Sato and Ikeda *Language Testing in Asia* (2015) 5:10
DOI 10.1186/s40468-015-0019-z

Language Testing in Asia
a SpringerOpen Journal

RESEARCH Open Access

# Test-taker perception of what test items measure: a potential impact of face validity on student learning

Takanori Sato[1*] and Naoki Ikeda[2]

* Correspondence:
taka-sato@lab.tamagawa.ac.jp
[1]Center for English as a Lingua Franca, Tamagawa University, Tokyo, Japan
Full list of author information is available at the end of the article

## Abstract

**Background:** High-stakes tests have an immense washback effect on what students learn and affect the content of student learning. However, if students fail to recognize the abilities that the test developers intend to measure, they are less likely to learn what the test developers wish them to learn. This study aims to investigate test-taker perception of the ability being measured by items (i.e., face validity) in high-stakes tests and examines the extent to which test-taker perception and test developer intention agree.

**Methods:** University students in Japan and Korea (N = 179) were given past entrance examinations administered in the respective countries and asked to read test items and record what ability they thought each item was measuring.

**Results:** Although the overall agreement rate was moderately high, items aiming to measure an ability to read between the lines were perceived to be measuring an ability to understand the content objectively. Furthermore, many participants perceived items designed to indirectly measure writing ability as those tapping into reading ability.

**Conclusions:** Face validity could be integrated for test development with the ultimate aim of promoting positive washback on students, which should be one of the intentions of test developers. In order to obtain the positive and intended washback effect on English learning, the present study suggests that the Japanese and Korean test committees need to (a) widely inform test-takers of the ability measured by each test item and (b) incorporate performance testing that measures test-takers' productive skills more directly.

**Keywords:** Face validity; Washback; Entrance examination; Test-taker perception

## Background

In some monolingual EFL countries such as Japan and Korea, high-stakes examinations are the primary opportunities to use English and a powerful instrumental motivation to learn English. Since high-stakes examinations are extremely influential to the students, having an immense impact on what is to be taught and learned (Cheng, 2008), test developers often develop tests in order to bring about the intended impact on language learning or *washback*. In many cases, the introduction or revision of examinations has been done to promote positive washback on teaching and learning. For example, developers of Test of English for Academic English (TEAP)—a high-stakes test jointly developed by Sophia University and EIKEN recently—explicitly stated their expectation that the newly introduced test

will draw attention to productive and cognitive skills and that it would lead to better English education and learning in Japan (Eiken Foundation of Japan, n.d.). In Korea, similarly, a key purpose of developing a new university entrance examination in Korea—the National English Ability Test (NEAT)—was claimed to be "to activate the teaching of speaking and writing of English at schools, which has not actually been conducted despite its compelling necessity" (Lee, 2012, p. 30). As can be seen, language tests are often used as de facto language education policies (Menken, 2008) and disciplinary tools (Shohamy, 2001), and test developers measure abilities and skills that they wish test-takers to focus on in their learning activities.

In this context, the gap between test-takers' perception of the skills being measured and test developers' intention in developing items is arguably one of the factors that undermines intended positive washback on learning. If test-takers fail to correctly recognize which skills the test developers intend to measure, they are not likely to focus on and spend time on the skills that the test developers wish them to pay attention. We investigate test-taker perception of the skills being measured by high-stakes examinations administered in two EFL countries, Japan and South Korea. More specifically, the study examines to what extent test-taker perception and test developer intention agree and discusses the importance of investigating test-taker perception or *face validity* when new tests are developed and incorporated.

## Face validity

Davies et al. (1999) define face validity as "the degree to which a test appears to measure the knowledge or abilities it claims to measure, as judged by an untrained observer" (p. 59). Accordingly, face validity is the non-experts' (e.g., test-takers') judgment of the test construct. For example, face validity is considered poor when a dictation task is used for measuring reading ability but the test-takers perceive it as a task measuring an unrelated ability (Davies et al., 1999).

Since it is based on lay people's intuitive judgments about the test, face validity has been dismissed as scientific evidence for construct validity and thus taken for granted in the field of educational measurement as well as language testing. For example, Stevenson (1985) criticizes the reliance on face validity in language tests, claiming that scientific evidence (e.g., criterion-related validation) is likely to be ignored. In fact, a test with low face validity does not necessarily lead to invalid score interpretation and use of the test. For instance, a cloze test may not be perceived as a test for proficiency by its appearance, but the theory underlying the test supports its use for measuring language proficiency (Akbari, 2012). Therefore, non-expert judgment of test appearance is not always a major concern of test developers and language testing researchers.

Meanwhile, a number of language testing researchers have acknowledged the importance of face validity and encouraged research into this factor. It is claimed that test-takers' performance might be negatively affected by low face validity. If the test is perceived as irrelevant to the claimed purposes, test-takers put less effort into the test and the scores may not accurately reflect their ability (Alderson, Clapham, & Wall, 1995; Bachman, 1990; Brown & Abeywickrama, 2010; Hughes, 2003; Kane, 2006). This is considered to undermine the validity of score use and interpretation, as Kane (2006) states,

"to the extent that students put less effort into their performance on a test than they would on the corresponding tasks in other settings, because the test seems irrelevant, the extrapolation inference would be weakened" (p. 36). Accordingly, Davies et al. (1999) argue that "the notion of test appeal is a practical consideration which test designers cannot afford to overlook" (p. 59).

Although the benefits of investigating face validity have been acknowledged by some researchers, there are only a few empirical studies on stakeholders' judgments of the construct of language tests. For example, Brown (1993) investigated test-takers' reactions to the Occupational Foreign Language Test that measures Japanese proficiency in the tourism and hospitality industry. A questionnaire survey was conducted to elicit the test-takers' perceptions of the skills measured by the test. The results indicated that the face validity of the test was not satisfactory for students undertaking general Japanese courses because some test tasks were not perceived to tap into their language proficiency. Brown (1993) claims that the test-taker perception of the test can be used to develop a fair and accessible test for all prospective test-takers. More recently, So (2014) collected language teachers' perspectives on the construct of TOEFL Junior and sought a way to incorporate them into test development. Concurring with Brown (1993), So concludes that stakeholders' involvement in test development helps to improve test quality and leads to their acceptance of the test.

### Washback and face validity

It is currently acknowledged that tests, in particular high-stakes tests, have a large impact on learning and teaching, referred to as *washback* (Cheng, 2008; Wall, 2012). Because of this characteristic, such tests are frequently used to influence education systems or curriculum in various contexts, including the content of lesson, textbooks, and pedagogical approaches used in the classroom. In fact, it has been widely believed that the improvement of high-stakes tests, including university entrance examinations, directly results in the improvement of education. This belief has been particularly espoused by the general public and is even popular among language teachers. However, it is also acknowledged that washback is a complex phenomenon and that tests do not necessarily bring about the improvement of education. In particular, empirical studies have shown that tests do not have direct impacts on teaching methods or approaches employed in the classroom (e.g., Watanabe, 1996, 2004).

Although the complexity of washback has been acknowledged, it has been claimed that tests seem to have direct impacts on the content of teaching and learning, which has been supported by various empirical studies (e.g., Ferman, 2004; Hawkey, 2006; Qi, 2004; Wall, 1996, 2005; Wall & Alderson, 1993). For example, Wall and Alderson (1993) investigated the washback effect of a newly employed English examination in Sri Lanka, focusing on its impact on language teaching. A series of classroom observations were conducted to examine the nature of classes. As a result, they found washback effects of the test on the content of teaching, but not on the methodology employed by the teachers. The effects were confirmed during both the normal instruction periods and the examination preparation periods (Wall, 2005). Language teachers appear to select the content of teaching dealt with in the classroom based on the content of tests

and tend to concentrate on the skills measured by tests. In a similar vein, the content of students' learning appears to be strongly influenced by language tests, in particular by high-stakes examinations. Ferman (2004) found that a new national EFL oral matriculation test employed in Israel had a direct washback effect on students' learning. The new test was developed in an attempt to emphasize oral communicative skills. A questionnaire survey to students showed that the majority of the students increased their focus on learning oral skills after the introduction of the test. Although the methods for improving them were not necessarily considered positive (e.g., it led to memorization and rote learning), students' attention to oral skills was at least achieved by the new high-stakes test. Moreover, students believed that the test improved their overall command of English.

Given the tests' potential washback on the content of teaching and learning, language tests are often used as a powerful tool for bringing about intended washback effects, in particular on the content of teaching and learning. In fact, the attainment of intended washback is a crucial part of test developers' responsibility. Fulcher and Davidson (2007) argue that language testers should carefully consider the intended impact on all the stakeholders in the initial phase of test development, which is referred to as *effect-driven testing*. Similarly, Bachman and Palmer (2010) claim that language testers need to take into account the intended consequences of the test in test-design.

In this context, the test users' perceptions of the test construct or skills measured in the test are considered to play a crucial role in attaining the intended washback. The content of teaching and learning is susceptive to their perception of the test construct rather than the actual test construct the test developer intends to measure. For example, even though academic English proficiency may be the construct of a high-stakes test that the test developer wishes to measure, students will not attempt to consciously improve academic English skills if they perceive that the test measures general English proficiency. Likewise, it can be considered that the test users are influenced by what skill or ability each item appears to be measuring. With regard to the role of face validity in washback, Watanabe (2004) argues,

> When we come to think of washback, the important test quality may not only be the validity and reliability of the psychometric tradition, nor the notion of "consequential validity" (Messick, 1989), but a type of "face validity" understood from the viewpoint of the test users. (p. 142)

This suggests that considering face validity is one of the crucial responsibilities of test developers. In particular, the ability that test developers claim to measure should correspond to the ability that test users (e.g., teachers, students) believe the test measures. At a minimum, this is a necessary condition for achieving the intended washback on the content of teaching and learning.

In summary, face validity has not always been regarded as scientific evidence for the validity of language tests since it is non-experts' (test-takers') perception of what the test measures. However, investigating test-takers' perception of the test construct is considered important for test developers if they wish to achieve intended washback on the content of English learning. Since in some monolingual EFL countries, English tests—in particular, high-stakes examinations—are often used to control students' English

learning, it is imperative to confirm the extent to which test-takers accurately recognize what test developers intend to measure in tests.

### The study

The present study investigates the face validity of nation-wide influential university entrance examinations administered in two EFL countries, Japan and Korea. More specifically, the study examines test-taker perception of the ability being measured by items in the examinations and determines the extent to which test-taker perception and test developer intention agree. The following two research questions are addressed:

1. To what extent do test-taker perceptions of the ability measured by each item and test committee intentions agree?
2. What types of test items induce gaps between test-taker perceptions and test committee intentions?

## Methods

### Examinations

This study collected test-taker perception of a sample of items derived from two high-stakes examinations administered in the past: the National Center Test for University Admissions of Japan (NCTUA) and the College Scholastic Ability Test of Korea (CSAT).

The NCTUA is developed by the Japan National Center for University Entrance Examinations (NCUEE) and administered annually in January. The primary purpose of the test is claimed to be to measure the university applicants' abilities concerning basic content learned in senior high school (NCUEE, n.d.). The English test in the NCTUA consists of a reading section (200 points) and a listening section (50 points), and one of the main constructs of the test is claimed to be *practical communication competence*. At the same time, it aims to improve English teaching and learning practice in Japan and help to improve English skills including speaking and writing skills (Watanabe, 2013). All items are multiple-choice items with four to five options. The reading section contains about 50 items measuring a wide range of knowledge including pronunciation, grammar, and reading comprehension, and the test-takers are required to read a relatively large amount of text and answer questions. In the listening section, each test-taker receives a portable player and individually listens to the prompts to answer the corresponding questions.

This study uses items in the reading section. The test committee claims that the reading section of the English test measures students' achievement of practical English communication abilities based on the senior high school study guideline. Additionally, it measures not only knowledge of vocabulary and grammar but also language proficiency including the sociolinguistic, discourse, and strategic aspects of language use.

The CSAT is developed by the Korea Institute of Curriculum and Evaluation (KICE) and administered annually in November. The purpose of the examination is to measure test-takers' academic ability as required for university education (KICE, 2012). Most universities use the CSAT as a first examination and administer their own in-house test as a second examination. Accordingly, most Korean students need to take the test to

enroll in a university in Korea. The English subtest of the CSAT contains a reading section (66 points) and a listening section (34 points). The English subtest measures the ability to use English as a communication tool, not simply to measure fragments of knowledge. Thus, it emphasizes measuring a communicative aspect of language ability. In addition, it aims to measure all four of the major skill areas: reading, listening, writing, and speaking. The test employs the multiple-choice format for all items to ensure fairness and objectivity.

This study uses the reading section of the English test. It is claimed to measure the ability to infer the main point of a passage and omitted content in the text and the ability to grasp details of conversations or paragraphs. The test also aims to indirectly assess writing ability by measuring the abilities to understand schematic organization and to summarize the passage.

### Participants

Data were collected from 80 Japanese and 98 Korean students who were studying at a Japanese or Korean national university at the time of data collection. These universities were located in rural areas, and their academic rankings were average among Japanese and Korean national universities. The students at each university were sophomores and juniors (19 to 20 years old) undertaking the same undergraduate coursework subject. Their majors varied, although approximately half of the Korean participants were English majors. All of them have taken the NCTUA or CSAT to enroll in their university.

Although the perspectives of the university students do not necessarily represent those of actual test-takers (i.e., third-grade high school students who aim to enroll in a national university), this study selected university students because of logistical reasons. This limitation should be acknowledged when generalizing the results of the study.

### Instruments

Booklets containing NCTUA and CSAT test items were developed. These contained 18 items selected from the NCTUA administered in January 2009 and 25 items selected from the CSAT administered in November 2008. These test items ($N = 43$)—including instructions, prompts, and options—were obtained from the official websites of both test committees. The instructions and options provided in either Japanese or Korean were translated into the participant's respective L1 so as not to cause misunderstanding due to the language. The prompts and options originally provided in English were written as they were in the booklet. All the items were randomly arranged in the booklet.

Response sheets were also developed for eliciting the participants' perceptions of the ability measured by each test item. They contained a list of 29 statements, which were published in the test committees' official reports, on the ability measured by each test item (KICE, 2009; NCUEE, 2009). Of these, there were 18 statements on the NCTUA items and 11 statements on the CSAT items (see Additional file 1: Appendices 1 and 2 for the statements on the ability measured by the NCTUA and CSAT). The statements were translated into the L1 of the participants, and those including jargon were paraphrased using general terms. Additionally, original statements that were considered too

general and abstract were divided into two or three separate specific statements. For example, the statement "vocabulary" was divided into "ability to understand the meaning of a word accurately" and "ability to identify the correct vocabulary fit in the context". This modification was made to two statements on the CSAT. Finally, to examine to what degree the participants perceive that the tests measure communicative skills, speaking and writing abilities were added to the statements for the NCTUA, and speaking ability was added to those for the CSAT, although the committee reports did not include these abilities. The response sheets also included item numbers (1 to 37) and boxes to fill out concerning the participants' perception.

### Procedures

The item booklets and response sheets were sent to Japanese and Korean instructors working at the participants' universities. The second author corresponded with them via e-mail to explain the purpose of the research and instruct on the data collection procedures. The instructors distributed the survey forms to the participants during one of their class periods in September and October in 2009.

To begin, the instructors explained the purpose of the survey. After consent was obtained from the participants, the instructors explained how to indicate their perceptions on the response sheet. Each of the participants was given a hard copy of the item booklet and response sheet. They were not instructed to answer the test items but asked to simply read the content and check all the ability statements that they thought each item was measuring. The participants thus gave their responses using their first impression of each test item. They were asked to choose applicable abilities from the 18 NCTUA statements for the NCTUA test items and from 11 CSAT statements for the CSAT test items. There was no time constraint to complete the survey. They were also asked to freely write about their general perceptions of entrance examinations administered in their countries.

After all the participants' responses were collected, the instructor at each university returned the response sheets to the second author. As a consequence, we elicited 80 Japanese students' and 98 Korean students' responses to the NCTUA and CSAT test items.

### Data analysis

We collected Japanese and Korean students' perceptions of both NCTUA and CSAT test items. This study analyzes the Japanese participants' responses to the 18 NCTUA items and the Korean participants' responses to the 25 CSAT items. This is because each of the examinations was developed and administered only for students in that country, who prepare for the examination and are familiar with its test items. The findings were not considered generalizable had we examined the perceptions of students who have never taken the examination and were not familiar with the item types.

The agreement rate between the test committee intention and the participant perception of each item was calculated. Whereas the NUCEE explicitly provides information on the ability that each item aims to measure, the KICE only states the ability that particular item types or formats are designed to measure. Thus, we closely examined the committee reports and identified what each Korean test item is intended to measure.

Each test item included one to four abilities that the test committees intended to measure, and most participants indicated more than two abilities for each item. We analyzed how many participants correctly identified the ability intended to be measured by the committees, and the types of test item likely to cause misperception of the ability intended to measure. The written comments were used to interpret the possible reason for the gap between test-taker perception and test committee intention.

## Results

### Agreement between test-taker perceptions and test committee intentions

First, to answer Research Question 1, we examined the percentages of the participants who judged the item to be measuring the abilities as intended by the test committees. Tables 1 and 2 present how many Japanese and Korean participants accurately recognized the abilities that the NUCEE and KICE intended to measure, respectively. The numbers in the tables correspond to those in Additional file 1.

Table 1 shows the degree to which the Japanese participants recognize the test committee intentions. The mean percentage was 71.8 %, and the range was from 22.5 % (Item 34) to 90.0 % (Items 2 and 12). Item 34 induced the largest gap between the test committee intention and test-taker perception. Although the NUCEE intended to measure the ability to understand the intended meaning that is not explicitly stated, only 22.5 % of the participants perceived that the item measures that ability.

Table 2 shows the degree to which the Korean participants recognize the test committee intentions. The mean percentage was 59.1 %, and the range was from 5.1 % (Item 17) to 90.8 % (Items 32). In contrast to the Japanese case, several items induced a large gap between the test committee intention and test-taker perception. For example, less than 10 % of the participants accurately recognized the abilities measured by three items: Items 17 (5.1 %), 33 (7.1 %), and 37A (6.1 %). These items were all intended to measure the writing ability of test-takers. Another item measuring this ability, Item 20, also indicated a low agreement rate (11.2 %). In addition, less than a quarter of the Korean participants accurately recognized one of the abilities measured by Items 7 and 9. These items lowered the overall agreement rate in the Korean data.

It should be noted that one factor for lowering the agreement rates was that several ability statements had similar meaning and overlapped in content. Thus, the participants possibly could not distinguish clearly between some of the statements. For

**Table 1** Agreement rates: Japanese data

| Item # | 1 | 2 | 3 | 4 | 5 | 5 | 10A | 10B | 10C |
|---|---|---|---|---|---|---|---|---|---|
| Intention | ① | ② | ③ | ④ | ⑤ | ⑦ | ⑥ | ⑥ | ⑥ |
| Rate | 78.6 | 90.0 | 85.0 | 73.8 | 68.8 | 71.3 | 52.5 | 62.5 | 62.5 |
| Item # | 12 | 13 | 16 | 18 | 21 | 24 | 26 | 28 | 29 |
| Intention | ⑦ | ⑧ | ⑨ | ⑩ | ⑪ | ⑫ | ⑬ | ⑬ | ⑬ |
| Rate | 90.0 | 70.0 | 65.0 | 70.0 | 87.5 | 68.8 | 86.3 | 81.3 | 80.0 |
| Item # | 34 | 34 | 34 | Mean | | | | | |
| Intention | ⑭ | ⑮ | ⑯ | | | | | | |
| Rate | 75.0 | 22.5 | 66.3 | 71.8 | | | | | |

Intention = the ability that the NUCEE intended to measure (the numbers corresponding to those in Additional file 1).
Rate = percentage of the Japanese participants who judged the item to be measuring the abilities as intended by the NUCEE

**Table 2** Agreement rates: Korean data

| Item # | 6 | 6 | 7 | 7 | 7 | 8 | 8 | 9 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| Intention | ⑦ | ⑨ | ⑦ | ⑧ | ⑨ | ① | ② | ① | ⑧ |
| Rate | 61.2 | 83.7 | 66.3 | 12.2 | 82.7 | 71.4 | 88.8 | 24.5 | 60.2 |
| Item # | 11 | 11 | 14 | 14 | 15 | 15 | 17 | 19 | 19 |
| Intention | ③ | ④ | ③ | ④ | ③ | ⑤ | ⑩ | ③ | ⑤ |
| Rate | 60.2 | 46.9 | 69.4 | 45.9 | 66.3 | 88.8 | 5.1 | 80.6 | 49.0 |
| Item # | 20 | 22 | 23 | 23 | 25 | 27 | 30 | 30 | 31 |
| Intention | ⑩ | ③ | ③ | ④ | ③ | ③ | ③ | ④ | ③ |
| Rate | 11.2 | 77.6 | 76.5 | 51.0 | 84.7 | 35.7 | 82.7 | 70.4 | 84.7 |
| Item # | 31 | 32 | 32 | 33 | 35 | 35 | 36A | 36A | 36B |
| Intention | ④ | ③ | ⑤ | ⑩ | ③ | ④ | ② | ③ | ① |
| Rate | 64.3 | 70.4 | 90.8 | 7.1 | 80.6 | 43.9 | 52.0 | 37.8 | 48.0 |
| Item # | 36B | 36B | 37A | 37B | 37C | 37C | Mean | | |
| Intention | ② | ③ | ⑩ | ③ | ③ | ⑤ | | | |
| Rate | 76.5 | 40.8 | 6.1 | 84.7 | 68.4 | 71.4 | 59.1 | | |

Intention = the ability that the KICE intended to measure (the numbers corresponding to those in Additional file 1).
Rate = percentage of the Korean participants who judged the item to be measuring the abilities as intended by the KICE

example, Statements ⑩, ⑭, and ⑯ provided by the Japanese test committee (see Additional file 1) are quite similar and overlapping in the sense that the ability to identify the main point is involved in all of the statements. Similarly, Statements ③, ④, ⑤, and ⑥ provided by the Korean test committee (Additional file 1) are all related to comprehension of the text. This suggests that the participants' perceptions might not be necessarily as distant from the test committees' intention as indicated by the results, even though their responses were not exactly the same as the intended ability.

Large discrepancies between the test committees and the participants were expected, given that empirical studies to date have confirmed that even language experts frequently disagree with each other about the abilities measured by test items (e.g., Alderson & Kremmel, 2013). However, excluding a small number of items, the participants seemed to moderately agree with the test committees and understand their intentions correctly. At least, they perceived that the ability the test committee intended to measure was one of the abilities measured by the item. One possible reason for this is that questions themselves are simple and straightforward since the ability measured is explicitly stated in the question statement. For example, the question statement of Item 2—an item measuring an ability to identify the accented syllable of a word—was "Choose the word whose primary accent is placed the same as the word given below". In such cases, the participants seemed to accurately understand the test developers' intentions without any great difficulty.

#### Items inducing gaps between test-taker perceptions and test committee intentions

Although the participants moderately agreed with the test committee about the abilities measured by items, some test items induced large gaps between these perceptions. To answer Research Question 2, this section closely examines the types of test item that induced large gaps. We focused on two types of items: items designed to measure the ability to read between the lines and writing ability.

An item that induced a large gap between test committee intentions and test-taker perceptions was Item 34 in the NUCEE (Fig. 1). Test-takers were required to read a passage on monolingual and bilingual dictionaries (eight paragraphs totaling approximately 780 words) and choose an option suitable to fill each blank (the passage and options are omitted due to limitations of space). According to the test committee, the item aimed to measure (a) the ability to understand the procedure, outline, and main point of the story, (b) the ability to understand the intended meaning that is not explicitly stated, and (c) grasp the main point of the paragraph. Although 75.0 % and 66.3 % of the Japanese participants perceived that the item measures the first and third abilities respectively, only 22.5 % of them felt that the second ability was an ability measured by Item 34.

It can be considered that, among the seven questions in the item, Questions 3 and 7 particularly required test-takers to comprehend the intended meaning that is not explicitly stated. Question 3 asked test-takers to identify the appropriate example describing "a Japanese 'equivalent' can never give you the real meaning of a word in English," which cannot be found in the text. Similarly, Question 7 required test-takers to grasp what the writer implies in the last paragraph. The answers for the other questions were found in the passage. One possible reason for the low percentage of recognition of ⑮ (the ability to understand the intended meaning that is not explicitly stated) is that the participants did not actually answer the questions but only indicated their perceptions by skimming the item and using their first impression. They might have missed or did not understand the word *implies* in Question 7. In other words, they did not closely examine what Questions 3 and 7 were asking and simply assumed that Item 34 measures reading comprehension as a whole.

Another possible explanation is that test-takers might perceive that reading always involves an understanding of the literal meaning of texts rather than an understanding of inferred meaning. A similar tendency was found in the Korean participants' perception of several CSAT items. Table 3 shows the participants' responses to six items designed to measure both (a) the ability to grasp the content quickly and objectively and (b) the

**Read the passage below and choose an appropriate option from ① to ④ to fill in the blanks.**

A passage on monolingual and bilingual dictionaries (8 paragraphs, approx. 780 words)

1 When the writer received the dictionary from his aunt, he did not find it easy to use because _____.
2 The type of dictionary described in paragraph (3) is different from those explained in paragraphs (2) and (4) in that it _____.
3 Which of the following examples best fits the aunt's view that "a Japanese 'equivalent' can never give you the real meaning of a word in English"? _____
4 By using the type of monolingual dictionary described in paragraph (6), the writer _____.
5 Through using a "find-the-right-word" dictionary, one can _____.
6 When paragraphs (1) through (8) are divided into four groups according to the topic of each, which grouping is most appropriate? _____
7 The writer implies that _____.

**Fig. 1** Item 34 from the NUCEE

**Table 3** Percentages of the Korean participants who chose ③ and ④ on six items

| Item # | Item type | ③ | ④ |
|---|---|---|---|
| 11 | Complete a topic sentence in the passage | 60.2 % | 46.9 % |
| 14 | Identify the referent of this (the main theme) | 69.4 % | 45.9 % |
| 23 | Identify the main point of the passage | 76.5 % | 51.0 % |
| 30 | Identify the claim of the passage | 82.7 % | 70.4 % |
| 31 | Identify the title of the passage | 84.7 % | 64.3 % |
| 35 | Identify the purpose of the passage | 80.6 % | 43.9 % |

③ = Ability to grasp the content quickly and objectively. ④ = Ability to understand the content and to infer the implicit content, including the main point, title, argument, and expression intentionally eliminated

ability to understand the content and to infer the implicit content, including the main point, title, argument, and expression intentionally eliminated. These items commonly required test-takers to read one paragraph (100 to 150 words) and infer the main theme that is not explicitly stated in the paragraph. Nevertheless, as Table 3 shows, more participants chose the former than the latter ability. They seemed to perceive that the items measure the ability to understand the literal meaning of the passage more strongly than the ability to infer the main theme that is not directly stated in the passage.

The data suggest that the participants tended to perceive that reading items typically measure the ability to comprehend what is written in the passage even though the ability to make an inference from the passage is also required. Comprehension of the literal meaning (reading the lines) was possibly deemed a prerequisite for comprehension of the unstated meaning (reading between the lines). This view is not completely naïve since even language teachers often prescribe this approach (Alderson, 2000). For the participants, therefore, reading the lines might be the fundamental ability or skill needed in order to solve the reading items which demand an additional skill. Alternatively, they might have thought that understanding of the literal meaning would suffice to solve the items. The participants' perception that these items measure the ability to comprehend the literal meaning is not considered distant from the test committees' intentions since reading the lines and reading between the lines are hard to distinguish clearly (Alderson, 2000). In this sense, the gap between the test committee intentions and participants' perception may not be crucial because students are likely to engage in reading to prepare for these reading items. However, if test developers wish to draw students' attention to various specific subskills of reading (Brown & Abeywickrama, 2010; Hubley, 2012; Hughes, 2003), reading items may not necessarily have a direct impact on students' attempt to acquire them.

Few Korean participants perceived that Items 17, 20, 33, and 37A measured writing ability (5.1 %, 11.2 %, 7.1 %, and 6.1 %, respectively). Instead, they tended to consider that these items measure (a) the ability to grasp the content quickly and objectively (③) and/or (b) the ability to understand the connection among sentences (⑧). Table 4 shows the ability statements most frequently chosen for these items.

According to the Korean test committee, items measuring *writing ability* include tasks to place given sentences and paragraphs in the appropriate order, tasks to summarize a paragraph in a single sentence, and tasks to place a given sentence in the appropriate place in the passage to suit the context (KICE, 2009). For example, Figs. 2 and 3 present Items 17 and 33, respectively. The former requires test-takers to reorder

**Table 4** Most frequently chosen ability statements for writing items

| Item # | Most frequent choice | | 2nd most frequent choice | |
|---|---|---|---|---|
| 17 | ③ | 69.4 % | ⑧ | 66.3 % |
| 20 | ③ | 67.4 % | ② | 52.0 % |
| 33 | ⑧ | 80.6 % | ③ | 55.1 % |
| 37A | ⑧ | 67.4 % | ③ | 65.3 % |

② = Ability to identify the correct vocabulary fit in the context. ③ = Ability to grasp the content quickly and objectively. ⑧ = Ability to understand the connection among sentences

groups of sentences (two to three sentences) in a coherent order, and the latter requires them to insert a sentence into an appropriate place in a paragraph consisting of six sentences.

The Korean test committee states that these items are designed to measure test-takers' writing ability *indirectly* (KICE, 2009). In general, indirect testing aims to make an inference about test-takers' language performance through artificial language tasks rather than tasks directly reflecting target language use (Davies et al., 1999). It is reasonable to consider that these items tap into part of the knowledge, abilities, and skills of writing. For example, Items 17 and 33 seem to primarily assess textual or discourse knowledge that includes cohesion of texts and rhetorical organization (Bachman & Palmer, 1996, 2010; Grabe & Kaplan, 1996). However, such an intention was not accurately recognized by the participants. Many of them perceived that these items measure textual or discourse knowledge of *reading* instead of writing. This may be because the passage is long or because the participants do not have to present their written products.

For large-scale high-stakes testing, indirect tests of writing have been widely used in Japan and Korea because of their high rater-reliability and practicality. Weigle (2002) argues, "Indirect tests of writing represented the domination of the agenda of testing firms and their clients, who wanted fast, reliable, and inexpensive ways of sorting students according to the status quo of existing social patterns" (p. 239). These indirect test items may successfully measure a component of test-takers' writing ability quickly and rather objectively. However, such items are not likely to be realized by test-takers

**Place sentences (i), (ii), (iii) in the most appropriate order to come after the following sentences. Choose the one best answer.**

Now many kinds of superior coffee beans are being decaffeinated in ways that conserve strong flavor. But the public suffers from a groundless fear of chemical decaffeination and prefers instead to buy water-processed decaf.

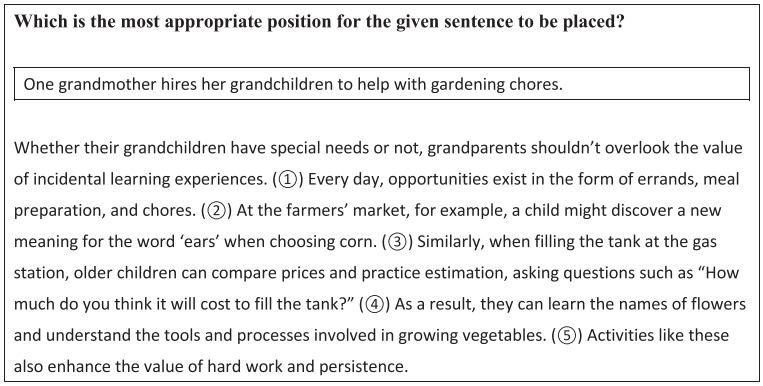
i The solvent comes into direct contact with them, carrying the caffeine with it. The drained solvent is then mixed with water, and the caffeine is drawn out to be sold.

ii In the water process, however, no solvent touches the beans. After the beans are steamed, they are soaked in water, which removes the caffeine along with all the soluble solids in the beans. The solution is drained off to a separate tank, where the caffeine is drawn out from it.

iii Every process of decaffeination, whether chemical- or water-based, starts with steaming the green beans to loosen the bonds of caffeine. In the chemical process, a solvent circulates through the beans.

① ( i )-( ii )-( iii ) ② ( ii )-( i )-( iii ) ③ ( ii )-( iii )-( i ) ④ ( iii )-( i )-( ii ) ⑤ ( iii )-( ii )-( i )

**Fig. 2** Item 17 from the CSAT


Which is the most appropriate position for the given sentence to be placed?

One grandmother hires her grandchildren to help with gardening chores.

Whether their grandchildren have special needs or not, grandparents shouldn't overlook the value of incidental learning experiences. (①) Every day, opportunities exist in the form of errands, meal preparation, and chores. (②) At the farmers' market, for example, a child might discover a new meaning for the word 'ears' when choosing corn. (③) Similarly, when filling the tank at the gas station, older children can compare prices and practice estimation, asking questions such as "How much do you think it will cost to fill the tank?" (④) As a result, they can learn the names of flowers and understand the tools and processes involved in growing vegetables. (⑤) Activities like these also enhance the value of hard work and persistence.

**Fig. 3** Item 33 from the CSAT

as those tapping into writing ability. The participants' written comments on entrance examinations in general reflected this perception. Some participants stated that examinations cannot measure performance skills (including writing and speaking) and focus solely on measuring reading skills:

> Most English tests only measure reading ability, and are unable to assess writing ability. They only assess how well test takers can analyze written passages. (Korean participant)

> There are some test items intended to measure speaking ability or writing ability, but I think they cannot sufficiently measure abilities for speaking and writing. (Korean participant)

> I believe that we need performance tests of writing and speaking. (Korean participant)

> I don't think multiple choice tests can sufficiently capture test takers' writing ability. (Japanese participant)

This finding suggests that indirect tests are less likely to be perceived as measuring part of productive performance ability and thus result in poor face validity. As a consequence of the low face validity, indirect testing of writing may influence students' efforts to improve reading ability rather than writing ability.

## Conclusions

This study examined test-taker perception of the ability being measured by high-stakes examinations and investigated the agreement between test-taker perception and test developer intention. University students in Japan and Korea provided their ideas on what each item measures in entrance examinations in Japan and Korea. It was found that the overall agreement was 71.8 % (Japanese data) and 59.1 % (Korean data), respectively. Some items generated disagreement and undermine the face validity of the tests. First, items aiming to measure an ability to read between the lines were perceived to be measuring an ability to understand the content objectively. Second, many participants perceived items created to measure writing skills as tapping into reading skills.

There are several limitations to be acknowledged in this study. First, the participants were university students who had passed entrance examinations (i.e., high score achievers

in general), rather than high school students who were preparing for the examinations. Thus, the sample of the population was limited, and the results might not be readily generalizable to the perception of real test-takers of the examinations. Second, the participants were not asked to answer the test items but instead indicated their perceptions using their first impression. This is possibly limiting because they might not have carefully read the questions and considered what the test items seemed to measure. As stated above, this could be a possible cause of the low percentage of some items measuring reading skills (e.g., Item 34). Third, some ability statements were highly overlapping and might not have been clearly distinguished by the participants, as pointed out previously. Additionally, some statements were too general (e.g., speaking and writing abilities).

Although there are limitations as stated above, we have argued that some types of test items potentially induce a gap between test committee intentions and test-taker perceptions, which does not necessarily result in the intended washback on students' learning. There are two practical implications for maximizing the test committees' intended washback effect. First, it is necessary to explicitly and widely inform test-takers and language teachers of the test construct and the specific abilities that the test items intend to measure. Test developers should prepare reader-friendly materials related to test construct for test-takers, or the subskills intended to be measured could be written in the test paper or included in question statements. At the least, teachers have to realize the specific subskills that the test intends to measure and should raise students' awareness of them. As an essential component of *teacher assessment literacy*—an understanding of "how to use assessment to maximize student motivation and learning" (Coombe, Troudi, & Al-Hamly, 2012, p. 25)—language teachers should realize the importance of conveying test developers' intention to students and raise their awareness of the abilities being measured throughout the test.

Second, direct assessment of performance is preferable to indirect assessment if test developers' intended washback includes students' attempts to improve oral or writing performance. Even when test developers' intention is accurately conveyed, intended washback on student learning might not be obtained if they are not convinced by the given information. The students' general perceptions of paper-pencil entrance examination suggest that they are less likely to perceive that multiple-choice type items tap into productive performance skills even though they receive the correct information about the items. In this sense, directly assessing writing is necessary in order to enhance intended washback. In fact, past literature supports the idea that performance testing can induce positive washback because students are likely to engage with performance to prepare themselves for the test (Kane, Crooks, & Cohen, 1999).

This study has argued that test-taker perception of test construct affects the content of student learning, and intended washback may not be successfully achieved if there is a gap between test-taker perception and the test committee's intention. Although this seems logical and plausible, the relationship between test-taker perception and their actual learning has not yet been empirically confirmed. In other words, this claim should be regarded as a hypothesis on washback, which should be researched in order to support or reject it, as Alderson and Wall (1993) argue with regard to various hypotheses of washback. Meanwhile, test-taker perception of what the test measures, or face validity, needs to be taken seriously as a potential factor affecting the content of learning and investigated further in a range of contexts.

## Additional file

**Additional file 1: Appendix 1.** The statements on the ability measured by the NCTUA. **Appendix 2.** The statements on the ability measured by the CSAT. (DOCX 15 kb)

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

#### Authors' contributions
Both authors read and approved the final manuscript.

Received: 8 April 2015 Accepted: 15 July 2015
Published online: 24 July 2015

#### References
Akbari, R. (2012). Validity in language testing. In C Coombe, P Davidson, B O'Sullivan, & S Stoynoff (Eds.), *The Cambridge guide to second language assessment* (pp. 30–36). Cambridge: Cambridge University Press.
Alderson, CJ. (2000). *Assessing reading*. Cambridge: Cambridge University Press.
Alderson, CJ, Clapham, C, & Wall, D. (1995). *Language test construction and evaluation*. Cambridge: Cambridge University Press.
Alderson, CJ, & Kremmel, B. (2013). Re-examining content validation of a grammar test: The (im)possibility of distinguishing vocabulary and structural knowledge. *Language Testing, 30*(4), 535–556.
Alderson, CJ, & Wall, D. (1993). Does washback exit? *Applied Linguistics, 14*(2), 115–129.
Bachman, LF. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.
Bachman, LF, & Palmer, AS. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford: Oxford University Press.
Bachman, LF, & Palmer, AS. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford: Oxford University Press.
Brown, A. (1993). The role of test-taker feedback in the test development process: Test-takers' reactions to a tape-mediated test of proficiency in spoken Japanese. *Language Testing, 10*, 277–301.
Brown, HD, & Abeywickrama, P. (2010). *Language assessment: Principles and classroom practices* (2nd ed.). White Plains, NY: Pearson Education.
Cheng, L. (2008). Washback, impact and consequences. In E Shohamy & NH Hornberger (Eds.), *Language testing and assessment* (2nd ed., pp. 349–364). New York: Springer Science + Business Media LLC.
Coombe, C, Troudi, S, & Al-Hamly, M. (2012). Foreign and second language teacher assessment literacy: Issues, challenges, and recommendations. In C Coombe, P Davidson, B O'Sullivan, & S Stoynoff (Eds.), *The Cambridge guide to second language assessment* (pp. 20–29). New York: Cambridge University Press.
Davies, A, Brown, A, Elder, C, Hill, K, Lumley, T, & McNamara, T. (1999). *Dictionary of language testing*. Cambridge: Cambridge University Press.
Eiken Foundation of Japan. (n.d.). Kaihatsusya no koe [The voice of the developers], from http://www.eiken.or.jp/teap/group/voice.html.
Ferman, I. (2004). The washback of an EFL national oral matriculation test to teaching and learning. In L Cheng, Y Watanabe, & A Curtis (Eds.), *Washback in language testing: Research contexts and methods* (pp. 191–210). Mahwah, NJ: Lawrence Erlbaum Associates.
Fulcher, G, & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. London: Routledge.
Grabe, W, & Kaplan, RB. (1996). *Theory and practice of writing: An applied linguistic perspective*. Essex: Addison Wesley Longman Limited.
Hawkey, R. (2006). *Impact theory and practice: Studies of the IELTS test and Progetto Lingue 2000*. Cambridge: Cambridge University Press.
Hubley, NJ. (2012). Assessing reading. In C Coombe, P Davidson, B O'Sullivan, & S Stoynoff (Eds.), *The Cambridge Guide to second language assessment* (pp. 211–217). New York: Cambridge University Press.
Hughes, A. (2003). *Testing for language teachers* (2nd ed.). Cambridge: Cambridge University Press.
Kane, MT. (2006). Validation. In RL Brennan (Ed.), *Educational Measurement* (4th ed., pp. 17–64). Westport, CT: Greenwood Publishing.
Kane, MT, Crooks, T, & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice, 18*(2), 5–17.
KICE. (2009). The College Scholastic Ability Test 2009: Media report, from http://www.kice.re.kr/ko/board/view.do?article_id=73465&menu_id=10087.
KICE. (2012). College Scholatic Ability Test, from http://kice.re.kr/en/contents.do?contentsNo=149&menuNo=405.
Lee, W. (2012). The speaking test of Korea's NEATs 2 & 3: Its usability for university admission qualifications. *Japan Language Testing Association Journal, 15*, 27–42.
Menken, K. (2008). High-stakes tests as de facto language education policies. In E Shohamy & NH Hornberger (Eds.), *Language testing and assessment* (2nd ed., pp. 401–413). New York: Springer Science + Business Media LLC.
Messick, S. (1989). Validity. In RL Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: American Council on Education and Macmillan.
NCUEE. (2009). The examination report of the National Center of Test for University Admission 2009, from http://www.dnc.ac.jp/old_data/exam_repo/21/gaikokugo.html.
NCUEE. (n.d.). The structure and administration of the National Center Examination, from http://www.dnc.ac.jp/center/shiken_gaiyou/.

Qi, L. (2004). Has a high-stakes test produced the intended changes? In L Cheng, Y Watanabe, & A Curtis (Eds.), *Washback in language testing: Research contexts and methods* (pp. 171–190). Mahwah, NJ: Lawrence Erlbaum Associates.
Shohamy, E. (2001). *The power of tests: A critical perspective on the uses of language tests*. Harlow: Pearson Education Limited.
So, Y. (2014). Are teacher perspectives useful? Incorporating EFL teacher feedback in the development of a large-scale international English test. *Language Assessment Quarterly, 11*(3), 283–303.
Stevenson, DK. (1985). Authenticity, validity and a tea party. *Language Testing, 2*(1), 41–47.
Wall, D. (1996). Introducing new tests into traditional systems: Insights from general education and from innovation theory. *Language Testing, 13*(3), 334–354.
Wall, D. (2005). *The impact of high-stakes examinations on classroom teaching: A case study using insights from testing and innovation theory*. Cambridge: Cambridge University Press.
Wall, D. (2012). Washback. In G Fulcher & F Davidson (Eds.), *The Routledge handbook of language testing* (pp. 79–92). Oxon: Routledge.
Wall, D, & Alderson, CJ. (1993). Examining washback: The Sri Lankan impact study. *Language Testing, 10*(1), 41–69.
Watanabe, Y. (1996). Does grammar translation come from the entrance examination? Preliminary findings from classroom-based research. *Language Testing, 13*, 318–333.
Watanabe, Y. (2004). Teacher factors mediating washback. In L Cheng, Y Watanabe, & A Curtis (Eds.), *Washback in language testing: Research contexts and methods* (pp. 129–146). Mahwah, NJ: Lawrence Erlbaum Associates.
Watanabe, Y. (2013). The National Center Test for University Admissions. *Language Testing, 30*(4), 565–573.
Weigle, SC. (2002). *Assessing writing*. Cambridge: Cambridge University Press.