Language Testing in Asia

RESEARCH Open Access

Investigating the factor structure of the Test of English for Academic Purposes (TEAP) and its relation to test takers' perceived test task value

Keita Nakamura*

*Correspondence:
ke-nakamura@akane.waseda.jp

Waseda University, Tokyo, Japan

Abstract

Background: This study investigated the scoring and criterion-related validity of the TEAP, a newly developed Test of English for Academic Purposes. In this study, scoring validity was examined by investigating the factor structure, while criterion-related validity was examined by first investigating the longitudinal change of test takers' perceived test task value toward the measured construct and then investigating the relationship of test takers' perceived value to the factor structure of the TEAP.

Methods: Confirmatory item-level factor analysis was conducted using the data obtained from 2217 first-year university students comparing four models (unitary, correlated, receptive-productive, higher order). Additional confirmatory factor analyses were conducted to first investigate the longitudinal change of perceived value toward the measured construct and then to investigate the relationship of test takers' perceived value of the construct measured by the test to the factor structure of the TEAP.

Results: The results show that the higher-order model was the best-fitting model. This confirmed a previous small-scale study suggesting the generalizability of the test's factor structure. Furthermore, it was found that test takers' perceived values measured at the start of university positively affected the values measured about 6 months later.

In addition, perceived values measured both at the start of university study and about 6 months later positively correlated with the higher-order factor of the test.

Conclusions: The results provide further support of the scoring validity of the test. In addition, a positive relationship between the higher-order factor of TEAP and the factors of perceived values provide evidence of the usefulness of test takers' perception to further support the criterion-related validity of the test.

Keywords: Factor structure, Scoring validity, Criterion-related validity, Longitudinal change of perceived value

Introduction

Educational reform efforts are observed around the world (Chalhoub-Deville, 2016) including Japan's plan to reform its English education (Allen, 2020; Kuramoto & Koizumi, 2016; Sasaki, 2008). In 2013, the Ministry of Education, Culture, Sports, Science and Technology (MEXT) launched a policy designed to improve English education

in Japan and revise the university entrance examination system. The first point of the revision concerns the balanced teaching and learning of the four English skills: reading, writing, listening, and speaking. High school graduates are expected to achieve a level of English competency equivalent to the B1/B2 levels of the Common European Framework of References for Languages (CEFR; Council of Europe, 2001), which defines the six common reference levels (A1, A2, B1, B2, C1, C2), while junior high school students are expected to achieve the A1/A2 levels.

The second point of the present reform in Japan concerns the revision of in-house English language college entrance examinations to measure the four English skills. In Japan, most universities develop and administer their own in-house entrance examinations for admission purposes annually, and the revision came from the concern that those in-house tests mostly use multiple-choice test formats focusing on receptive aspects of English skills such as reading and listening. Under the new policy, universities are expected to develop and administer the four skills through in-house examinations, but they are also allowed to utilize externally developed standardized English tests of the four skills in addition to in-house examinations. Thus, under the new plan, students who wish to be matriculated to universities could choose English tests and take them up to two times within the April–December period for admission purposes in addition to the in-house tests.

The third point concerns the revision of the National Center Test (NCT), which was first introduced in 1990 and has been taken by over 500,000 students (National Center for University Entrance Examinations, 2017) who wish to be matriculated in mainly national or public universities. The revision came from the concern that the NCT only measures written and listening skills of English by a multiple-choice test format, while the new policy required well-balanced teaching and learning of the four skills of English. As of this writing, a newly revised NCT, now called the Common Test (CT), was launched and is expected to be continued until 2023 (Ministry of Education, Culture, Sports, Science and Technology, 2020).

This study concerns the second point of revision or the plan to use standardized English tests for college entrance purposes. In the original plan, MEXT selected the authorized tests based on criteria such as the appropriate link between the test content and the national curriculum standards and the CEFR. Such tests include the Test of English as a Foreign Language Internet-based Test (TOEFL iBT®; Educational Testing Service, 2022), International English Language Testing System (IELTS; British Council, IDP, IELTS Australia, and Cambridge English Language Assessment, 2022), Cambridge English exams (University of Cambridge Local Examinations Syndicate, 2022), the Test in Practical English Proficiency (EIKEN; Eiken Foundation of Japan, 2022a, 2022b), Global Test of English Communication (GTEC; Benesse, 2022), and the Test of English for Academic Purposes (TEAP; Eiken Foundation of Japan, 2022a, 2022b). Though this point of revision was expected to start in 2020, the plan was officially postponed in 2019 due to various fairness-issue reasons such as the issue of test fee, test center locations, and the number of test administrations (Allen, 2020). Yet, even after the announcement, universities still have the freedom to choose standardized tests to be used for admission purposes so that students can choose and take tests in addition to the option of taking in-house tests.

Test of English for Academic Purposes (TEAP)

The TEAP test was developed by the Eiken Foundation of Japan together with Sophia University, a private university in Tokyo known for its English language program, and the Center for Research in English Language Learning and Assessment (CRELLA) of the University of Bedfordshire in the UK. There have been many standardized English tests which measure the four skills of English such as TOEFL iBT, IELTS, and EIKEN, but the TEAP was designed specifically for college admission purposes in Japan. Some desirable features of the TEAP include the appropriate task difficulty for university applicants in Japan, or CEFR A2-B2 (MEXT, 20120, and the link to the national course of study, while most of other tests have original test purposes and designs other than Japanese college admission. In order to collect evidence to support the use of the TEAP for Japanese college entrance purpose, the TEAP has gone through a series of validation studies conducted in collaboration with CRELLA and reported on the website of the Eiken Foundation of Japan, including studies of contextual and cognitive aspects of the reading and listening sections (Taylor, 2014), the writing section (Weir, 2014), speaking section (Nakatsuhara, 2014; Nakatsuhara et al., 2014), stakeholders' perceptions of university entrance examinations, and the expected washback from the introduction of the TEAP (Green, 2014; Nakamura, 2014), and factor structure (In'nami et al., 2016).

Dimensionality of foreign language ability

In the socio-cognitive framework (O'Sullivan & Weir, 2011), scoring validity is defined as "to what extent can we depend on the scores on the test? What do the numbers or grades mean?" Large-scale EFL tests which measure multiple skills (e.g., reading, listening, writing, and speaking) have been the target of investigation in the previous research (Gu, 2015; In'nami & Koizumi, 2011; In'nami et al., 2016; Sasaki, 1993; Sawaki et al., 2009), yet other studies have investigated the validity of single-skill assessments including Bachman and Palmer (1981), which investigated the scoring validity of FSI (Foreign Service Institute) speaking assessment. The rationale behind those studies has been not only to validate the scoring procedure but also to elucidate the fundamental issue related to the dimensionality of a language ability (Bachman & Palmer, 1981; Gu, 2015; In'nami et al., 2016; Oller Jr., 1980; Sawaki et al., 2009). This aspect of validity is important because the actual score reporting policy (e.g., computing an overall score from individual skill scores) must be supported by the psychometric properties of the target assessment.

Oller Jr. (1980) once argued that one's language proficiency, like one's intelligence, can be explained by the existence of a unitary factor from the studies which showed the strong correlations among various language tasks. However, later studies (Bachman & Palmer, 1981; Gu, 2015; In'nami et al., 2016; Sawaki et al., 2009) have found the plausibility of multi-componential factor structure to explain one's language proficiency supporting the validity of producing a separate score for each skill and a composite score.

Sawaki et al. (2009) investigated the factor structure of the TOEFL iBT using CFA and found that the higher-order factor model was the best-fitting model ($CFI = 0.98$, $RMSEA = .022$ [.021, .022]). In'nami and Koizumi (2011) investigated the factor structure of the TOEIC listening and reading test using CFA and found that the correlated factor model, which hypothesized reading and listening as two correlated factors, was

the best-fitting model ($CFI = 0.972$, $RMSEA = .065$ [.042, .088]). Due to the fact that the higher-order and correlated models cannot be statistically distinguished when the number of correlated first-order factors is three or less, the researchers only evaluated the reading and listening components; it was difficult to compare the correlated model with more complex models, including the higher-order factor model. However, the authors argued that the high correlation ($r = 0.87$) between the reading and listening factors suggested the existence of a higher-order factor.

Following the previous studies on the factor structure of standardized English tests, investigating the factor structure of TEAP, which is the focus of this study, was an essential aspect of validation. In'nami et al. (2016) hypothesized four competing models (i.e., a unitary model, a correlated factor model, a higher-order model, and a receptive and productive model). The single-factor model or a unitary model hypothesizes that all items from all four skills load on one factor. The correlated four-factor model hypothesizes each item can be explained by a skill-specific factor that is correlated with the other skill-specific factors. The higher-order factor model hypothesizes the presence of a general higher-order factor with four lower-order factors corresponding to the assessed four language skills. The receptive-productive model hypothesizes the presence of two factors, a receptive factor and a productive factor under. The former corresponds to the receptive skills (i.e., reading and listening), and the latter corresponds to the productive skills (i.e., writing and speaking). The authors confirmed the existence of a higher-order structure of EFL test by investigating the factor structure of the TEAP and its relationship to the TOEFL iBT based on data gathered from 100 college students in order to validate the proposed use of scores derived from the TEAP for college admission purposes as well. The researchers found that a structural model that hypothesized the existence of a higher-order factor, which controls the four first-order factors, fits the data best ($CFI = 0.932$, $RMSEA = .014$ [.000, .022]) when compared to a unitary model, a correlated factor model, and a receptive and productive model. They argued that the results supported computing and reporting a score for each of the four English skills of reading, listening, writing, and speaking on the TEAP test and computing a composite score for admission purposes.

A major part of previous studies on large-scale tests (e.g., the TOEFL iBT and the TEAP) suggest the existence of a higher-order general English proficiency factor under which skill-based factors of reading, listening, writing, and speaking are located (Gu, 2015; In'nami et al., 2016; Sawaki et al., 2009).

Test-taker perception

In the socio-cognitive framework (O'Sullivan & Weir, 2011), criterion-related validity is defined as "What external evidence is there outside the test scores themselves that the test is doing a good job?" As this aspect of validity focuses on the extent to which test scores reflect a suitable externally measured variable of performance or demonstration of the similar abilities as are included in the test, the appropriateness of the selected variable determines the result of investigation. Previous studies on large-scale tests have used variables such as scores from other tests which measure similar construct (e.g., ETS, 2010), students' evaluation of classroom activities in term of importance (e.g., Sawaki et al., 2009), or students' evaluation of their own English proficiency

(e.g., Powers & Powers, 2015; Runnels, 2016). Those chosen variables together with the target test scores have shown the criterion-related aspect of validity of the target test. One of important variables when considering the criterion-related validity (O'Sullivan & Weir, 2011) of a test is learner beliefs or students' perceptions of the target language or test because what language learners believe affects how they engage with daily learning activity or practicing for target test (Dornyei, 2005). This variable is especially important for this study context where the new policy of English education (MEXT, 2015) aims to balanced teaching and learning of the four skills of English in the classroom. Dornyei (2005) argued that learner beliefs greatly affect behavior when a learner believes in a particular method of learning might reject another. Previous studies (Horwitz, 1988; Eccles & Wigfield, 2002; Sawaki et al., 2009; Xie & Andrews, 2013; Xie, 2015) on learner beliefs have posited various constructs regarding learners' beliefs on language learning. Horwitz (1988) defined learner beliefs as learners' opinions on a variety of issues and controversies related to language learning and developed the Beliefs about Language Learning Inventory (BALLI). BALLI was invented to validate the existence of learner beliefs and their impact on language learning with a scale of five types of learner beliefs as (a) difficulty of language learning, (b) foreign language aptitude, (c) the nature of language learning, (d) learning and communication strategies, and (e) motivation and expectations. In the field of psychology, Eccles and Wigfield (2002) posited an expectancy-value model in which learners' achievement performance, persistence, and choice are affected by their expectancy-related (e.g., "Can I do it?") and task-value ("Do I want to do it?") beliefs. In the field of language testing, based on the expectancy-value model, Xie and Andrews (2013) defined expectation and value as learners' beliefs about how well they will do on upcoming tasks and how much they value the upcoming tasks as desirable. In their study (Xie & Andrews, 2013), which was conducted to 872 Chinese test takers of College English Test (CET), participants responded to a questionnaire asking their perceived expectations (e.g., "If I prepare for it in appropriate ways, I believe I will pass CET4.") and values (e.g., "In order to answer questions correctly, I must understand the key points in reading.") toward the target test together with their test preparation activities. Based on SEM, the authors found a significant effect of learners' perceived test values on the actual test preparation activity confirming the positive impact of language learners' beliefs on their language learning practices before the test. Sawaki et al. (2009) examined the criterion-related validity of the Test of English as a Foreign Language Internet-based Test (TOEFL iBT) listening section by examining its relationship to a criterion measure designed to reflect language-use tasks that university students encounter in everyday academic life. The design of the criterion measure was based on students' responses to a survey on the frequency (i.e., how often learners engage in a task) and importance (i.e., how important a task is for learners to performance well in the class) of various classroom tasks that require academic listening. The authors found significant positive correlation between the listening section score of the TOEFL iBT test and the students' responses to the survey ($r = 0.64$ for all participants) suggesting the usefulness of using students' survey results to examine the criterion aspect of test validity.

Xie (2015) investigated the impact of learner beliefs on the test scores of the CET test by conducting a study in which about 800 Chinese test takers responded to two questionnaires asking (1) their perception toward skills necessary to answer test questions

correctly and (2) their test preparation activities before taking the CET test. In test based on SEM, the author found a significant path ($\beta = 0.389$, $p < .01$) from the test-taker perception to the test preparation.

Finally, previous studies (Peacock, 2001; Li, 2021) have focused on the nature of learner beliefs focusing on the longitudinal change. Peacock (2001) examined the longitudinal change of beliefs about second language learning of ESL teachers over their 3-year program using BALLI and found nonsignificant differences over the 3 years. In the similar vein, Li (2021) investigated the longitudinal change of Chinese EFL learners' beliefs upon arrival at the university (survey 1) and a year after the arrival (survey 2) by asking their degree of agreement to (1) difficulty of language (perceptions about the difficulty of learning a foreign language in general [e.g., it is difficult for me to take part in group discussion in English]), (2) nature of language learning (perceptions about a wide range of issues concerning the nature of learning a foreign language [to learn English means doing a lot of repetition and practice]), and (3) autonomy in language learning (perceptions about readiness to be autonomous in learning a foreign language [I believe that I should find my own opportunities to use English.]). The study found a significant ($p < .001$) increase of agreement to all three types of questions possibly reflecting the course of study they had gone through for a year.

In summary, investigating the degree of the relationship of test takers' test score to test takers' self-assessment or beliefs has been an important part of validation because their English proficiency measured at a test should match with their perception or belief toward the target test and study practice. Previous studies on large scale tests (Powers & Powers, 2015; Ross, 1998; Xie, 2015; Xie & Andrews, 2013; Sawaki & Nissan, 2009) have found significant relationships between test score and learners' self-assessments and beliefs. In addition, previous studies (Li, 2021; Peacock, 2001) examining the longitudinal change of learner beliefs have found mixed results in terms of the degree of change over a certain period of time, yet these studies examined the change in a cross-sectional way in which the same participants were asked to respond to questions only once.

Purposes of the study

As an attempt to address the research gap, this study is intended to connect the often separately investigated aspects, scoring and criterion related, of validity (O'Sullivan & Weir, 2011) of the TEAP. This study is also intended to further investigate the dimensionality of EFL proficiency based on the four-skill English tests, the longitudinal change of test takers' perceived values toward a test, and the link of test takers' perceived values to the measured constructs of the TEAP.

Thus, this study poses the following three research questions:

1) Which of the four models (unitary, correlated four factor, higher order, and receptive-productive) best represents the test construct of the TEAP?
2) To what degree do test takers' perceptions (task value) toward the measured construct of the TEAP change over time before and after entering university?
3) To what degree are test takers' perceptions (task value) toward the measured construct of the TEAP related to the factor structure of the TEAP?

The first question is expected to shed light on the test dimensionality issue (Gu, 2015; In'nami et al., 2016; Kamiya, 2017; Oller Jr., 1980; Sawaki et al., 2009) which can further contribute to the debate on the existence of a higher-order factor structure of one's EFL proficiency measured by the four skills' test. The first research question will pave the way for the following research questions because these research questions will be conducted based on the findings of the first research question in terms of the factor structure of the TEAP. The second question will contribute to the discussion of longitudinal change of test takers' perceived value of a test (Li, 2021; Peacock, 2001; Xie & Andrews, 2013), which is found to be one of the factors positively affecting test takers' test preparation activity. Finally, the third question, which is based on the discussion of the first and second research questions, will also contribute to the discussion on the link between test takers' perceptions and the measured English proficiency (Xie, 2011; Xie & Andrews, 2013) by investigating the longitudinal change of test takers' perceptions and the relationship to the proficiency measured by the TEAP.

Method

Participants

A total of 2490 first-year undergraduate Japanese learners of English enrolled at a private university in Tokyo participated in this study. Of these students, the data of 273 students were excluded case wise because they did not complete the study. The English proficiency of these students, based on the self-reported overall score (full mark of written and listening sections is 200 and 50, respectively) of the NCT from 1532 of the participants (written mean = 165.3, $SD = 28.2$; listening mean = 43.7, $SD = 8.02$), could be considered to be on the higher end of the overall Japanese third-year high school students (written mean = 118.9, $SD = 41.1$; listening mean = 33.16, $SD = 9.4$) (National Center for University Entrance Examinations, 2017) though this only represents about 70% of all participating students. A questionnaire on the study participants' most focused skill of English at high school showed that 80%, 5%, 8%, and 7% of participants focused mostly on reading, listening, speaking, and writing, respectively. As for the most focused skill of English at university, the participants responded that 19%, 18%, 48%, and 15% focused mostly on reading, listening, speaking, and writing, respectively.

Instrument

The Eiken Foundation of Japan provided a mock version of the TEAP which was equivalent to the actual TEAP tests administered at test centers in terms of the format, administration, content, and difficulty. The reading and listening sections are designed to measure the understanding of short and long passages with visual information including graphs and charts. The speaking section is based on face-to-face, one-on-one interviews and includes both monologue and dialogue tasks on various issues. The writing section includes both summary and integrated tasks. In this study, item-level dichotomous responses were obtained for reading and listening skills, while criterion-level ratings were obtained for speaking and writing skills (see Table 1).

A set of questionnaires (a total of 10 questions) was prepared based on the previous studies (Sawaki & Nissan, 2009; Xie & Andrews, 2013) to investigate test takers' perceived value toward each section of the TEAP. Instead of asking the generic value

Table 1 Structure of the TEAP

Section	Items	Scores
Reading	60 dichotomous items	60 maximum raw score points Reading section score on a scale of 20–100
Listening	50 dichotomous items	50 maximum raw score points Listening section score on a scale of 20–100
Writing	2 items scored on a 0–3 scale	27 maximum raw score points Writing section score on a scale of 20–100
Speaking	4 items scored on a 0–3 scale	15 maximum raw score points Speaking section score on a scale of 20–100

Table 2 Questionnaire to measure test takers' perception toward test construct

	Question
1	*To what degree do you think that knowledge of vocabulary and word usage is important when taking English classes at university?*
2	*To what degree do you think that being able to comprehend the main ideas of English reading texts is important when taking English classes at university?*
3	*To what degree do you think that being able to accurately comprehend the details of English reading texts is important when taking English classes at university?*
4	*To what degree do you think that being able to comprehend graphs and charts, etc. in English is important when taking English classes at university?*
5	*To what degree do you think that being able to understand the main ideas in conversations or lectures in which English is used is important when taking English classes at university?*
6	*To what degree do you think that being able to comprehend accurately the details of conversations and lectures in English is important when taking English classes at university?*
7	*To what degree do you think that being able to ask questions and take part in discussions in English is important when taking English classes at university?*
8	*To what degree do you think that being able to express your own opinions in English about social issues is important when taking English classes at university?*
9	*To what degree do you think that being able to write a summary in English of the main ideas in an English text is important when taking English classes at university?*
10	*To what degree do you think that being able to write an essay in English by integrating information from multiple English texts is important when taking English classes at university?*

toward the test as a whole as was designed in Xie and Andrews (2013), each question asks the usefulness of a measured construct of each section of the TEAP on a 6-point Likert scale of agreement: 1 = I strongly think so and 6 = I strongly do not think so.

Procedure

First, in April 2015, just a few days after entering university, students responded to a set of questionnaire items (see Table 2) which asked their perceptions (as perceived value at high school or PVH) toward the construct measured by the TEAP by section level. Then, in January 2016, 8 months after they started taking university courses, the same group of students took the mock version of the TEAP and responded to the same questionnaire (as perceived value at university or PVU). Students were asked to participate in this series of study as a mandatory part of the university' educational program outside the regular class hours.

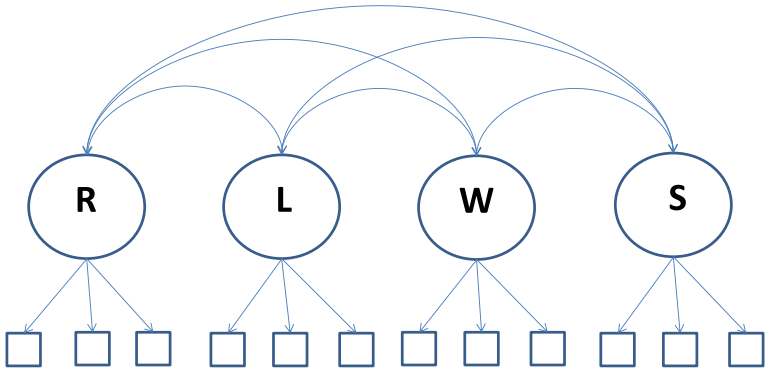

Fig. 1 TEAP correlated four-factor model. R, reading factor. L, listening factor. W, writing factor. S, speaking factor

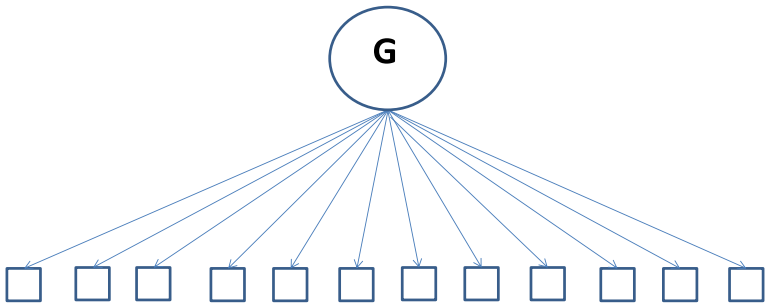

Fig. 2 TEAP single-factor model. G, general factor

Analysis

To address the first research question, the TEAP item-level raw scores were obtained for each participant. Based on the previous discussion on the dimensionality of one's EFL proficiency measured by the four skills' test (Gu, 2015; In'nami & Koizumi, 2011; In'nami et al., 2016; Kamiya, 2017; Sawaki et al., 2009) and on the design of the TEAP which outputs four separate skill scores (reading, listening, writing, and speaking) and an accumulated overall score, four models were hypothesized: a correlated four-factor model, a single-factor model, a higher-order factor model, and a receptive-productive model. The correlated four-factor model (Fig. 1) hypothesizes the presence of four correlated factors corresponding to the four assessed skills. This model assumes that the variance from each item can be explained by a skill-specific factor that is correlated with other skill specific. The correlated four-factor model (Fig. 1) hypothesizes the presence of four correlated factors corresponding to the four assessed skill-specific factors. Previous studies have found this model to have as good a statistical fit to the model as the higher-order factor model (Gu, 2015; In'nami et al., 2016; Sawaki et al., 2009).

The single-factor model specifies that all items from all four skills load on one factor. This model assumes that all the variance from all items can be explained by a single factor. This model assumes that the variance from each item can be explained by a single general factor as posed by Oller Jr. (1980). Figure 2 shows the TEAP single-factor model.

The higher-order factor model hypothesizes the presence of a higher-order factor under which four other factors corresponding to the assessed four language skills are

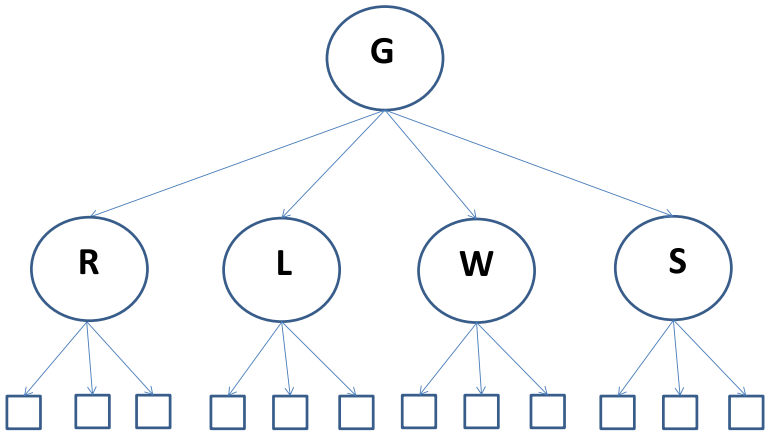

Fig. 3 TEAP higher-order factor model. G, general factor. R, reading factor. L, listening factor. W, writing factor. S, speaking factor

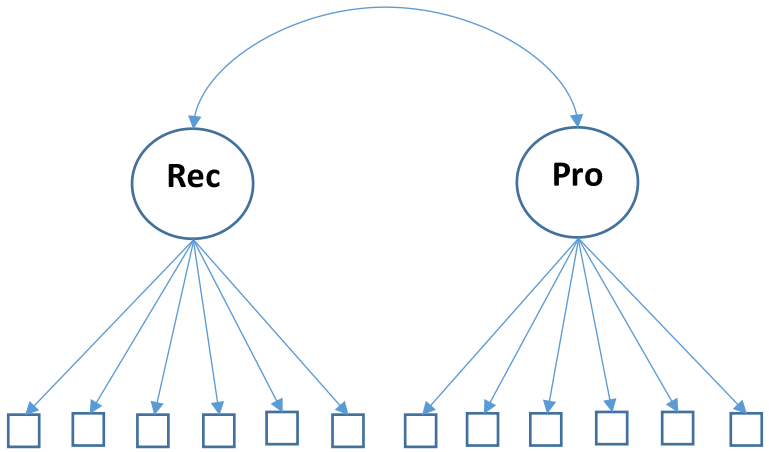

Fig. 4 TEAP receptive-productive factor model. Rec, receptive skill factor. Pro, productive skill factor

controlled. This model assumes that the variance from each item can be explained by a skill-specific factor that is governed by the higher-order model, and previous studies have chosen this model as the final model based on both statistical and theoretical aspects (Gu, 2015; In'nami et al., 2016; Sawaki et al., 2009). Figure 3 shows the TEAP higher-order factor model.

The receptive-productive model hypothesizes the presence of two factors: a receptive factor and a productive factor. This model assumes that the ability or the factor structure of the TEAP is separable into receptive (i.e., reading and listening) and productive (i.e., writing and speaking) components, and previous studies on the factor structure of the four-skill English tests have posed this as one of the competing models (In'nami et al., 2016; Kamiya, 2017). Figure 4 shows the TEAP receptive-productive factor model.

To address the second research question, test takers' responses to the questionnaire, both PVH and PVU, were analyzed by evaluating the model-data fit of the proposed model (see Fig. 5). Figure 5 shows the model in which the PVH affects the PVU

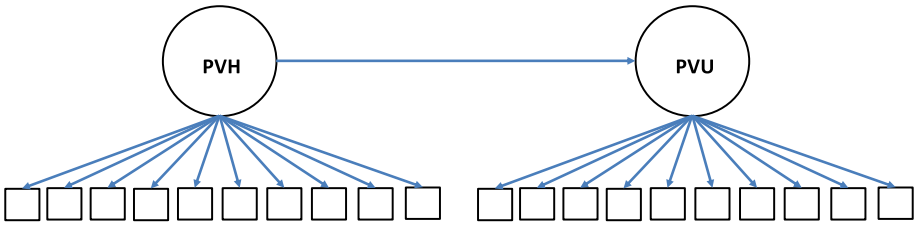

Fig. 5 Test takers' perception longitudinal model. PVH, perceived value at high school. PVU, perceived value at university

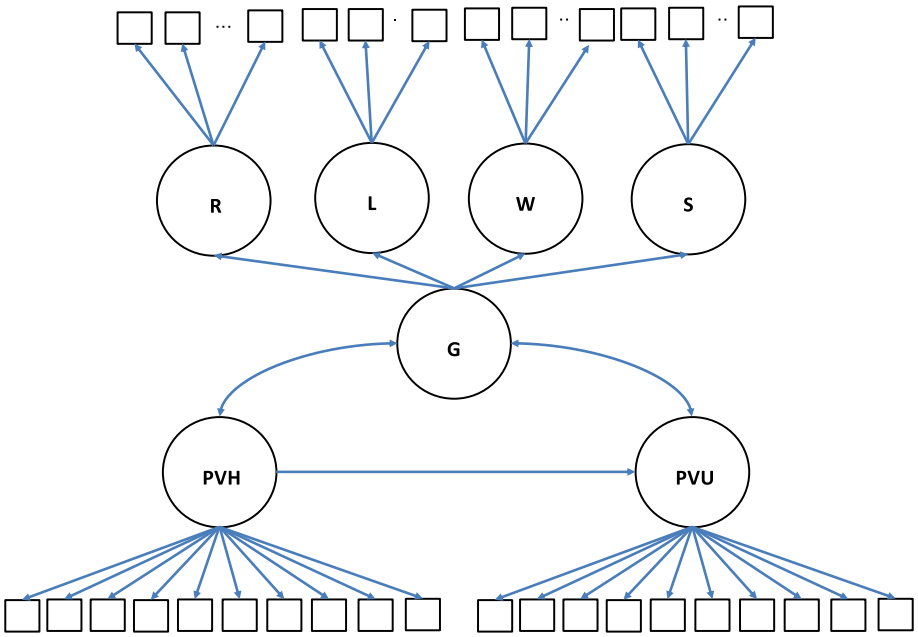

Fig. 6 Test takers' perception and TEAP higher-order model. PVH, perceived value at high school. PVU, perceived value at university. G, general factor. R, reading factor. L, listening factor. W, writing factor. S, speaking factor

For the third research question, the model-data fit of the proposed model in which the final model from the first research question was correlated with the factor structure of the PVH and PVU (see Fig. 6) was investigated.

For the investigation of all research questions of this study, Mplus 8.4 (Muthén & Muthén, 1998–2022) was employed to estimate the parameters and evaluate the model fit. The parameters were estimated using the robust weighted least squares (WLSMV) in order to deal with item-level categorical data of the TEAP. This estimation method was chosen because this WLSMV estimator has shown accurate test statistics, parameter estimates, and standard errors under both normal and non-normal latent response distributions across various sample sizes (Byrne, 2012. In order to ensure the model identification, factor loadings of all the first observed variables across skills (reading, listening, writing, and speaking) were fixed to 1.0. The chi-square, comparative fit index (CFI), Tucker–Lewis index (TLI), root-mean-square error of approximation (RMSEA), and standardized root-mean-square residual (SRMR) were employed to evaluate model

fit. Those indices were chosen because, based on the meta-analysis of CFA studies, they were the most frequently reported (In'nami et al., 2016; In'nami & Koizumi, 2011; Kamiya, 2017; Muthén, 2004). Model fit was evaluated by a nonsignificant chi-square, the CFU and the TLI of 0.95 or above, the RMSEA of 0.06 or below, and the weighted root-mean-square residual (SRMR) of 0.08 or below.

Results

Table 3 shows the results concerning the first research question, CFA for the four proposed models (correlated, high order, single, and receptive-productive). Overall, the correlated model and higher-order model showed the better model-data fit compared to the other two models (correlated and single).

The difference between the correlated and higher-order models was small. This finding was consistent with that of In'nami et al. (2016), who found a very similar degree of model-data fit indices between the two models. In addition, the result was also similar to Sawaki et al. (2009) who found similar model-data fit indices between the TOEFL iBT correlated four-factor model and higher-order model.

For the higher-order model, all the estimated parameters were found to be significant (see Table 9 in Appendix), including the paths from the general higher-order factor to the individual skill factors (reading, listening, writing, and speaking; 0.619–0.946). Standardized parameter estimates for the TEAP correlated-factor model were all significant (see Table 10 in Appendix) including the correlation between factors (from 0.457 to 0.766).

Following the previous literature in which two nesting models were compared (In'nami et al., 2016; Kamiya, 2017; Sawaki et al., 2009), these two models (correlated and higher order) were further compared using the Mplus DIFFTEST command (Muthén & Muthén, 1998–2022), which is often used when the data contain dichotomous or categorical data. The DIFFTEST result showed that the correlated-factor model was significantly better than the higher-order model (χ^2 difference = 57.231, *df* difference = 2, $p = 0.000$).

However, the difference between the two models in terms of CFI, TLI, RMSEA, and SRMR was minimal (Table 2), showing that these two models were practically equivalent. As the higher-order model was more parsimonious than the correlated-factor model and it was consistent with previous research (In'nami et al., 2016), the higher-order model was chosen as the final model. This argument was in line with Sawaki et al. (2009) that, even though a significant chi-square difference test (correlated-factor model vs. higher-order model) implied that the correlated-factor model was better than the higher-order model, and the difference of fit indices between the

Table 3 Fit indices for the models

Model	Chi-square	*df*	CFI	TLI	RMSEA [CI]	SRMR
Correlated	14851.478*	7496	0.948	0.947	0.021 [0.021, 0.022]	0.051
Single	32137.384*	7502	0.825	0.822	0.038 [0.038, 0.039]	0.072
Higher order	15048.220*	7498	0.946	0.945	0.021 [0.021, 0.022]	0.051
Receptive-productive	23026.457*	7501	0.890	0.888	0.031 [0.030, 0.031]	0.062

df degrees of freedom, *CI* 90% confidence interval

* $p < 0.05$

Table 4 Descriptive statistics for questionnaire items on test-taker perception

	High school	University
Vocabulary and word usage	5.32 (0.92)	4.69 (1.44)
Comprehending main ideas by reading	5.63 (0.67)	4.92 (1.39)
Comprehending details by reading	5.19 (0.99)	4.54 (1.39)
Comprehending graphs/charts by reading	5.14 (0.97)	4.47 (1.38)
Comprehending main ideas by listening	5.69 (0.61)	5.06 (1.38)
Comprehending details by listening	5.30 (0.91)	4.78 (1.36)
Asking questions and taking part in discussion	5.55 (0.77)	4.99 (1.39)
Expressing opinions	5.55 (0.77)	4.88 (1.40)
Writing a summary	5.44 (0.82)	4.85 (1.40)
Writing an essay	5.13 (1.00)	4.61 (1.40)

Table 5 Fit indices for the proposed model

Model	Chi-square	*df*	CFI	TLI	RMSEA [CI]	SRMR
Test takers' perception longitudinal model	2186.617*	169	0.986	0.984	0.073 [0.071, 0.076]	0.040

TOEFL iBT higher-order model and correlated-factor model was minimal, which allowed them to make the final decision that the TOEFL iBT higher-order model was the best model. In addition, the result fits well with the current score reporting policy in which a composite score is reported together with scores of each individual skill.

Table 4 shows the results of descriptive statistics for questionnaire items on test-taker perception measured at high school and university. The result shows that across all 10 questionnaire items, on average, the degrees of perceived value by test takers at university became lower than the value perceived at high school, while degrees of variance at university became greater than at high school.

Table 4 shows the results of structural equation model (SEM) for the test takers' perception longitudinal model.

Overall, the model showed a decent model-data fit. Table 5 shows the estimated standardized parameter estimates for the test takers' perception longitudinal model. All the estimated parameters as well as the path from PVH ($\alpha = 0.90$) to PVU ($\alpha = 0.98$) were significant, showing that the value perceived by test takers at high school positively affected the value perceived at university.

Table 6 shows the results of SEM for the test takers' perception and the TEAP higher-order model. Overall, the model showed decent model-data fit. Table 7 shows the estimated standardized parameter estimates for the test takers' perception longitudinal model. All the estimated parameters were found to be positive and also significant, and the path from the general English proficiency to the PVU, the correlations between the general PVH and general PVU, was also significant. The correlation between the general PVH ($r = 0.178$) was higher than the general PVU ($r = 0.131$), showing that the value perceived at high school has more impact on the measured construct than the perceived value at university. In addition, parameter estimates from each skill factor were all significant together with those from the higher-order general factor to skill factors (Table 8).

Table 6 Standardized parameter estimates for the test takers' perception longitudinal model

		Estimate	SE	Est./SE	Two tailed *p*-Value
PVH	By				
U3		0.658	0.015	44.486	0.000
U4		0.821	0.011	74.429	0.000
U5		0.765	0.011	70.699	0.000
U6		0.681	0.014	50.412	0.000
U7		0.861	0.010	83.631	0.000
U8		0.801	0.010	79.182	0.000
U9		0.870	0.008	104.624	0.000
U10		0.868	0.009	99.045	0.000
U11		0.835	0.009	91.260	0.000
U12		0.781	0.010	76.876	0.000
PVU	By				
U138		0.850	0.007	121.822	0.000
U139		0.915	0.004	230.758	0.000
U140		0.897	0.005	190.051	0.000
U141		0.860	0.006	145.234	0.000
U142		0.935	0.004	247.588	0.000
U143		0.900	0.004	208.411	0.000
U144		0.931	0.004	252.137	0.000
U145		0.915	0.004	225.613	0.000
U146		0.933	0.003	286.205	0.000
U147		0.871	0.005	161.088	0.000
PVU	On				
PVH		0.248	0.021	11.532	0.000

Table 7 Fit indices for the test takers' perception and the TEAP higher-order model

Model	Chi-square	*df*	CFI	TLI	RMSEA [CI]	SRMR
Test takers' perception longitudinal model	18077.159*	10145	0.970	0.970	0.019 [0.018, 0.019]	0.049

* $p < 0.05$

Discussion and conclusion

In order to further collect validity evidence for the TEAP test, the author examined the following: (1) the factor structure of the TEAP to investigate scoring validity, (2) the degree of longitudinal (high school and university) change of test takers' perceived value toward the measured tasks, and (3) the relationship between the perceived value factors and the TEAP factor to investigate criterion-related validity.

For the first research question, confirmatory factor analysis was conducted on the data collected from 2217 first-year Japanese university students. This research question asked which factor model best explains the TEAP data (single-factor, correlated four-factor, higher-order factor, and receptive-productive factor). Of these four item-level response models, the higher-order model best explained the TEAP data. This result also supports the current score reporting policy in which a composite overall score is computed from scores from individual skill scores. This result replicated a previous study (In'nami

Table 8 Standardized parameter estimates for the test takers' perception and the TEAP higher-order model

		Estimate	SE	Est./SE	Two tailed *p*-Value
PVH	By				
U3		0.637	0.017	38.098	0.000
U4		0.819	0.012	68.023	0.000
U5		0.748	0.012	62.189	0.000
U6		0.669	0.015	45.730	0.000
U7		0.872	0.011	80.828	0.000
U8		0.792	0.011	71.394	0.000
U9		0.875	0.009	98.849	0.000
U10		0.878	0.009	96.153	0.000
U11		0.839	0.010	86.134	0.000
U12		0.792	0.011	74.999	0.000
PVU	By				
U138		0.848	0.007	121.049	0.000
U139		0.917	0.004	229.688	0.000
U140		0.895	0.005	187.323	0.000
U141		0.857	0.006	144.135	0.000
U142		0.937	0.004	244.570	0.000
U143		0.898	0.004	203.572	0.000
U144		0.932	0.004	248.186	0.000
U145		0.916	0.004	223.741	0.000
U146		0.933	0.003	279.472	0.000
U147		0.872	0.006	158.531	0.000
Reading	By				
U14		0.389	0.026	14.906	0.000
U15		0.318	0.028	11.434	0.000
U16		0.409	0.029	14.176	0.000
U17		0.300	0.029	10.445	0.000
U18		0.247	0.031	7.977	0.000
U19		0.495	0.025	19.496	0.000
U20		0.382	0.029	12.971	0.000
U21		0.347	0.027	12.647	0.000
U22		0.440	0.035	12.664	0.000
U23		0.645	0.021	31.275	0.000
U24		0.505	0.024	20.981	0.000
U25		0.398	0.028	14.294	0.000
U26		0.122	0.029	4.233	0.000
U27		0.371	0.027	13.694	0.000
U28		0.489	0.026	19.115	0.000
U29		0.414	0.026	15.798	0.000
U30		0.373	0.028	13.546	0.000
U31		0.373	0.027	13.786	0.000
U32		0.190	0.038	4.956	0.000
U33		0.440	0.028	15.505	0.000
U34		0.536	0.027	19.926	0.000
U35		0.290	0.032	9.041	0.000
U36		0.532	0.027	19.984	0.000
U37		0.272	0.028	9.820	0.000

Table 8 (continued)

		Estimate	SE	Est./SE	Two tailed *p*-Value
U38		0.442	0.026	16.815	0.000
U39		0.396	0.028	14.088	0.000
U40		0.630	0.025	25.723	0.000
U41		0.617	0.027	22.527	0.000
U42		0.707	0.027	26.179	0.000
U43		0.451	0.025	17.914	0.000
U44		0.503	0.025	20.056	0.000
U45		0.678	0.022	30.946	0.000
U46		0.650	0.021	31.636	0.000
U47		0.612	0.022	27.661	0.000
U48		0.379	0.027	14.013	0.000
U49		0.280	0.028	10.143	0.000
U50		0.637	0.021	29.812	0.000
U51		0.728	0.019	38.199	0.000
U52		0.782	0.016	47.642	0.000
U53		0.756	0.018	43.130	0.000
U54		0.427	0.026	16.167	0.000
U55		0.554	0.023	24.051	0.000
U56		0.459	0.025	18.210	0.000
U57		0.572	0.022	25.771	0.000
U58		0.180	0.029	6.108	0.000
U59		0.684	0.023	29.586	0.000
U60		0.238	0.028	8.433	0.000
U61		0.578	0.022	25.915	0.000
U62		0.203	0.030	6.726	0.000
U63		0.442	0.026	17.287	0.000
U64		0.642	0.021	31.092	0.000
U65		0.676	0.020	33.137	0.000
U66		0.352	0.027	13.017	0.000
U67		0.643	0.021	30.153	0.000
U68		0.293	0.028	10.539	0.000
U69		0.320	0.027	11.834	0.000
U70		0.599	0.026	23.145	0.000
U71		0.429	0.028	15.128	0.000
U72		0.604	0.025	24.616	0.000
U73		0.429	0.029	15.034	0.000
Listening	By				
U74		0.471	0.024	19.666	0.000
U75		0.643	0.021	31.191	0.000
U76		0.645	0.022	29.684	0.000
U77		0.630	0.022	28.220	0.000
U78		0.556	0.022	25.034	0.000
U79		0.572	0.024	23.608	0.000
U80		0.631	0.020	31.337	0.000
U81		0.523	0.023	22.344	0.000
U82		0.478	0.025	18.945	0.000
U83		0.256	0.030	8.612	0.000
U84		0.593	0.022	26.712	0.000

Table 8 (continued)

		Estimate	SE	Est./SE	Two tailed *p*-Value
U85		0.443	0.025	17.849	0.000
U86		0.540	0.023	23.973	0.000
U87		0.420	0.028	15.221	0.000
U88		0.540	0.022	24.051	0.000
U89		0.439	0.027	16.308	0.000
U90		0.302	0.027	11.115	0.000
U91		0.322	0.030	10.688	0.000
U92		0.434	0.025	17.611	0.000
U93		0.716	0.019	38.246	0.000
U94		0.362	0.035	10.227	0.000
U95		0.370	0.026	14.151	0.000
U96		0.503	0.025	19.894	0.000
U97		0.426	0.026	16.671	0.000
U98		0.452	0.025	18.428	0.000
U99		0.613	0.021	29.834	0.000
U100		0.510	0.024	21.252	0.000
U101		0.569	0.024	23.630	0.000
U102		0.474	0.024	20.105	0.000
U103		0.644	0.020	32.260	0.000
U104		0.411	0.032	12.971	0.000
U105		0.572	0.024	23.754	0.000
U106		0.506	0.023	21.795	0.000
U107		0.592	0.024	25.058	0.000
U108		0.455	0.029	15.591	0.000
U109		0.292	0.028	10.562	0.000
U110		0.367	0.026	13.840	0.000
U111		0.589	0.023	25.672	0.000
U112		0.134	0.036	3.731	0.000
U113		0.422	0.028	15.185	0.000
U114		0.078	0.032	2.451	0.014
U115		0.286	0.027	10.516	0.000
U116		0.428	0.026	16.779	0.000
U117		0.557	0.026	21.723	0.000
U118		0.489	0.025	19.622	0.000
U119		0.318	0.030	10.732	0.000
U120		0.377	0.027	14.095	0.000
U121		0.285	0.027	10.468	0.000
U122		0.409	0.026	15.450	0.000
U123		0.334	0.027	12.467	0.000
Writing	By				
U124		0.677	0.017	40.114	0.000
U125		0.837	0.012	70.640	0.000
U126		0.816	0.013	62.515	0.000
U127		0.833	0.012	67.512	0.000
U128		0.824	0.010	82.839	0.000
U129		0.864	0.008	103.508	0.000
U130		0.924	0.007	138.923	0.000
U131		0.938	0.007	137.338	0.000

Table 8 (continued)

		Estimate	SE	Est./SE	Two tailed *p*-Value
U132		0.902	0.008	117.339	0.000
Speaking	By				
U133		0.828	0.012	68.835	0.000
U134		0.953	0.007	140.641	0.000
U135		0.959	0.006	153.998	0.000
U136		0.930	0.008	118.319	0.000
U137		0.848	0.010	83.382	0.000
General	By				
Reading		0.795	0.013	63.461	0.000
Listening		0.955	0.010	100.182	0.000
Writing		0.619	0.015	42.091	0.000
Speaking		0.726	0.014	52.350	0.000
PVU	On				
PVH		0.248	0.022	11.521	0.000
General	With				
PVU		0.131	0.024	5.487	0.000
PVH		0.178	0.024	7.299	0.000

et al., 2016) which also examined the factor structure of the TEAP, providing added evidence to the scoring validity of the TEAP. The estimated loadings of first-order factors on higher-order factor (0.807, 0.946, 0.619, and 0.718 for reading, listening, writing, and speaking, respectively) were similar to those found in the previous study (In'nami et al., 2016), also suggesting the generalizability of the factor structure of the TEAP.

One explanation for this finding is that both studies were conducted on data from Japanese university students with similar regional representativeness (i.e., Tokyo area), while the results could be different if the data were collected from other age groups of students (e.g., high school students). In addition, this study also adds evidence to the existence of a higher-order factor structure of the four-skill English tests in previous studies (e.g., Sawaki et al., 2009) on the issue of dimensionality of a language ability (Gu, 2015; Oller Jr., 1980; Sawaki et al., 2009). As for the second research question, confirmatory factor analysis was used on the same data as the first research question. The second research question concerned the degree to which test takers' perception (task value) changes over time before and after entering university. Descriptive statistics of questionnaire items found that the degree of agreement at high school was higher than at university, while the variance of agreement was greater at university than at high school across all 10 items. Standardized factor loadings across tasks for the perceived value at university were higher (from 0.848 to 0.937) than the perceived value at high school (from 0.637 to 0.878) with more varying degrees of loadings. In addition, the perceived value factor at high school toward measured tasks (PVH) positively ($\beta = 0.248$) affected the perceived value at university (PVU). Based on the result, this study shed light on the nature of perceived value across different tasks and its longitudinal change (from high school to university). The result suggests that students perceived the values for tasks with a varying degree, and that the perceived value at high school positively affected that at university. The reason behind this longitudinal change of perceived value might be

explained by the fact that students went through a series of courses taught in English for 8 months at university, perceiving various aspects of each skill of English. The majority of students (80%) reported that high school English language instruction focused on the reading skill the most, while the trend changed after 8 months when the four skills were more equally focused upon. This trend probably reflects the nature of actual courses at university which gave students the opportunity to learn various aspects of each skill of English. This result might suggest a positive washback of the TEAP which assesses the four skills of English leading to more balanced teaching and learning of English at high school.

Regarding the third question, this study found that the TEAP higher-order general factor positively correlated with the perceived value factors both at high school ($r = 0.178$) and university ($r = 0.131$). This result is consistent with Xie and Andrews (2013), who found a positive impact of perceived test value on students' test preparation, while this study found a positive relationship of perceived skill-based value and measured test result. This result also adds evidence to the discussion of criterion-related aspects of test validity suggesting the possibility of including the self-reported perceived value as a criterion to measure a test's validity.

Implications and future research

First, this study identified the higher-order model as the best representation of the underlying factor structure of the TEAP test, which supports the scoring validity of the TEAP as test takers receive not only a score for each skill but also a composite score. This is because universities usually consider admissions based on the submitted composite scores rather than looking at each skill score. However, future research is required to investigate the generalizability of this study by extending the participants to broader populations including students with more diverse English proficiency.

Second, this study found that students' perceived value toward the tasks on a test changes over time before and after entering university in terms of strength and variation. This suggests that we need to take into account that students tend to have a varying degree of perceived value toward different task types reflecting their study practice at each stage of their English study. Additionally, it could be important to raise students' awareness toward each skill of English at high school because the perceived value at high school could positively affect that at university. However, future studies are required to investigate the longitudinal change of students' perceived value over time further by possibly extending the interval between university entrance and the time they graduate. In addition, more qualitative aspect of students' perceived value could be investigated by conducting interviews. Such a study would shed light on the rationale behind test takers' perceptions and the changes overtime after entrance to university.

Third, this study identified a positive relationship between test takers' value toward tasks measured by the TEAP and the measured construct of the TEAP, showing the importance of perceived value toward each aspect of the test constructs when considering the validity of a test. Future study is needed to further include the amount of time spent on test preparation activities.

Appendix

Tables 9 and 10

Table 9 Standardized parameter estimates for the TEAP higher-order model

		Estimate	SE	Est./SE	Two tailed *p*-Value
Reading	By				
U14		0.388	0.026	14.932	0.000
U15		0.323	0.028	11.663	0.000
U16		0.412	0.029	14.373	0.000
U17		0.302	0.029	10.545	0.000
U18		0.250	0.031	8.152	0.000
U19		0.499	0.025	19.772	0.000
U20		0.384	0.029	13.076	0.000
U21		0.348	0.027	12.758	0.000
U22		0.446	0.035	12.873	0.000
U23		0.642	0.021	31.135	0.000
U24		0.506	0.024	21.146	0.000
U25		0.400	0.028	14.466	0.000
U26		0.128	0.029	4.437	0.000
U27		0.369	0.027	13.698	0.000
U28		0.488	0.025	19.194	0.000
U29		0.412	0.026	15.767	0.000
U30		0.368	0.027	13.381	0.000
U31		0.373	0.027	13.858	0.000
U32		0.183	0.038	4.804	0.000
U33		0.444	0.028	15.770	0.000
U34		0.539	0.027	20.043	0.000
U35		0.291	0.032	9.106	0.000
U36		0.536	0.026	20.313	0.000
U37		0.272	0.028	9.848	0.000
U38		0.448	0.026	17.124	0.000
U39		0.401	0.028	14.388	0.000
U40		0.631	0.024	25.917	0.000
U41		0.616	0.027	22.645	0.000
U42		0.705	0.027	26.341	0.000
U43		0.451	0.025	18.001	0.000
U44		0.503	0.025	20.114	0.000
U45		0.677	0.022	30.843	0.000
U46		0.652	0.020	31.940	0.000
U47		0.616	0.022	28.061	0.000
U48		0.377	0.027	14.057	0.000
U49		0.281	0.027	10.229	0.000
U50		0.637	0.021	29.963	0.000
U51		0.728	0.019	38.267	0.000
U52		0.785	0.016	48.277	0.000
U53		0.753	0.018	42.879	0.000
U54		0.428	0.026	16.281	0.000
U55		0.555	0.023	24.375	0.000
U56		0.460	0.025	18.367	0.000

Table 9 (continued)

		Estimate	SE	Est./SE	Two tailed *p*-Value
U57		0.571	0.022	25.874	0.000
U58		0.183	0.029	6.258	0.000
U59		0.683	0.023	29.554	0.000
U60		0.238	0.028	8.478	0.000
U61		0.581	0.022	26.301	0.000
U62		0.202	0.030	6.732	0.000
U63		0.443	0.025	17.451	0.000
U64		0.640	0.021	31.025	0.000
U65		0.675	0.020	33.497	0.000
U66		0.351	0.027	13.029	0.000
U67		0.640	0.021	30.158	0.000
U68		0.294	0.028	10.642	0.000
U69		0.325	0.027	12.084	0.000
U70		0.593	0.026	23.043	0.000
U71		0.422	0.028	14.981	0.000
U72		0.599	0.024	24.566	0.000
U73		0.423	0.028	14.943	0.000
Listening	By				
U74		0.471	0.024	19.638	0.000
U75		0.638	0.021	30.689	0.000
U76		0.641	0.022	29.378	0.000
U77		0.630	0.022	28.009	0.000
U78		0.560	0.022	25.199	0.000
U79		0.572	0.024	23.569	0.000
U80		0.630	0.020	31.206	0.000
U81		0.524	0.023	22.399	0.000
U82		0.475	0.025	18.831	0.000
U83		0.254	0.030	8.566	0.000
U84		0.595	0.022	26.748	0.000
U85		0.446	0.025	18.033	0.000
U86		0.545	0.022	24.229	0.000
U87		0.418	0.028	15.147	0.000
U88		0.538	0.022	23.913	0.000
U89		0.438	0.027	16.307	0.000
U90		0.302	0.027	11.123	0.000
U91		0.323	0.030	10.742	0.000
U92		0.434	0.025	17.650	0.000
U93		0.713	0.019	37.886	0.000
U94		0.365	0.035	10.396	0.000
U95		0.370	0.026	14.209	0.000
U96		0.501	0.025	19.834	0.000
U97		0.425	0.026	16.603	0.000
U98		0.457	0.024	18.691	0.000
U99		0.614	0.021	29.866	0.000
U100		0.514	0.024	21.413	0.000
U101		0.571	0.024	23.705	0.000
U102		0.478	0.024	20.302	0.000
U103		0.642	0.020	31.995	0.000

Table 9 (continued)

		Estimate	SE	Est./SE	Two tailed *p*-Value
U104		0.415	0.032	13.146	0.000
U105		0.570	0.024	23.656	0.000
U106		0.502	0.023	21.586	0.000
U107		0.588	0.024	24.803	0.000
U108		0.460	0.029	15.777	0.000
U109		0.295	0.028	10.689	0.000
U110		0.361	0.027	13.589	0.000
U111		0.591	0.023	25.908	0.000
U112		0.137	0.036	3.804	0.000
U113		0.422	0.028	15.218	0.000
U114		0.082	0.032	2.581	0.010
U115		0.291	0.027	10.734	0.000
U116		0.425	0.026	16.678	0.000
U117		0.553	0.026	21.622	0.000
U118		0.493	0.025	19.870	0.000
U119		0.313	0.030	10.540	0.000
U120		0.378	0.027	14.100	0.000
U121		0.286	0.027	10.495	0.000
U122		0.408	0.026	15.403	0.000
U123		0.335	0.027	12.500	0.000
Writing	By				
U124		0.678	0.017	40.232	0.000
U125		0.837	0.012	71.248	0.000
U126		0.817	0.013	63.257	0.000
U127		0.836	0.012	68.467	0.000
U128		0.824	0.010	83.570	0.000
U129		0.863	0.008	103.589	0.000
U130		0.925	0.007	139.631	0.000
U131		0.938	0.007	137.876	0.000
U132		0.901	0.008	118.213	0.000
Speaking	By				
U133		0.825	0.012	68.466	0.000
U134		0.954	0.007	141.560	0.000
U135		0.958	0.006	152.879	0.000
U136		0.930	0.008	118.529	0.000
U137		0.849	0.010	84.008	0.000
General	By				
Reading		0.807	0.012	65.649	0.000
Listening		0.946	0.010	99.454	0.000
Writing		0.619	0.015	41.796	0.000
Speaking		0.718	0.014	50.658	0.000

Table 10 Standardized parameter estimates for the TEAP correlated-factor model

		Estimate	SE	Est./SE	Two tailed *p*-Value
Reading	By				
U14		0.387	0.026	14.906	0.000
U15		0.323	0.028	11.675	0.000
U16		0.412	0.029	14.375	0.000
U17		0.303	0.029	10.593	0.000
U18		0.252	0.031	8.206	0.000
U19		0.499	0.025	19.824	0.000
U20		0.386	0.029	13.143	0.000
U21		0.349	0.027	12.799	0.000
U22		0.447	0.035	12.926	0.000
U23		0.641	0.021	31.070	0.000
U24		0.507	0.024	21.172	0.000
U25		0.402	0.028	14.538	0.000
U26		0.129	0.029	4.478	0.000
U27		0.369	0.027	13.706	0.000
U28		0.488	0.025	19.189	0.000
U29		0.412	0.026	15.775	0.000
U30		0.368	0.027	13.395	0.000
U31		0.373	0.027	13.884	0.000
U32		0.183	0.038	4.804	0.000
U33		0.444	0.028	15.795	0.000
U34		0.540	0.027	20.048	0.000
U35		0.291	0.032	9.125	0.000
U36		0.537	0.026	20.336	0.000
U37		0.272	0.028	9.858	0.000
U38		0.448	0.026	17.147	0.000
U39		0.401	0.028	14.410	0.000
U40		0.631	0.024	25.902	0.000
U41		0.616	0.027	22.652	0.000
U42		0.705	0.027	26.367	0.000
U43		0.451	0.025	17.988	0.000
U44		0.503	0.025	20.115	0.000
U45		0.677	0.022	30.869	0.000
U46		0.652	0.020	31.976	0.000
U47		0.616	0.022	28.070	0.000
U48		0.377	0.027	14.061	0.000
U49		0.281	0.027	10.255	0.000
U50		0.637	0.021	29.979	0.000
U51		0.728	0.019	38.212	0.000
U52		0.785	0.016	48.229	0.000
U53		0.753	0.018	42.831	0.000
U54		0.427	0.026	16.285	0.000
U55		0.556	0.023	24.403	0.000
U56		0.460	0.025	18.377	0.000
U57		0.571	0.022	25.859	0.000
U58		0.184	0.029	6.292	0.000
U59		0.683	0.023	29.538	0.000
U60		0.238	0.028	8.473	0.000

Table 10 (continued)

		Estimate	SE	Est./SE	Two tailed *p*-Value
U61		0.581	0.022	26.313	0.000
U62		0.202	0.030	6.733	0.000
U63		0.442	0.025	17.441	0.000
U64		0.639	0.021	30.987	0.000
U65		0.675	0.020	33.447	0.000
U66		0.350	0.027	13.007	0.000
U67		0.640	0.021	30.113	0.000
U68		0.294	0.028	10.644	0.000
U69		0.325	0.027	12.083	0.000
U70		0.593	0.026	23.006	0.000
U71		0.422	0.028	14.951	0.000
U72		0.599	0.024	24.507	0.000
U73		0.422	0.028	14.901	0.000
Listening	By				
U74		0.471	0.024	19.701	0.000
U75		0.639	0.021	30.810	0.000
U76		0.642	0.022	29.443	0.000
U77		0.630	0.022	28.043	0.000
U78		0.560	0.022	25.239	0.000
U79		0.572	0.024	23.578	0.000
U80		0.629	0.020	31.249	0.000
U81		0.524	0.023	22.426	0.000
U82		0.475	0.025	18.880	0.000
U83		0.255	0.030	8.594	0.000
U84		0.595	0.022	26.811	0.000
U85		0.446	0.025	18.073	0.000
U86		0.545	0.022	24.302	0.000
U87		0.418	0.028	15.178	0.000
U88		0.538	0.022	23.974	0.000
U89		0.438	0.027	16.323	0.000
U90		0.302	0.027	11.144	0.000
U91		0.323	0.030	10.764	0.000
U92		0.434	0.025	17.655	0.000
U93		0.713	0.019	37.900	0.000
U94		0.364	0.035	10.395	0.000
U95		0.370	0.026	14.198	0.000
U96		0.501	0.025	19.823	0.000
U97		0.425	0.026	16.632	0.000
U98		0.457	0.024	18.703	0.000
U99		0.614	0.021	29.938	0.000
U100		0.514	0.024	21.487	0.000
U101		0.571	0.024	23.740	0.000
U102		0.478	0.024	20.319	0.000
U103		0.641	0.020	32.033	0.000
U104		0.415	0.032	13.168	0.000
U105		0.570	0.024	23.735	0.000
U106		0.502	0.023	21.606	0.000
U107		0.588	0.024	24.825	0.000

Table 10 (continued)

		Estimate	SE	Est./SE	Two tailed *p*-Value
U108		0.460	0.029	15.784	0.000
U109		0.295	0.028	10.691	0.000
U110		0.361	0.027	13.612	0.000
U111		0.591	0.023	25.959	0.000
U112		0.137	0.036	3.821	0.000
U113		0.422	0.028	15.219	0.000
U114		0.083	0.032	2.609	0.009
U115		0.291	0.027	10.739	0.000
U116		0.425	0.025	16.687	0.000
U117		0.553	0.026	21.641	0.000
U118		0.493	0.025	19.879	0.000
U119		0.314	0.030	10.575	0.000
U120		0.378	0.027	14.099	0.000
U121		0.286	0.027	10.481	0.000
U122		0.408	0.026	15.438	0.000
U123		0.335	0.027	12.497	0.000
Writing	By				
U124		0.678	0.017	40.327	0.000
U125		0.838	0.012	71.303	0.000
U126		0.817	0.013	63.255	0.000
U127		0.836	0.012	68.435	0.000
U128		0.824	0.010	83.560	0.000
U129		0.863	0.008	103.432	0.000
U130		0.925	0.007	139.486	0.000
U131		0.938	0.007	137.967	0.000
U132		0.901	0.008	118.335	0.000
Speaking	By				
U133		0.825	0.012	69.035	0.000
U134		0.954	0.007	142.546	0.000
U135		0.958	0.006	153.501	0.000
U136		0.930	0.008	119.192	0.000
U137		0.849	0.010	84.289	0.000
Listening	With				
Reading		0.766	0.012	66.040	0.000
Writing	With				
Reading		0.530	0.016	32.785	0.000
Listening		0.547	0.016	34.326	0.000
Speaking	With				
Reading		0.534	0.018	30.089	0.000
Listening		0.716	0.014	52.875	0.000
Writing		0.457	0.018	25.596	0.000

Abbreviations

TEAP	Test of English for Academic Purposes
MEXT	Ministry of Education, Culture, Sports, Science and Technology
CEFR	Common European Framework of References for Languages
NCT	National Center Test
CT	Common Test
IELTS	International English Language Testing System
GTEC	Global Test of English Communication
TOEFL iBT	e Test of English as a Foreign Language Internet-based Test
CRELLA	Center for Research in English Language Learning and Assessment
EFL	English as a Foreign Language
CFA	Confirmatory factor analysis
TOEIC	Test of English for International Communication
CFI	Comparative fit index
RMSEA	Root-mean-square error of approximation
ETS	Educational Testing Service
BALLI	Beliefs about Language Learning Inventory
CET	College English Test
SEM	Structural equation modeling
ESL	English as a Second Language
SD	Standard deviation
PVH	Perceived value at high school
PVU	Perceived value at university
WLSMV	Weighted least squares mean and variance adjusted
TLI	Tucker–Lewis index
SRMR	Standardized root -mean-square residual

Acknowledgements
The author thanks all the students who participated in this study. The author would also like to express my sincere gratitude for Dr. Yasuyo Sawaki at the Waseda University Graduate School of Education for her helpful discussions and insightful comments on the manuscript.

Author's contributions
The author read and approved the final manuscript.

Funding
Not applicable

Availability of data and materials
The data that support the findings of this study are available from the Eiken Foundation of Japan, but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data may however be available from the author upon reasonable request and with permission of the Eiken Foundation of Japan.

Declarations

Competing interests
The author is employed by the Eiken Foundation of Japan, which develops and administers the TEAP test.

Received: 7 June 2022 Accepted: 12 August 2022
Published online: 23 September 2022

References

Allen, D. (2020). Proposing change in university entrance examinations: A tale of two metaphors. *TEVAL - Shiken: A Journal of Language Testing and Evaluation in Japan*, *24*(2), 23–38.
Bachman, L., & Palmer, A. (1981). The construct validation of the FSI oral interview. *Language Learning*, *31*(1), 67–77.
Benesse. (2022). GTEC CBT. http://www.benesse-gtec.com/cbt/en. Accessed 11 May 2022.
British Council, IDP, IELTS Australia, & Cambridge English Language Assessment (2022). IELTS. http://www.ielts.org/. Accessed 11 May 2022.
Byrne, B. (2012). *Structural equation modeling with Mplus*. New York: Routledge.
Chalhoub-Deville, M. (2016). Validity theory: Reform policies, accountabilities testing, and consequences. *Language Testing*, *33*(4), 453–472. https://doi.org/10.1177/0265532215593312.
Council of Europe (2001). Common European Framework of Reference for Languages: Learning, teaching, assessment. Retrieved May 2022 from https://rm.coe.int/1680459f97
Dornyei, Z. (2005). The Psychology of the Language Learner. New Jersey: Routledge
Eccles, J. and Wigfield, A. (2002). Motivational Beliefs, Values, and Goals. Annual Review Psychology, 53, 109-132.
Educational Testing Service. (2010). Linking TOEFL iBT Scores to IELTS scores. A Research Report. Princeton, NJ: ETS.

Educational Testing Service. (2022). About the TOEFL iBT® test. https://www.ets.org/toefl/ibt/about. Accessed 11 May 2022.
Eiken Foundation of Japan. (2022a). EIKEN tests. http://www.eiken.or.jp/eiken/en/eiken-tests/. Accessed 11 May 2022.
Eiken Foundation of Japan. (2022b). TEAP kenkyu report [TEAP research reports]. http://www.eiken.or.jp/teap/group/report.html. Accessed 11 May 2022.
Green, A. (2014). The Test of English for Academic Purposes (TEAP) impact study: Report 1—Preliminary questionnaires to Japanese high school students and teachers. http://www.eiken.or.jp/teap/group/pdf/teap_washback_study.pdf. Accessed 11 May 2022.
Gu, L. (2015). Language ability of young English language learners: Definition, configuration, and implications. *Language Testing*, *32*(1), 21–38. https://doi.org/10.1177/0265532214542670.
Horwitz, E. (1988). The Beliefs about Language Learning of Beginning University Foreign Language Students. *The Modern Language Journal, 72*(3), 283–294.
In'nami, Y., & Koizumi, R. (2011). Factor structure of the revised TOEIC test: A multiple-sample analysis. *Language Testing*, *29*(1), 131–152. https://doi.org/10.1177/0265532211413444.
In'nami, Y., Koizumi, R., & Nakamura, K. (2016). Factor structure of the Test of English for Academic Purposes (TEAP) test in relation to the TOEFL iBT test. *Language Testing in Asia*, *6*(3), 1–23. https://doi.org/10.1186/s40468-016-0025-9.
Kamiya, N. (2017). Can the National Center Test in Japan be replaced by commercially available private English tests of four skills? In the case of TOEFL Junior Comprehensive. *Language Testing in Asia*, *7*(15), 1–22. https://doi.org/10.1186/s40468-017-0046-z.
Kuramoto, N., & Koizumi, R. (2016). Current issues in large-scale educational assessment in Japan: Focus on national assessment of academic ability and university entrance examinations. *Assessment in Education: Principles, Policy & Practice*. https://doi.org/10.1080/0969594X.2016.1225667.
Li, C. (2021). *Understanding EAP learners' beliefs about language learning from a socio-cultural perspective*. Singapore: Springer.
Ministry of Education, Culture, Sports, Science and Technology. (2020). Daigaku Nyushi Seido no Genjou to Koudai Setsuzoku Kaikaku no Keii ni tsuite [On the history of the upper secondary school-university articulation reform and current status of college entrance exam, January 15, 2020]. https://www.mext.go.jp/content/20200116-mxt_daigakuc02-000004136_5.pdf. Accessed 11 May 2022.
Ministry of Education,Culture, Sports, Science and Technology (2015). Seito no eiryoku koujou puran [The plan for improving students' English proficiency]. Retrieved June 28, 2022 from https://www.mext.go.jp/a_menu/kokusai/gaikokugo/__icsFiles/afieldfile/2015/07/21/1358906_01_1.pdf
Muthén, B. O. (2004). Mplus Technical Appendices. https://www.statmodel.com/download/techappen.pdf. Accessed 11 May 2022.
Muthén, L. K., & Muthén, B. O. (1998–2022). *Mplus [computer software]*. Los Angeles: Muthén & Muthén.
Nakamura, K. (2014). Examination of possible consequences of a new test within the context of university entrance exam reform in Japan. Paper presented at the 36th Language Testing Research Colloquium, VU University Amsterdam, the Netherlands. http://www.eiken.or.jp/teap/group/pdf/teap_ltrcpresentation20140620.pdf. Accessed 11 May 2022.
Nakatsuhara, F. (2014). A research report on the development of the Test of English for Academic Purposes (TEAP) speaking test for Japanese university entrants—Study 1 & study 2. http://www.eiken.or.jp/teap/group/pdf/teap_speaking_report1.pdf. Accessed 11 May 2022.
Nakatsuhara, F., Joyce, D., & Fouts, T. (2014). A research report on the development of the Test of English for Academic Purposes (TEAP) speaking test for Japanese university entrants—Study 3 & study 4. http://www.eiken.or.jp/teap/group/pdf/teap_speaking_report2.pdf. Accessed 11 May 2022.
National Center for University Entrance Examinations. (2017). Outline of the National Center for University Entrance Examinations. www.dnc.ac.jp/albums/abm00033004.pdf (dnc.ac.jp). Accessed 11 May 2022.
O'Sullivan, B., & Weir, C. (2011). Test development and validation. In B. O'Sullivan (Ed.), *Language testing: theories and practices*, (pp. 13–32). Oxford: Palgrave.
Oller Jr., J. W. (1980). Language testing research (1979–1980). *Annual Review of Applied Linguistics*, *1*, 124–150.
Peacock, M. (2001). Pre-service ESL teachers' beliefs about second language learning: A longitudinal study. *System*, *29*, 177–195.
Powers, D.E., & Powers, A. (2015). The incremental contribution of TOEIC® Listening, Reading, Speaking, and Writing tests to predicting performance on real-life English language tasks. *Language Testing, 32*(2), 151–167. https://doi.org/10.1177/0265532214551855
Ross, S. (1998). Self-assessment in second language testing: a meta-analysis and analysis of experiential factors. Language Testing, 15(1), 1–20.
Runnels, J. (2016). Self-assessment accuracy: correlations between Japanese English learners' self-assessment on the CEFR-Japan's Can do statements and scores on the TOEIC. Taiwan Journal of TESOL, 13(1), 105–137.
Sasaki, M. (1993). Relationships among second language proficiency, foreign language aptitude, and intelligence: A structural equation modeling approach. *Language Learning*, *43*(3), 313–344.
Sasaki, M. (2008). The 150-year history of English language assessment in Japanese education. *Language Testing*, *25*(1), 63–83. https://doi.org/10.1177/0265532207083745.
Sawaki, Y., & Nissan, S. (2009). *Criterion-related validity of the TOEFL iBT listening section (TOEFL iBT Research Report)*. Princeton: ETS.
Sawaki, Y., Stricker, L., & Oranje, H. A. (2009). Factor structure of the TOEFL Internet-based test. Language Testing, 25(1), 53–0. https://doi.org/10.1177/0265532208097335
Taylor, L. (2014). A report on the review of test specifications for the reading and listening papers of the Test of English for Academic Purposes (TEAP) for Japanese university entrants. http://www.eiken.or.jp/teap/group/pdf/teap_rlspecreview_report.pdf. Accessed 11 May 2022.
University of Cambridge Local Examinations Syndicate. (2022). Cambridge English exams. http://www.cambridgeenglish.org/exams/. Accessed 11 May 2022.

Weir, C. (2014). A research report on the development of the Test of English for Academic Purposes (TEAP) writing test for Japanese university entrants. http://www.eiken.or.jp/teap/group/pdf/teap_writing_report.pdf. Accessed 11 May 2022.
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.
Xie, Q. (2015). Do component weighting and testing method affect time management and approaches to test preparation? A study on the washback mechanism. *System*, *50*, 56–68. https://doi.org/10.1016/j.system.2015.03.002.
Xie, Q., & Andrews, S. (2013). Do test design and uses influence test preparation? Testing a model of washback with structural equation modeling. *Language Testing*, *30*(1), 49–70. https://doi.org/10.1177/0265532212442634.

Publisher's Note