tend to consider the test as more important, which in turn leads to increased test-taking effort. The findings are consistent with previous studies (e.g., Wise, 2009) and lend empirical support to the expectancy-value theory, suggesting that test-takers base their judgments of test importance on test validity, rather than on computer delivery.

Conversely, perceived computer delivery seemed directly related to test-taking effort. The more positive attitudes test-takers showed toward using computers to test speaking skill, the more effort they reported expending to complete the test. This result could mean that, due to the lack of feedback from the computer, the computer-delivered speaking test made it difficult for participants to concentrate on responses; thus, they reported lower levels of perseverance in completing the test. According to the interviewees' comments, the sources of distraction they felt (e.g., other test-taker voices (cf. excerpt 17), confusion about how to use extra time (cf. excerpt 16), and frustration at being unable to ask questions (cf. excerpt 6)) were primarily associated with using computers during testing.

Based on the above discussion, it was hypothesized that test-taker perception influences test performance through the mediation of test-taking motivation. Specifically, as shown in Fig. 1, perceived test validity influences test-taking effort indirectly through

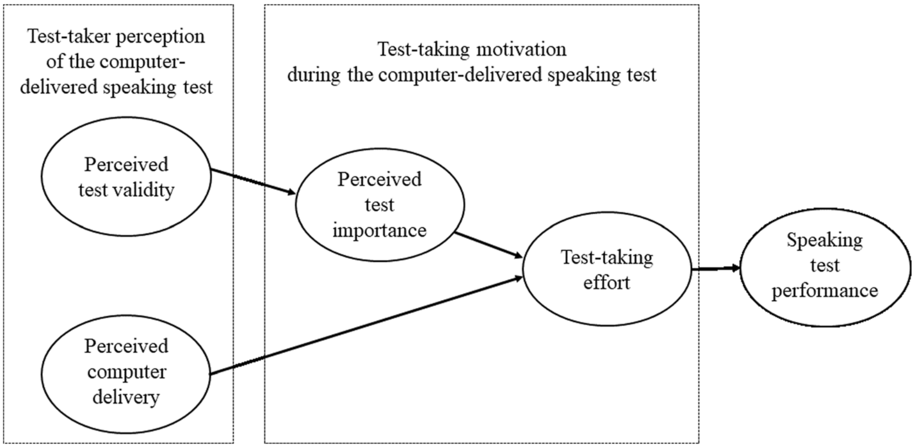
Test-taker perception of the computer-delivered speaking test
Test-taking motivation during the computer-delivered speaking test
Perceived test validity
Perceived computer delivery
Perceived test importance
Test-taking effort
Speaking test performance

Fig. 1 Proposed relationships between test-taker perception, test-taking motivation, and test performance. It is proposed that perceived test validity influences test-taking effort indirectly through the mediation of perceived test importance. Unfavorable perception of computer delivery will lead to decreased test-taking effort, which in turn affects test performance.

the mediation of perceived test importance, and perceived computer delivery has a direct effect on test-taking effort, which in turn affects test performance.

Note that the proposed relationships in Fig. 1 are exploratory in nature and should therefore be interpreted with the following methodological caveats in mind. First, measures of test-taker perception and test-taking motivation were collected after participants took the test, which did not satisfy the temporal precedence needed for a causal claim. Participants may have expressed different perceptions because of different levels of test-taking motivation. To draw a causal inference, future research should recruit only those who have previously taken the test and record their test-taker perception prior to the study. Those without previous test-taking experiences should indicate their test-taker perception after reading sample test items or taking a sample test, but before starting the actual test. Second, given our relatively small sample size, only correlation analyses were performed, which did not permit the validation of indirect relationships between variables. Further investigation is needed using a larger sample and advanced statistical analysis methods, such as path analysis or structural equation modeling.

Limitations and directions for future research

A potential limitation to this study was the measurement of perceived computer delivery. In addition to the narrow representation of the construct as discussed earlier, it is possible that the questionnaire items may have been too generic. For example, the item "Taking the speaking test on a computer was a pleasant experience" could refer to pleasant reactions to various aspects of the test-taking experience. This statement could be more effectively revised as "Taking the speaking test on a computer made me feel confident about my English skills," "I felt relaxed while taking the speaking test," or "I enjoyed talking to a computer in the speaking test." Moreover, the item "It is unnatural to talk to a computer during a speaking test" could be improved by adding specific reasons such as "because I was not able to ask questions or interrupt" or "because I could not speak at my own pace."

Caution is also warranted in generalizing these findings to different populations and computer-delivered speaking tests. As the TOEIC Speaking test targets a wide range of test-takers, the findings of this study using a sample of students may not be generalizable to other groups (e.g., business persons and office workers or students with a wider range of English abilities). In addition, the results may not apply to speaking tests using different types of tasks from those in the TOEIC Speaking test (particularly video-based tasks, which are likely to be perceived differently from those using only written or audio prompts).

Further research is needed to investigate other related affective variables, such as test anxiety. As Cheng et al. (2014) has argued, test anxiety is usually related to test-taking motivation, so they should be investigated together. Moreover, all the interviewees in our study reported feeling motivated to improve their speaking skill after taking the TOEIC Speaking test. Longitudinal studies, therefore, would provide more insights into how test-taking experience may affect test-taker perception, test-taking motivation, and performance on the subsequent tests to take.

Conclusion

This study was conducted to enhance our understanding of whether and how test-taker reactions to computer-delivered speaking tests may affect test performance. Contrary to our predictions, the influence of test-taker perception and test-taking motivation on test performance was negligible. Our findings seemingly eliminate the concern that these two variables represent important sources of irrelevant variance in computer-delivered speaking test scores, thereby validating inferences made from such tests.

Nevertheless, it is important for test developers to monitor test-taker perception, particularly regarding computer delivery, as our findings imply that test-takers hold less favorable attitudes toward computers. To promote more positive attitudes, a tutorial could be added to tests.[10] Also, given the findings of the relationship between test-taker perception and test-taking motivation, it is important for testing agencies to promote tests, which may lead to increased perception of test importance and may thus, help test-takers succeed.

Despite the aforementioned limitations, we believe that this study is an important initial attempt to explore the relationship between test-taker perception and test performance that provides tentative evidence for the mediational role of test-taking motivation in language testing. We hope that this study will add to the growing body of research about computer-delivered speaking test score interpretations and will provide the groundwork for continued research into test-taker reactions toward computer-delivered speaking tests.

Endnotes

[1] Expectancy-value theory asserts that effort in the case of testing is a function of expectancy and task value. Although expectancy-value theorists agree on these two determinants of motivation, there are subtle differences in the precise description of expectancies and values. Eccles and Wigfield (2002) theorized that expectancies consist of expectancies for success and ability beliefs. Value consists of attainment value, intrinsic value, utility value, and cost.

[2] Self-serving bias means that test-takers may attribute their poor performance to the test being invalid, or to their low motivation in taking the test (Chan, Schmitt, Jennings, Clause, & Delbridge, 1998).

[3] We used the Student Opinion Scale because it is based on expectancy-value theory and has shown satisfactory estimates of reliability and substantial evidence of validity (Sundre, 1999; Sundre & Moore, 2002).

[4] We did not ask participants to respond to the questionnaire prior to the test, as they were not expected to have any previous experience of taking the test.

[5] As this study was a part of a project to collect data on the effects of the test-taking experience on English learning, time between the administration of the test and the interview was required.

[6] Although the TOEIC Speaking test is intended to measure English-speaking ability in both daily life and global workplace, we did not include "the global workplace" in the statement, considering that our participants who are university students may not be familiar with the global workplace. Therefore, participants who showed disagreement with this statement might consider the test a good way to learn how well people can speak English in the global workplace.

[7] Evidence has suggested that the type of items test-takers are presented with influences the relationship between test-taking motivation and performance (e.g., DeMars, 2000).

[8] The average TOEIC speaking score was 134.6 for the present sample, whereas it was 95.3 for Japanese university students, 103.3 for English majors, and 104.0 for non-English foreign language majors in fiscal year 2015 (Institute for International Business Communication, 2016). The average TOEIC LR score was 767 for the present sample, whereas it was 443 for Japanese university students, 503 for English majors, and 450 for non-English foreign language majors in fiscal year 2015 (Institute for International Business Communication, 2016).

[9] We failed to calculate attenuated correlations between the TOEIC speaking score and other variables, due to lack of access to information on the reliability of TOEIC speaking scores.

[10] Evidence has shown that adding a computer-administered tutorial to the TOEFL computer-based test was effective to increase test-takers' acceptance of the test (Jamieson, Taylor, Kirsch, & Eignor, 1999).

Abbreviations

IP: Institutional Program; RQ: Research question; TOEFL iBT: Test of English as a Foreign Language Internet-based Test; TOEFL: Test of English as a Foreign Language; TOEIC LR: Test of English for International Communication Listening and Reading; TOEIC SW: Test of English for International Communication Speaking and Writing; TOEIC: Test of English for International Communication

Acknowledgments

We are grateful to Professor Yasuyo Sawaki for her constructive feedback on an earlier version of the manuscript. We also thank Dr. Mutsumi Inokawa for her help with data collection and Mariko Yoshida for checking the translation of the questionnaire. Finally, we are grateful to all the students who participated in the study.

Funding

This research was funded by the Grant-in-Aid for Scientific Research in Japan (No. 26284076 and 18H00685).

Authors' contributions

Both authors were responsible for the research design and data interpretation. YZ performed the statistical analysis and drafted the manuscript, and AY provided comments and edited the manuscript. Both authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 11 March 2019 Accepted: 16 May 2019
Published online: 26 June 2019

References

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge: Cambridge University Press.

American Educational Research Association, American Psychological Association, & National Council for Measurement in Education. (2004). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Atkinson, J. W. (1957). Motivational determinants of risk taking behavior. *Psychology Review, 64*, 359–372. https://doi.org/10.1037/h0043445.

Bachman, L., & Palmer, A. (2010). *Language assessment in practice*. Oxford: Oxford University Press.

Baumert, J., & Demmrich, A. (2001). Test motivation in the assessment of student skills: the effects of incentives on motivation and performance. *European Journal of Psychology of Education, 16*(3), 441–462. https://doi.org/10.1007/BF03173192.

Brooks, L., & Swain, M. (2015). Students' voices: the challenge of measuring speaking for academic contexts. In B. Spolsky, O. Inbar, & M. Tannenbaum (Eds.), *Challenges for language education and policy: making space for people* (pp. 65–80). New York: Routledge.

Chan, D., Schmitt, N., Jennings, D., Clause, C. S., & Delbridge, K. (1998). Applicant perceptions of test fairness integrating justice and self-serving bias perspectives. *International Journal of Selection and Assessment, 6*(4), 232–239. https://doi.org/10.1111/1468-2389.00094.

Cheng, L. Y., Klinger, D., Fox, J., Doe, C., Jin, Y., & Wu, J. (2014). Motivation and test anxiety in test performance across three testing contexts: the CAEL, CET, and GEPT. *TESOL Quarterly, 48*(2), 300–330. https://doi.org/10.1002/tesq.105.

Cole, J. S., & Osterlind, S. J. (2008). Investigating differences between low- and high-stakes test performance on a general education exam. *The Journal of General Education, 57*(2), 119–130. https://doi.org/10.1353/jge.0.0018.

DeMars, C. E. (2000). Test stakes and item format interactions. *Applied Measurement in Education, 13*(1), 55–77. https://doi.org/10.1207/s15324818ame1301_3.

Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology, 53*, 109–132. https://doi.org/10.1146/annurev.psych.53.100901.135153.

Educational Testing Service. (2016). *Examinee handbook: speaking and writing*. Retrieved from http://www.etsglobal.org/content/download/828/12618/version/5/file/Examinee+Handbook+-+TOEIC+Speaking+and+Writing-LR.pdf

Educational Testing Service. (2019). *TOEIC user guide–speaking & writing*. Retrieved from https://www.ets.org/s/toeic/pdf/toeic_sw_score_user_guide.pdf.

Elder, C., Iwashita, N., & McNamara, T. (2002). Estimating the difficulty of oral proficiency tasks: what does the test-taker have to offer? *Language Testing, 19*(4), 347–368. https://doi.org/10.1191/0265532202lt235oa.

Fan, J. S. (2014). Chinese test takers' attitudes towards the Versant English test: a mixed-methods approach. *Language Testing in Asia, 4*(1), 1–17. https://doi.org/10.1186/s40468-014-0006-9.

Fan, J. S., & Ji, P. Y. (2014). Test candidates' attitudes and their test performance: the case of the Fudan English test. *University of Sydney Papers in TESOL, 9*, 1–35. Retrieved from http://faculty.edfac.usyd.edu.au/projects/usp_in_tesol/pdf/volume09/Article01.pdf.

Guilford, J. P. (1956). *Fundamental statistics in psychology and education*. New York: McGraw-Hill Book.

Institute for International Business Communication. (2016). *TOEIC program: data & analysis 2016*. Retrieved from http://www.iibc-global.org/library/toeic_data/toeic/pdf/DAA.pdf.

Institute for International Business Communication. (2018). *TOEIC program: data & analysis 2018*. Retrieved from https://www.iibc-global.org/library/default/toeic/official_data/pdf/DAA.pdf

Jamieson, J., Taylor, C., Kirsch, I., & Eignor, D. (1999). *Design and evaluation of a computer-based TOEFL tutorial (TOEFL research report no. 62)*. Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1999.tb01799.x.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Washington, DC: American Council on Education/Praeger.

Kiddle, T., & Kormos, J. (2011). The effect of mode of response on a semidirect test of oral proficiency. *Language Assessment Quarterly, 8*(4), 342–360. https://doi.org/10.1080/15434303.2011.613503.

Liao, C. W., & Qu, Y. X. (2010). *Alternate form test-retest reliability and test score changes for the TOEIC Speaking and Writing tests (TOEIC compendium no. 10.2)*. Princeton: Educational Testing Service. Retrieved from https://www.ets.org/Media/Research/pdf/TC-10-10.pdf.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: Macmillan.

Ministry of Education, Culture, Sports, Science, & Technology. (2017). *Daigaku nyugaku kyoutsu tesuto jishi houshin* [guidelines for implementing the Common Test]. Retrieved from http://www.mext.go.jp/component/a_menu/education/micro_detail/__icsFiles/afieldfile/2017/10/24/1397731_001.pdf

Murray, J. C., Riazi, A. M., & Cross, J. L. (2012). Test candidates' attitudes and their relationship to demographic and experiential variables: the case of overseas trained teachers in NSW, Australia. *Language Testing, 29*(4), 577–595. https://doi.org/10.1177/0265532212440690.

National Center of University Entrance Examinations. (2018). *Daigaku nyushi eigo seiseki teikyo shisutemu sanka youken wo mitashiteiru koto ga kakuninsareta shikaku/kentei shiken* [a list of English tests approved to be compatible with requirements of providing test results under the university admission exam system]. Retrieved from https://www.dnc.ac.jp/news/20180326-02.html

Pintrich, P. R. (1999). The role of motivation in promoting and sustaining self-regulated learning. *International Journal of Educational Research, 31*(6), 459–470. https://doi.org/10.1016/S0883-0355(99)00015-4.

Pintrich, P. R., & Schunk, D. H. (2002). *Motivation in education: theory, research, and applications (2nd ed.)*. Columbus: Merrill.

Qian, D. D. (2009). Comparing direct and semi-direct modes for speaking assessment: affective effects on test takers. *Language Assessment Quarterly, 6*(2), 113–125. https://doi.org/10.1080/15434300902800059.

Sackett, P. R., Laczo, R. M., & Arvey, R. D. (2002). The effects of range restriction on estimates of criterion interrater reliability: implications for validation research. *Personnel Psychology, 55*(4), 807–825. https://doi.org/10.1111/j.1744-6570.2002.tb00130.x.

Sanchez, R., Truxillo, D., & Bauer, T. (2000). Development and examination of an expectancy-based measure of test-taking motivation. *The Journal of Applied Psychology, 85*(5), 739–750. https://doi.org/10.1037/0021-9010.85.5.739.

Sato, T., & Ikeda, N. (2015). Test-taker perceptions of what test items measure: a potential impact of face validity on student learning. *Language Testing in Asia, 5*(10), 1–16. https://doi.org/10.1186/s40468-015-0019-z.

Stricker, L. J., & Attali, Y. (2010). *Test takers' attitudes about the TOEFL iBT (TOEFL iBT research report no. 13)*. Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2010.tb02209.x.

Sundre, D. L. (1999). *Does examinee motivation moderate the relationship between test consequences and test performance?* Retrieved from ERIC database. (ED432588).

Sundre, D. L., & Moore, D. L. (2002). The Student opinion Scale: a measure of examinee motivation. *Assessment Update, 14*(1), 8–9.

Swing, R. L. (2001). Dedicated assessment days: mobilizing a campus' efforts. *Assessment Update, 13*(6), 1–15.

Tannenbaum, R. J., & Wylie, C. E. (2008). *Linking English-language test scores onto the Common European Framework of Reference: an application of standard-setting methodology (TOEFL iBT research report no. 6)*. Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2008.tb02120.x.

Wigfield, A., & Eccles, J. S. (2000). Expectancy-value theory of achievement motivation. *Contemporary Educational Psychology, 25*(1), 68–81. https://doi.org/10.1006/ceps.1999.1015.

Wise, S. L. (2009). Strategies for managing the problem of unmotivated examinees in low-stakes testing programs. *Journal of General Education, 58*(3), 152–166. https://doi.org/10.1353/jge.0.0042.

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: problems and potential solutions. *Educational Assessment, 10*(1), 1–17. https://doi.org/10.1207/s15326977ea1001_1.

Xie, Q. (2011). Is test taker perception of assessment related to construct validity? *International Journal of Testing, 11*(4), 324–348. https://doi.org/10.1080/15305058.2011.589018.

Zhao, J., & Cheng, L. (2010). Exploring the relationship between Chinese university students' attitude towards the college English test and their test performance. In L. Cheng & A. Curtis (Eds.), *English language assessment and the Chinese learner* (pp. 190–201). New York: Routledge, Taylor & Francis Group.

Zhou, Y. J. (2012). Test-takers' affective reactions to a computer-delivered speaking test and their test performance. In M. Minegishi, O. Hieda, E. Hayatsu, & Y. Kawaguchi (Eds.), *Working papers in corpus-based linguistics and language education, no. 9* (pp. 295–310). Tokyo: Tokyo University of Foreign Studies. Retrieved from http://cblle.tufs.ac.jp/assets/files/publications/working_papers_09/section/295-310.pdf.

Zilberberg, A., Finneya, S. J., Marsha, K. R., & Anderson, R. D. (2014). The role of students' attitudes and test-taking motivation on the validity of college institutional accountability tests: a path analytic model. *International Journal of Testing, 14*(4), 360–384. https://doi.org/10.1080/15305058.2014.928301.