Language Testing in Asia

RESEARCH Open Access

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

Nobuhiro Kamiya

Correspondence:
kamiya@fic.gpwu.ac.jp
Department of International Communication, Gunma Prefectural Women's University, 1395-1 Kaminote, Tamamura-machi, Sawa-gun, Gunma 370-1193, Japan

Abstract

Background: The National Center Test is a test that high school graduates take in order to be matriculated into the majority of universities in Japan. Its English section (NCT), though, is scheduled to be replaced by commercially available private English tests of four skills (reading, listening, speaking, writing) due to the fact that NCT currently measures only the receptive skills of reading and listening. A concern has been raised, however, regarding the score comparability between NCT and those tests of four skills. Thus, this study was conducted in order to examine to what extent the score of NCT and TOEFL Junior Comprehensive (JC)—one of the commercially available private English tests of four skills—correlate with each other and to what extent their test constructs overlap from the viewpoint of L2 competence structure.

Methods: One hundred forty-four twelfth graders in Japan took NCT and JC. Pearson's correlations and an exploratory factor analysis (maximum likelihood) were conducted. Moreover, confirmatory factor analyses and chi-square difference tests were performed in order to compare several models.

Results: The results show that the scores of the two tests are highly correlated with each other, indicating that JC can be a proper candidate to replace NCT. Moreover, an exploratory factor analysis revealed that all of the six scores can be subsumed under a single factor. According to the results of a series of confirmatory factor analyses and chi-square difference tests, the correlated skill model was shown to be the best fit model, but the unitary model was almost equally a good fit as well.

Conclusions: The results overall indicate that, regardless of tests and skills to measure, English proficiency as a general construct may determine the major portion of these scores. This poses the question of whether, in order to assess learners' English proficiency, all four skills need to be measured separately.

Keywords: National Center Test, TOEFL Junior Comprehensive, Language testing, L2 competence structures

Background

In Japan, most high school students who wish to be matriculated in universities are required to take a common examination titled the National Center Test for university entrance examinations. Commencing in 1990, its English section (NCT; henceforth) started with a reading section only, to which the listening section was added in 2006. After 20 years of implementation, however, NCT is scheduled to be replaced by commercially available private tests due to the criticism of its incapability to measure four skills (reading, listening, speaking, writing) despite the fact that the ministry's curriculum guidelines stipulate that these skills are expected to be acquired in a balanced manner in English education in school (Ministry of Education 2010).

Traditionally, there were only a handful of English tests in which four skills could be measured; however, following the abovementioned trend, several new tests have recently emerged, and as far as the author knows, there are now over 10 kinds extant as can be seen in Table 1.

Among these, TOEFL Junior Comprehensive (JC; henceforth) was listed as one of the prospective candidates as a replacement for NCT (Eigo yonginou shikaku kentei kondankai 2016). However, a concern has been raised regarding the degree of

Table 1 Commercially available private tests of four skills

Official names	Common names	Organizations
Business Language Testing Service	BULATS	Eiken Foundation of Japan (n.d.-a)
Cambridge English exams	Cambridge English	Cambridge English Language Assessment (2017)
Eiken Test in Practical English Proficiency	Eiken	Eiken Foundation of Japan (n.d.-b)
Global Test of English Communication	GTEC	Benesse (1997–2017a)
Global Test of English Communication Academic	GTEC Academic	Benesse (n.d.-a)
Global Test of English Communication Computer Based Testing	GTEC CBT	Benesse (2000-2017)
Global Test of English Communication Corporate Test Edition	GTEC CTE	Benesse (1997–2017b)
Global Test of English Communication Junior	GTEC Junior	Benesse (n.d.-b)
International English Language Testing System	IELTS	British Council, IDP: IELTS Australia, and Cambridge English Language Assessment (2017)
Pearson Test of English Academic	PTE Academic	Pearson (2014)
Progress	Progress	Pearson (n.d.)
Test of English for Academic Purposes	TEAP	Eiken Foundation of Japan (n.d.-c)
Test of English for Academic Purposes Computer Based Testing	TEAP CBT	Eiken Foundation of Japan (n.d.-d)
Test of English as a Foreign Language Internet-Based Testing	TOEFL iBT	Educational Testing Service (2017a)
Test of English as a Foreign Language Junior Comprehensive	TOEFL Junior Comprehensive	Educational Testing Service (2017b)[a]
Test of English for International Communication	TOEIC (LR & SW)	Educational Testing Service (2017c)
United Nations Associations Test of English	Kokuren Eiken	United Nations Associations of Japan (n.d.)

Note. In alphabetical order. No intention to be an exhaustive list

[a]TOEFL Junior Comprehensive has already been excluded from this page

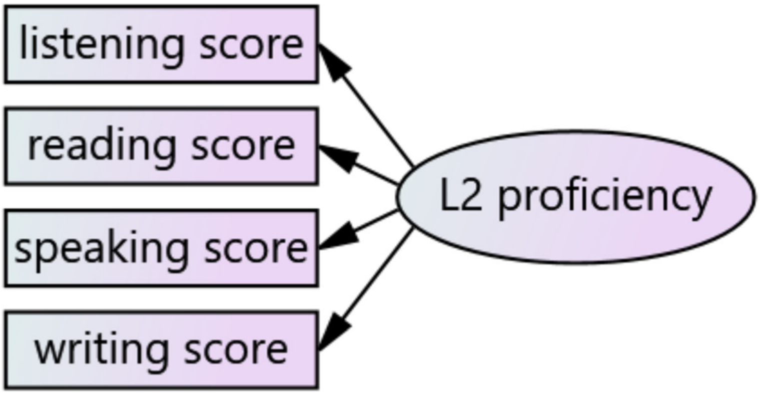

Fig. 1 A sample of unitary model

correspondence in scores between these tests and NCT for three reasons. Firstly, whereas NCT measures only the receptive skills, others additionally measure productive skills. Secondly, whereas NCT strictly follows the guidelines on high school curriculum published by the Ministry of Education, Culture, Sports, Science, and Technology (MEXT; henceforth), others have no such constraints (see In'nami et al. 2016). GTEC CBT, TEAP, and TEAP CBT are developed specifically geared toward Japanese high school students, so their test constructs loosely follow the guidelines; however, GTEC CBT also asks questions beyond them (Benesse 2000–2017), and TEAP (CBT) does not refer to this issue at all (Eiken Foundation of Japan n.d.-c, n.d.-d), implying that their restrictions are not as rigorous as that on NCT. Finally, the aim of these two tests significantly differ from each other: JC aims to "measure the full range of language uses that students encounter in English-medium school settings" (So et al. 2015, p. 4) whereas that of NCT is to "improve the selection of candidates for admission to Japanese universities, and thereby to promote education at universities, high schools, and other educational institutions" (National Center for University Entrance Examinations 2015, p. 1). These three issues pose a question of whether not only the total scores but also the test constructs are comparable between them. Therefore, the current study was conducted with the aim to examine the concurrent validity of NCT and JC, namely,

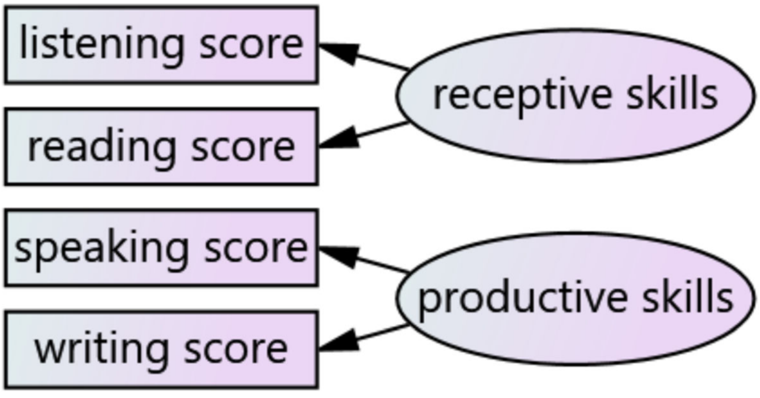

Fig. 2 A sample of uncorrelated model

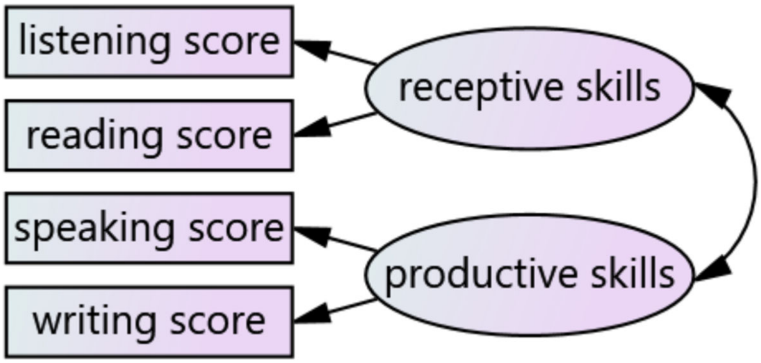

Fig. 3 A sample of correlated model

to what extent the scores of these two tests correlate with each other and to what extent the test constructs of these tests overlap from the viewpoint of L2 competence structure.

Structure of L2 competence

Language tests are supposed to represent L2 competence that learners have. If a test measures a single dimension of L2 competence, the test score (an observed variable) should be subsumed under a single factor of L2 competence (an unobserved latent variable). For instance, a score from a speaking test may have a one-on-one connection with a factor of speaking skills. However, if multiple tests measure a single dimension, all of them should be subsumed under a single factor of L2 competence. For example, the scores of two distinct listening tests may be measuring a single construct of listening skills. By examining the scores of various tests taken by the same test takers via factor analysis, researchers have sought to discover the structure of L2 competence. Thus far, four major models have been proposed:

1. A unitary or unidimensional model: L2 learners have only a single construct of L2 competence, which can explain all of test scores, regardless of skills or knowledge to be measured. For example, all of the test scores of four skills are subsumed under a single factor of L2 proficiency (Fig. 1; squares represent observed variables whereas circles represent unobserved latent variables, henceforth).

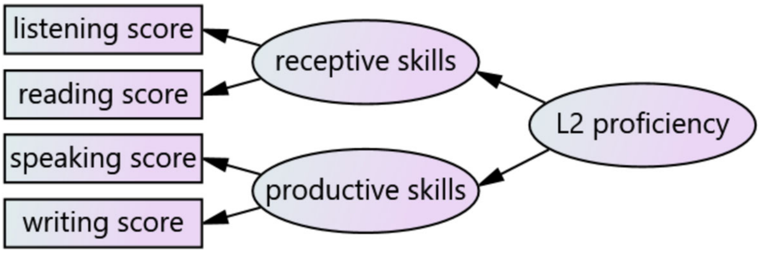

Fig. 4 A sample of higher-order model

Table 2 Structure of NCT

Sections	No. of questions	Testing time (min)	Raw scores
Reading	55	80	0–200
Listening	25	30	0–50
Total	80	110	0–250

Source: National Center for University Entrance Examinations (2015)

2. An uncorrelated model: Rather than one, multiple constructs exist among L2 learners and each of them separately subsumes the test scores purported to measure the same construct, but those factors are uncorrelated. For example, the scores of the listening test and reading test are subsumed under a factor of receptive skills whereas those of the speaking test and writing test are subsumed under a factor of productive skills, but the two factors of receptive skills and productive skills are unrelated and, thus, not correlated with each other (Fig. 2). This validates such a claim that receptive skills and productive skills need to be measured separately on the assumption that each of them develops independently.
3. A correlated model: Similar to the uncorrelated model, multiple constructs exist among L2 learners; however, the unobserved latent factors are correlated with each other. For example, the two factors of receptive skills and productive skills are connected with a line with arrowheads at both ends (Fig. 3). This also validates such a claim that receptive skills and productive skills need to be measured separately because, although they develop in tandem rather than independently, their development is still not perfectly synchronized.
4. A higher-order, second-order, or hierarchical model: Similar to the correlated model, multiple constructs exist among L2 learners; however, instead of directly connecting those factors, they are subsumed under another unobserved latent factor placed on the higher level. For example, the two factors of receptive skills and productive skills (first-order factors) are connected with another factor of L2 proficiency newly added on the right (Fig. 4; please note that the lines are single arrow-headed this time). This model also validates such a claim that receptive skills and productive skills need to be measured separately because, although they are both affected by another higher factor, how the factor influences each of them is not related to each other; thus, their development is not perfectly synchronized either.

Although the past literature has not reached an agreement concerning which model best explains L2 competence, at least, the unitary model and the uncorrelated model

Table 3 Structure of JC

Sections	No. of questions	Testing time (min)	Raw scores
Reading	36	41	140–160
Listening	36	36	140–160
Speaking	4	18	0–16
Writing	4	39	0–16
Total	80	134	

Source: GC&T (n.d.)

Table 4 Descriptive statistics of two NCTs on the national level and the number of participants

Years	Reading (/200)	Listening (/50)	Total (/250)	No. of participants
2016	112.4 (42.2)	30.8 (9.4)	143.2	64
2017	123.7 (45.0)	28.1 (10.2)	151.8	80

Figures in parentheses represent standard deviations. Source: National Center for University Entrance Examinations (2016, 2017)

have been scantly supported thus far. As one of the proponents of the former, Oller (1979, 1983) reported a series of proofs that a single trait (Oller himself does not define what the trait is in a clear and consistent manner though) could explain the major portions (over 70%) of the total variance of various tests, including TOEFL PBT. He goes so far as to say, "the current practice of many ESL programs, textbooks, and curricula of separating listening, speaking, and reading and writing activities is probably not just pointless but in fact detrimental" (p. 458). However, his argument has been refuted, partly owing to the fact that he used principal component analysis (see Fouly et al. 1990, for the criticism), which prompted studies onward to adopt confirmatory factor analysis and exploratory factor analysis with the Schmid-Leiman procedure instead, resulting in negating at least the strong version of the unitary model.

Most of the literature since the 1980s advocates adopting either the correlated model or the higher-order model, which are statistically indistinguishable from each other when the number of first-order factors is three (for examples, see Fouly et al. 1990; Sasaki 1996; Shin 2005). Fouly et al. (1990) examined the scores of various tests, including TOEFL PBT, collected from 334 university students in the USA with diverse L1s, and found that the three-correlated-factor model and the higher-order model were equally well fit. Similarly, Sasaki (1996) analyzed scores of several English tests taken by 160 Japanese university students and concluded that both the higher-order model and the correlated model equally fit. Shin (2005) investigated scores of TOEFL PBT and SPEAK (Speaking Proficiency in English Assessment Kit) taken by 738 participants divided into three different levels based on their English proficiency, which resulted in the higher-order model with three first-order factors of listening, speaking, and writing modes being the best model across the three groups, i.e., regardless of their proficiency. Investigating to what extent two distinct assessments measured the same language ability construct taken by 1224 English learners in the fourth grade in California, Llosa (2007) also found a model with three first-order factors (listening and speaking, reading, writing) under a higher-order factor (English proficiency) to be the best fit. More recently, In'nami et al. (2016) analyzed the factor structure of TEAP and compared it to that of TOEFL iBT. The higher-order model was found to be the best fit for TEAP, and also, the test constructs were similar between these two tests, both measuring academic English skills.

Table 5 Score conversions of JC

Sections	Raw scores (X)	Converting equations	Converted scores
Reading	140–160	$X - 140$	0–20
Listening	140–160	$X - 140$	0–20
Speaking	0–16	$X/4 \times 5$	0–20
Writing	0–16	$X/4 \times 5$	0–20
Total			0–80

Table 6 Descriptive statistics of NCT and JC

	NCT			JC				
	Reading	Listening	Total	Reading	Listening	Speaking	Writing	Total
Average	141.05	36.69	177.74	151.97	149.40	7.97	7.27	40.41
SD	42.19	9.25	49.83	5.12	5.20	2.72	2.70	14.75
Max	200	50	248	160	160	13	13	70
Min	30	10	52	140	140	0	1	5
Skewness	−0.82	−0.70	−0.79	−0.78	−0.09	−0.55	−0.11	−0.41
Kurtosis	−0.17	−0.42	−0.23	−0.25	−0.78	0.05	−0.53	−0.47

SD Standard deviation

When the number of first-order factor is as few as two, the higher-order factor is not even susceptible to verification, and a decision is often made based on some other criteria. As one of the earliest studies that used confirmatory factor analysis, Bachman and Palmer (1981) compared over 20 models using the data of speaking and reading collected via three distinct methods for each from 75 Mandarin native speakers aged between 19 and 35 residing in the USA. They found that the correlated model with two factors and the higher-order model were equally well fit but, based on the fact that the former is simpler with one less factor, concluded the former to be the best choice. Another study conducted by Bachman and Palmer (1982) examined the three traits of communicative competence (grammatical, pragmatic, and sociolinguistic) collected via four methods with 116 participants aged between 17 and 67 with various L1s residing in the USA and reached the conclusion that the higher-order model is superior to the correlated model based on the results of chi-square tests (However, see Fouly et al. (1990), for criticism on Bachman and Palmer’s studies.). More recently, In’nami and Koizumi (2012) analyzed TOEIC scores of four listening and five reading subskills collected from 569 EFL learners, most of whom were Japanese university students. Among several models that they tested, they found the correlated two-factor model (listening and reading) to be the best fit. Due to the identification problem, the higher-order model was not considered.

Somewhat more pertinent to the present study in terms of the same TOEFL tests of four skills, a series of large-scale studies with 2720 participants have been conducted to explore the factor structure of TOEFL iBT (Sawaki et al. 2008, 2009; Stricker and Rock 2008). They commonly adopted the higher-order model with four first-order factors of

Table 7 Results of correlation matrix

		NCT		JC				
		Listening	Total	Reading	Listening	Speaking	Writing	Total
NCT	Reading	.79	.99	.78	.75	.64	.68	.84
	Listening	–	.86	.71	.75	.59	.70	.81
	Total		–	.79	.78	.65	.71	.86
JC	Reading			–	.71	.58	.64	.88
	Listening				–	.64	.68	.90
	Speaking					–	.64	.81
	Writing						–	.84

All figures are significant at $p < .001$

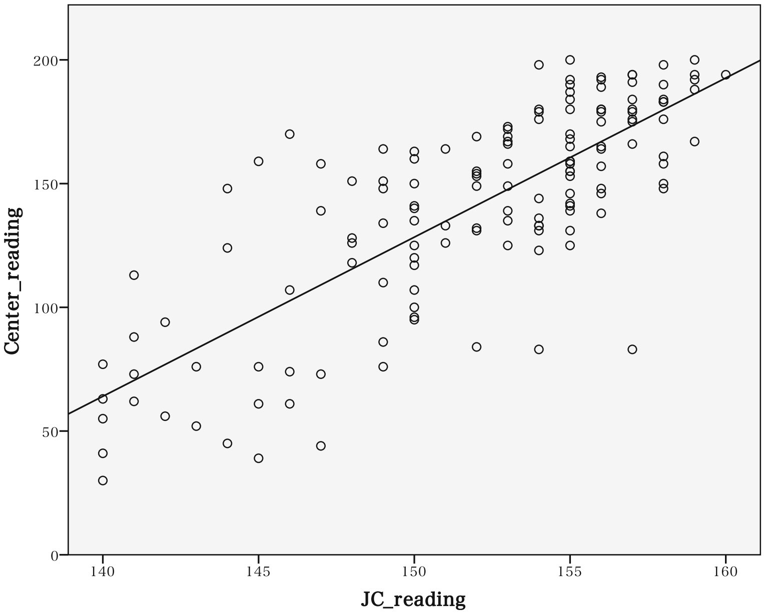

Fig. 5 Scatter plot of reading scores

four skills, which is in line with the TOEFL score convention to report each skill-based score along with a single composite score. Lastly, Gu (2015) is perhaps the only study that has examined the factor structure of JC (although its pilot form), in which a total of 436 EFL learners aged 11–15 took the test. Using item-level raw scores, three models were compared: the unitary model, the correlated four-factor model (reading, listening, speaking, writing), and the higher-order model with those four first-order factors. The results show that the higher-order model was the best fit with a high-order factor being assumed as general learning ability.

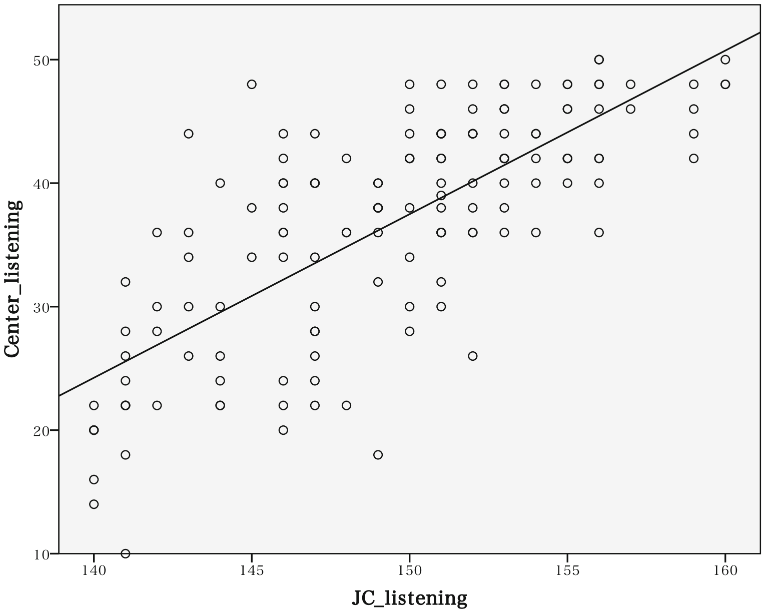

Fig. 6 Scatter plot of listening scores

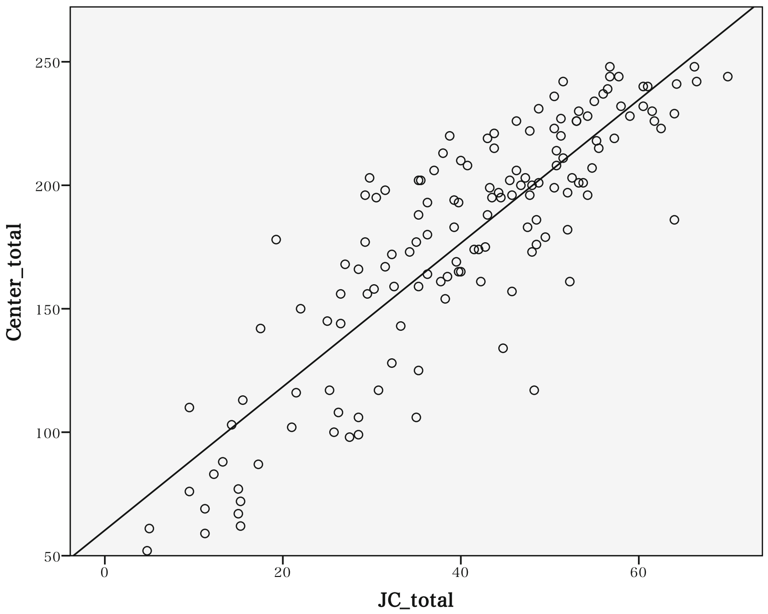

Fig. 7 Scatter plot of total scores

All in all, the accumulation of the past literature leads to the hypothesis that the structure of L2 competence is most likely to be either a correlated model with several first-order factors or a higher-order model subsuming them.

National Center Test

All of the Japanese high school students and graduates who wish to be matriculated in public universities and most private universities take the same test concurrently on the same day nationwide once a year in January. Among over 1100 four-year and junior colleges in Japan, around 850, including all of the national and public universities, currently require applicants to take the test (National Center for University Entrance Examinations n.d.-a), and over 550,000 applicants, comprising over a half of new high school graduates, take the multiple-choice tests of various subjects (National Center for University Entrance Examinations n.d.-b). The test items are made strictly in accordance with the guidelines on the high school curriculum stipulated by MEXT.

NCT consists of reading and listening sections (Table 2). Test booklets are printed and distributed to test takers. For the listening section, IC audio recorders are distributed to test takers, each of whom listens to the audio via earphones individually. Test takers complete multiple-choice questions by marking their answers on answer sheets.

As far as the author knows, there are only few studies that investigated the correlations of scores between NCT and other commercially available private tests (Eiken

Table 8 Results of exploratory factor analysis

	NCT		JC			
	Reading	Listening	Reading	Listening	Speaking	Writing
Communalities	0.81	0.75	0.70	0.73	0.52	0.63
Factor loadings	0.90	0.87	0.85	0.84	0.79	0.72

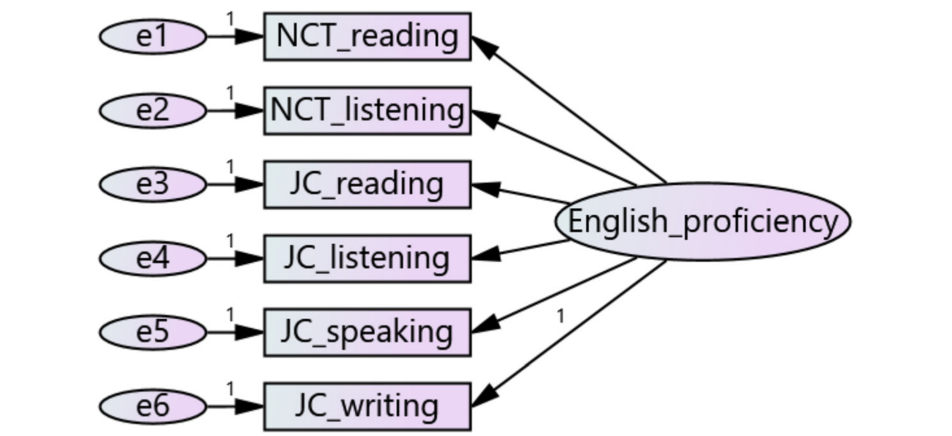

Fig. 8 Unitary model

Foundation of Japan 2015; Otsu 2013, 2014; Yamanishi 2001). For instance, Otsu (2014) analyzed the relationships of scores between TOEIC, TOEFL PBT, and NCT and found that its correlation was 0.80 between TOEIC and NCT and 0.70 between TOEFL and NCT. However, the number of participants was as few as 63 and 28, respectively, and their TOEIC and TOEFL scores were self-reported. Besides, the time lag between when they took those tests and NCT was not controlled for. Thus, the reliability of the scores is not ensured. Also, the English proficiency of the participants was skewed to advanced level, so they may not represent the typical sample. Remedying these shortcomings, Eiken Foundation of Japan (2015) analyzed the data in a more robust procedure. As many as around 1000 participants took TEAP and Eiken 3 or 4 months after they took NCT. The study found that the correlation between NCT and TEAP is 0.80 and that between NCT and Eiken is 0.89. However, none of the tests examined in those abovementioned studies (i.e., TOEIC, TOEFL PBT, TEAP, and Eiken) contained speaking or writing scores even if the test had those sections; thus, it is unknown to what extent the scores derived from the accumulation of four skills correlates with NCT scores, which is currently an issue of more importance. Finally, as far as the author knows, the factor structure of NCT has not been examined in any of the previous studies thus far.

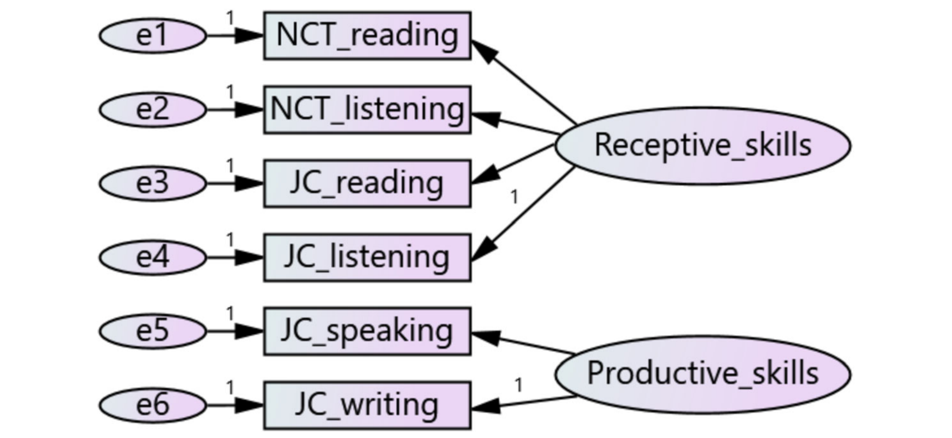

Fig. 9 Uncorrelated skill model

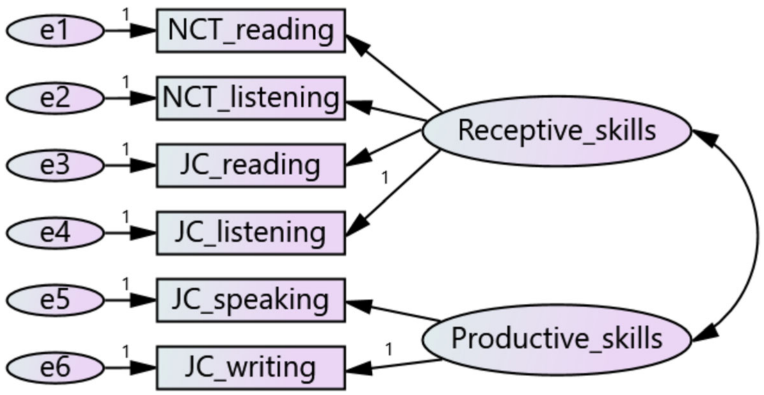

Fig. 10 Correlated skill model

TOEFL Junior Comprehensive

JC was created by the Educational Testing Service (ETS) and first implemented in 2014. Primarily, the test was targeted toward a group of English learners who find TOEFL iBT too advanced, and thus, it was "developed to address the increasing need for objective measures of English language proficiency for young adolescent learners, who are being introduced to English as a second or foreign language at a much younger age than ever before" (So et al. 2015, p. 1), measuring "how successfully a test taker can complete test tasks that are designed to represent the range of communicative tasks encountered in English-medium middle schools" (p. 6). Because it was intended to be implemented worldwide and also created by the Education Testing Service in the USA, it had no restrictions imposed by the guidelines on the high school curriculum stipulated by MEXT. It consisted of four sections: reading, listening, speaking, and writing (Table 3). As can be seen in Table 3, the score bands of reading and listening are quite distinct from those of speaking and writing, which So et al. (2015) explain as follows:

> Because the speaking and writing section scores will not be equated, the scores are not strictly comparable … To avoid any incorrect impression on the part of stakeholders that the speaking and writing scores are comparable across forms,

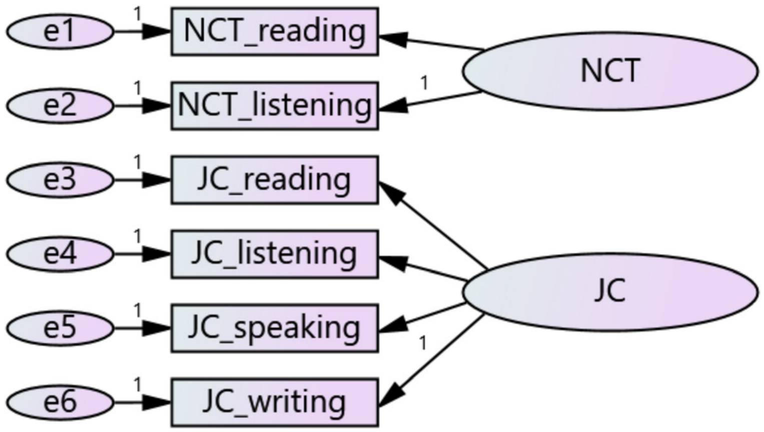

Fig. 11 Uncorrelated test model

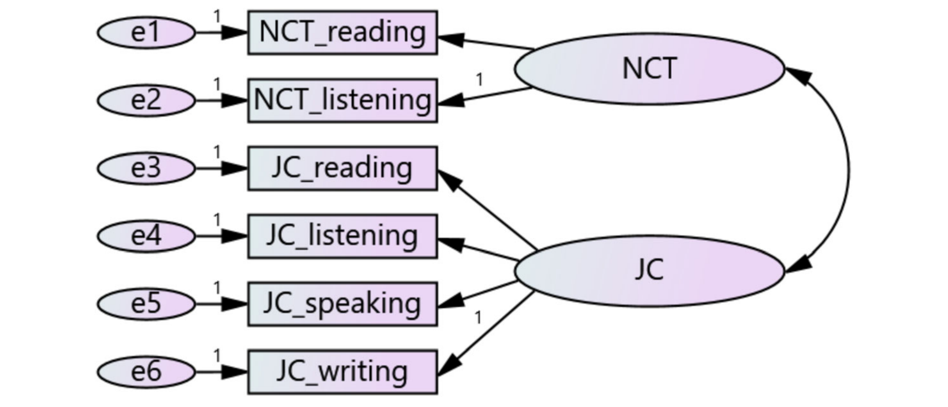

Fig. 12 Correlated test model

> as the reading and listening section scores are, it was decided that the speaking and writing scales would be made clearly distinguishable from the reading and listening scales. (p. 17)

Following this concept, instead of total scaled scores, overall score levels compatible with CEFR (see Tannenbaum and Baron 2015, for details) are reported in JC ranging from 1 to 6 (see Papageorgiou et al. 2015, for details). The use of different scoring scales for receptive skills (reading and listening) and productive skills (speaking and writing) imply that at least these two constructs need to be separately posited in the model.

The test is conducted via computers, and all of the questions are displayed on a computer screen. Some of the test items in the speaking and writing sections are integrated with listening. Unfortunately, the test became defunct at the end of 2016.[1,2] Although GC&T (Global Communication & Testing), the organization that operated JC in Japan, has a single data set that explored the relationship between JC scores and NCT scores, they are not publicized, thus inaccessible (personal communication, June 23, 2017).

Research questions

All in all, the present study sought to answer the following two research questions.

1. How are the scores of NCT and JC correlated with each other?
2. Which of the five models (unitary, uncorrelated skill, correlated skill, uncorrelated test, and correlated test) best represents the test constructs of NCT and JC?

Table 9 Results of model fits

	χ^2	*df*	χ^2/df	CFI	NFI	TLI	AIC	CAIC	RMSEA [CI]	SRMR
Unitary	11.921	9	1.325	0.995	0.982	0.992	35.921	85.559	0.048 [0, 0.112]	0.0211
Correlated skill	7.882	8	0.985	1.000	0.988	1.000	33.882	85.489	0 [0, 0.097]	0.0154
Correlated test	11.175	8	1.397	0.995	0.983	0.991	37.175	88.783	0.053 [0, 0.119]	0.0201

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

Table 10 Unstandardized regression weights of unitary model

	Estimate	SE	CR	*p*
NCT reading <– English proficiency	17.798	1.428	12.461	<.001
NCT listening <– English proficiency	3.770	0.317	11.896	<.001
JC reading <– English proficiency	2.013	0.178	11.321	<.001
JC listening <– English proficiency	2.083	0.179	11.621	<.001
JC speaking <– English proficiency	0.921	0.099	9.337	<.001
JC writing <– English proficiency	1.000			

SE standard error, *CR* critical ratio

Method

A total of 144 twelfth graders (F = 114, M = 30) in nine Japanese high schools took part in this study. All of them were native speakers of Japanese. Efforts were made to recruit students with diverse L2 proficiencies in order to secure a high reliability of data (Mizumoto 2014). They all officially took NCT in January of either 2016 or 2017. Their scores were validated by the official score reports sent by the National Center for University Entrance Examinations. The total scores were calculated simply by adding the scores of reading and listening as it is the conventional way. Their descriptive statistics on the national level along with the number of participants are shown in Table 4. Due to the relatively small discrepancies, no score adjustments were implemented between the scores of these 2 years. The participants also took JC in either March or May after an interval of 2–4 months following NCT. JC reports a composite score in addition to each of the four skills; however, the composite score ranges only from 1 to 6. Such a rough scoring with a small scale inevitably incapacitates the analysis from being fine-grained, which likely masks the precise relationship between the variables. Thus, a decision was made to calculate a total score that better reflects score variations. Because JC reports two different score bands for skills, in order to equalize their weights, the raw scores were converted to the scale of 0–20 (Table 5). The scale of 0–20 was adopted in order to utilize the original range of the raw scores as much as possible (20 for reading and listening and 16 for speaking and writing) because using a smaller range will conceal the difference in the raw scores. For analyses, Pearson’s correlations, an exploratory factor analysis (maximum likelihood), confirmatory factor analyses, and chi-square difference tests were conducted.

Table 11 Unstandardized regression weights of correlated skill model

	Estimate	SE	CR	*p*
NCT reading <– receptive skills	8.590	0.593	14.483	<.001
NCT listening <– receptive skills	1.816	0.134	13.585	<.001
JC reading <– receptive skills	0.970	0.076	12.773	<.001
JC listening <– receptive skills	1.000			<.001
JC speaking <– productive skills	0.916	0.092	9.958	<.001
JC writing <– productive skills	1.000			

SE standard error, *CR* critical ratio

Table 12 Unstandardized regression weights of correlated test model

	Estimate	SE	CR	*p*
NCT reading <− NCT	4.729	0.311	15.206	<.001
NCT listening <− NCT	1.000			
JC reading <− JC	2.007	0.177	11.363	<.001
JC listening <− JC	2.084	0.178	11.715	<.001
JC speaking <− JC	0.923	0.098	9.418	<.001
JC writing <− JC	1.000			

SE standard error, *CR* critical ratio

Results

Descriptive statistics and correlations

Table 6 shows descriptive statistics of the results of the two tests. A wide score range was observed for each section of both tests. As can be seen in Table 7, the Pearson's product-moment correlation matrix indicates that the two tests are highly correlated with each other not only for the scores of identical skills (reading: r = .78; listening: r = .75; Figs. 5 and 6, respectively) but also for the total scores (r = .86; Fig. 7), indicating a high concurrent validity between them. Excluding total scores, the lowest figure was observed in the combination between NCT listening and JC speaking at 0.59 and the highest was between the two readings at 0.78.

An exploratory factor analysis

Because all of the six scores were highly correlated, an exploratory factor analysis was conducted in order to detect the number of latent factors, using raw scores. Although strong correlations were observed between reading and total scores of NCT (r = .99) and between listening and total scores of JC (r = .90), since total scores were excluded for factor analysis, none of the variables tested in a factor analysis had a correlation of over 0.8. Also, the determinant was 0.01. Thus, multicollinearity was not at issue. A Bartlett's test was significant at p < .001, indicating that it is significantly different from an identity matrix. The value of Kaiser-Meyer-Olkin statistic was 0.917, verifying the sampling adequacy. No univariate or multivariate outlier was detected using z scores and Mahalanobis distance (max = 20.309), respectively. All of the values for skewness and kurtosis fall within the range of |3.30| (z score at p < .01), securing univariate normality (Tabachnick and Fidell 2007). Mardia's normalized estimate of multivariate kurtosis was 0.871, which is well below 5, securing multivariate normality (Byrne 2016).

Maximum likelihood extracted a single factor, explaining 68.97% of total variance. Furthermore, as can be seen in Table 8, all of the communalities were over 0.50, and all of the factor loadings were over 0.70. Judging from these figures, the reading score of NCT represents the largest portion of the factor.

Confirmatory factor analyses

The results of the exploratory factor analysis imply that, regardless of tests and skills to measure, a general construct, supposedly general English ability, determined the major portion of these scores, which supports the unitary model. Thus, in order to verify whether the unitary model would show the best fit over others, confirmatory factor

Table 13 Variances of unitary model

	Estimate	SE	CR	*p*
English proficiency	4.506	0.807	5.587	<.001
e6	2.691	0.359	7.500	<.001
e5	3.518	0.449	7.840	<.001
e4	7.277	1.053	6.909	<.001
e3	7.785	1.095	7.109	<.001
e2	20.878	3.124	6.684	<.001
e1	340.378	56.436	6.031	<.001

SE standard error, *CR* critical ratio

analyses were conducted, comparing three models: unitary (Fig. 8), two uncorrelated latent factors of receptive and productive skills (Fig. 9), and two correlated latent factors of receptive and productive skills (Fig. 10). Moreover, although high correlations were observed between these two tests, each of them was made for different aims and, thus, may have divergent test constructs from each other. Thus, for the sake of comparison with the three models above, two additional models, two uncorrelated latent factors of NCT and JC (Fig. 11), and two correlated latent factors of NCT and JC (Fig. 12) were analyzed. The higher-order model was not considered due to the fact that the number of first-order factors was two.

First of all, the two uncorrelated models (Figs. 9 and 11) were rejected as they were unidentified. Although this issue was solved by setting the weight of another regression path as one, such a treatment is not theoretically supported, and additionally, the models were found to be a poor fit (p < .001). Thus, these models are not considered further.

The results of the remaining three models are shown in Tables 9, 10, 11, 12, 13, 14, 15, and 16 and Figs. 13, 14, and 15. Table 9 indicates that both the unitary and the correlated skill models are satisfactorily well fit in terms of all of the indices (i.e., χ^2/df. < 1.5, CFI > 0.95, NFI > 0.95, TLI > 0.95, RMSEA < 0.05, SRMR < 0.05). Besides, the differences in AIC and CAIC are quite small and the lowest value in the confidence interval of RMSEA was 0, indicating that neither of the two models is significantly better than the other. Thus, because the unitary model is nested in the correlated skill model, a chi-square difference test was performed, whose results show that the latter was slightly better than the former (χ^2 difference = 4.039, *df* difference = 1, p = .044). The correlation between the two latent factors of receptive skills and productive skills was 0.93.

Table 9 also shows that the correlated test model was also as well fit as the unitary model for all of the indices except RMSEA at 0.053 > 0.05. A chi-square difference test indicates that the degree of fit of these two models are not significantly different from each other (χ^2 difference = 0.746, *df* difference = 1, p = .388). The correlation between the two latent factors of NCT and JC was 0.98.

Discussion

RQ1: How are the scores of NCT and JC correlated with each other?

The correlation of scores between NCT and JC turned out to be all high. The high correlation of the total scores (r = .86) indicates that JC could have been a proper candidate to replace NCT for their total scores (however, see In'nami et al., 2016, for a criticism of such a claim). This figure is somewhat compatible with those found in past

Table 14 Variances of correlated skill model

	Estimate	SE	CR	*p*
Receptive skills	19.495	3.113	6.261	<.001
Productive skills	5.037	0.874	5.762	<.001
e6	2.160	0.413	5.233	<.001
e5	3.114	0.458	6.802	<.001
e4	7.327	1.065	6.882	<.001
e3	7.684	1.090	7.047	<.001
e2	20.650	3.125	6.607	<.001
e1	329.335	56.216	5.858	<.001

SE standard error, *CR* critical ratio

studies that investigated correlations between NCT scores and other commercially available tests (r = .70 for TOEFL, r = .80 for TOEIC and TEAP, r = .89 for Eiken) (Eiken Foundation of Japan 2015; Otsu 2014). These results imply that whether the test follows the restriction of the guidelines on high school curriculum published by the MEXT exerts little impact on the test results. However, the unique contribution of the present study lies in the fact that this is perhaps the first study that proved that this applies to a commercially available test of not only two receptive skills but also those measuring four skills. Unfortunately, JC has already stopped its operation and, thus, cannot be a replacement for NCT. Still, the high compatibility between NCT and JC scores may be a good omen for other four-skill tests without the curriculum restriction to be named as prime candidates to replace NCT. To be cautious, however, such verifications need to be empirically tested in future studies. That is partly because the difficulty of those tests may not be compatible from the perspective of preparation. It is relatively easy to prepare for NCT because (a) the test content falls within the scope of curriculum guidelines, which limits the coverage of what to study; (b) a large number of exercise books are constantly published; and (c) the test becomes open to the public once the test is over every year, and the accumulation of the questions used in the past makes it easy to anticipate the types of questions to be asked. In contrast, those other tests of four skills lack all or at least one of these three points. Perhaps the test with the least disadvantage is Eiken because numerous exercise books are published, and it publicizes the questions used on the website; still, it is without the restriction of the curriculum guidelines, making it rather difficult to anticipate what to study for the test. All

Table 15 Variances of correlated test model

	Estimate	SE	CR	*p*
NCT	64.765	9.987	6.485	<.001
JC	4.543	0.810	5.606	<.001
e6	2.654	0.360	7.378	<.001
e5	3.469	0.447	7.764	<.001
e4	7.098	1.067	6.653	<.001
e3	7.733	1.115	6.938	<.001
e2	20.158	3.199	6.301	<.001
e1	319.622	60.869	5.251	<.001

SE standard error, *CR* critical ratio in all, more studies are expected to be conducted to see the compatibilities of the scores of those tests and NCT.

RQ2: Which of the five models (unitary, uncorrelated skill, correlated skill, uncorrelated test, and correlated test) best represents the test constructs of NCT and JC?

In the present study, the unitary model and the correlated test model were equally well fit according to the chi-square difference test. However, when two models are equally well fit, the model should be chosen based upon the principle of parsimony. That is to say, the simpler model with more degrees of freedom (unitary model) should be preferred over a more saturated model with fewer degrees of freedom (correlated test model). Together with the fact that the RMSEA value was over 0.05 in the correlated test model, this model should be rejected. This indicates that these two tests are not only similar but also even close to being identical in terms of their test constructs; thus, they do not warrant being distinguished from each other by test types. The high correlation between the two factors of tests (r = .98) in the model also corroborates this decision because previous studies adopted the correlation of over 0.9 as a criterion to detect extremely highly correlated factors as they are statistically indistinguishable. In fact, based on this notion, Gu (2015) discarded the correlated four-factor model.

Furthermore, the two models, the unitary model and the correlated skill model, were equally well fit. The chi-square difference test showed the latter was marginally better than the former if the conventional alpha level of 0.05 is adopted. Actually, some of the figures in Table 9 indicate the superiority of the latter (e.g., CFI = 0.995 vs 1, TLI = 0.992 vs 1, RMSEA = 0.048 vs 0). However, because the P value was only slightly lower than 0.05 (p = .044), if the more restricted alpha level is adopted (even at 0.04), both models fit equally well. Then, following the principle of parsimony, the unitary model should be chosen over the correlated skill model. Additionally, the two latent factors are estimated to be correlated at as high as 0.93, and for the abovementioned reason, these two factors are statistically indistinguishable.

It should be noted that, in In'nami et al. (2016), the unitary model was found to be the best fit both for TEAP and TOEFL iBT when the endogenous (observed) variables were skill-based scores: reading, listening, speaking, and writing of TEAP and TOEFL iBT, and claimed that this evidence indirectly supported the higher-order structures. However, an important difference should be noted between their study and the current one. At least for TEAP, In'nami et al. rejected the correlated model with two first-order factors of receptive skills and productive skills due to the low TLI value (0.718). On the contrary, in the current study, it was exactly this model that was found to be the best choice (TLI = 1), which indirectly negates the possibility of the higher-order model.

Thus, depending upon what perspective to take, either the unitary model or the correlated skill model can be the best candidate. If the correlated skill model is to be adopted in line with numerous previous studies that examined the factor structure of language competence, this study suggests that at least receptive skills and productive skills warrant being measured separately. (Due to the fact that a single variable was obtained for speaking and writing, this study does not avail to claim that each of these two skills should be measured separately.) However, if the unitary model is to be adopted, this is quite an unexpected result because the model has been almost uniformly rejected in most of the past literature on empirical grounds. This study is most likely the first study to analyze the factor structure of NCT, so there is no study to refer

Table 16 Squared multiple correlations

	Unitary	Correlated skill	Correlated test
NCT reading	0.807	0.814	0.819
NCT listening	0.754	0.757	0.763
JC reading	0.701	0.705	0.703
JC listening	0.729	0.727	0.735
JC speaking	0.521	0.576	0.527
JC writing	0.626	0.700	0.631

to; however, at least one study, Gu (2015), examined that of JC, so it warrants comparing with the current study.

Table 17 shows some contrastive features between these two studies. Regarding participants' ages, Gu's study is perhaps the study targeting one of the youngest cohorts in this line of studies, but the results were in accordance with other studies that targeted older learners, typically university students. Furthermore, ages are often intertwined with proficiency, but Shin (2005) negated its effects on the model choice. Thus, neither age nor proficiency can account for the discrepancy. Next, although the participants in both Gu (2015) and the present study were learning English as a foreign language, the former come from 15 different countries, over a half of them were Koreans, whereas all of the learners in the present study were Japanese, which could potentially be a culprit. However, Sasaki (1996), targeting only Japanese university students, also rejected the unitary model outright as it was poorly fit. Finally, according to Gu's (2015) descriptions, the pilot version and the authentic version of JC have some differences; however, assumingly, which are rather minor. To summarize, all of these abovementioned scenarios must be ruled out.

Gu (2015) speculated that those test items that require integrative skills to tackle may load on multiple factors, tested a model that allowed cross-loadings, but found that this model was not tenable as all of the loadings on the secondary factors turned out to be non-significant. Such a verification was possible because the analysis was done on the item level. Here, although speculatively, lies the reason for the discrepancy in the models to be adopted in the two studies. Perhaps, on each item level, the loading on the secondary factor is too weak to be significant, but as a composite score of a few

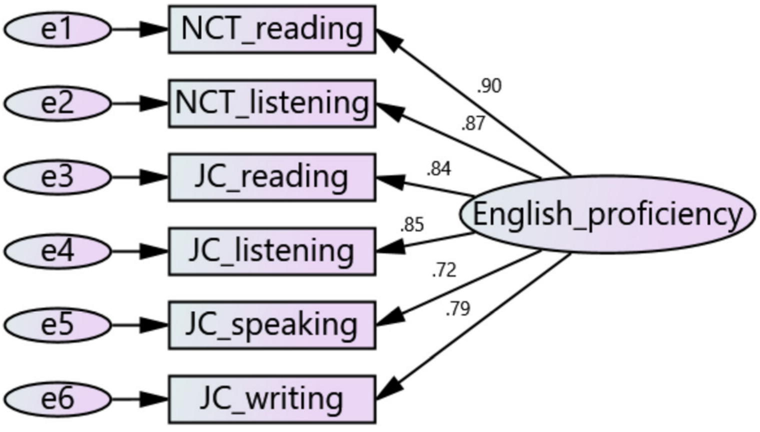

Fig. 13 Standardized regression weights of unitary model

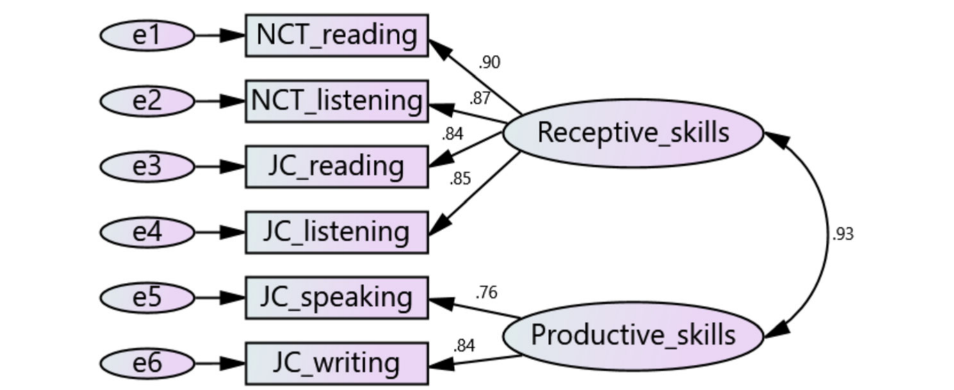

Fig. 14 Standardized regression weights of correlated skill model

items on multiple factors, their effects may have become significant, which may explain why the correlations among the first-order factors of the four skills were all highly correlated at around 0.9, and thus, the correlated model was consequently discarded in Gu's study. It should be noted that, in contrast to Gu's study, the present study adopted even broader categories of two factors for the first order: receptive skills and productive skills, and this practice even likely facilitates blurring the boundaries between first-order factors. In JC, some of the test items in the speaking and writing sections are integrated with listening. Thus, the receptive skill (of listening) is also measured in productive skills (of speaking and writing). Unfortunately, the uncorrelated/correlated models in which receptive skills are also loaded on JC speaking and writing were unidentified in this study. However, setting the regression weights from productive skills to both JC speaking and writing at 1 in the uncorrelated model, the estimated value from receptive skills to JC speaking and to JC writing turned out to be significant at 0.435 (p < .001) and 0.475 (p < .001), respectively. Thus, although tentatively, the abovementioned hypothesis is supported.

As a final note, Alderson (1991) claims that, when exploring the nature of language proficiency, a homogeneous group should be recruited because test takers with a variety of background of ethnics, cultures, languages, and education may yield different outcomes, leading to confounding results. Sasaki (1996) also posits that learning contexts may exert on disproportionate development of each skill. The present study

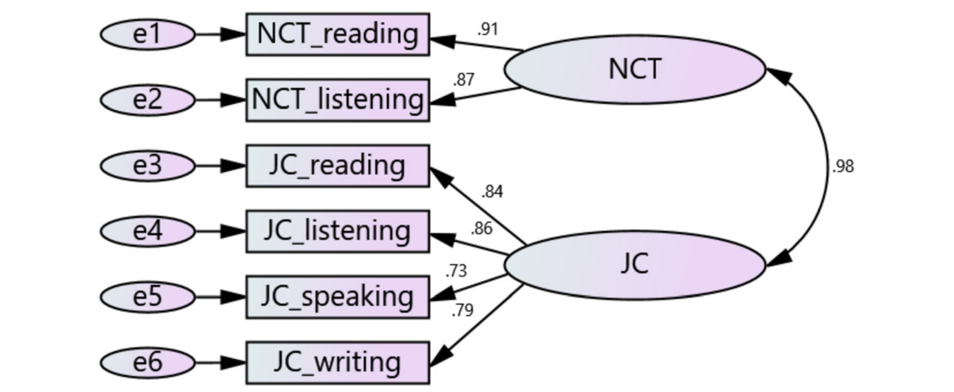

Fig. 15 Standardized regression weights of correlated test model

Table 17 Comparison between Gu (2015) and the present study

	Gu (2015)	Present study
Participants		
Numbers	436	144
Ages	11–15	17–19
Learning environments	EFL	EFL
Native countries	15 different countries	Japan
Tests formats	Pilot	Authentic
Analysis levels	Item raw scores	Total skill scores

followed Alderson's (1991) advice, but this practice, of course, comes at the expense of generalizability; thus, the results of the present study cannot be generalized beyond Japanese high school students. Then, it is somewhat surprising that the unitary model could be adopted for them because, generally speaking, English education in Japan tends to emphasize receptive skills over productive skills partly owing to the fact that the NCT measures only the former. However, the results of the present study indicate that even productive skills develop relatively evenly with receptive skills among Japanese high school students, which is in line with Sasaki (1996). All in all, this study showed that what Sasaki found among Japanese university students may be applicable even among Japanese high school students.

Conclusions

Assuming that all of the six sections in the two tests are measuring a single construct, a question one could ask is, in order to assess learners' English proficiency, do receptive skills and productive skills need to be measured separately? Since the communality of NCT reading was 0.81, 81% of NCT reading scores can be explained by a single factor. JC speaking had the lowest communality, but still, it was over a half (0.52). Then, it is possible to claim that, by assessing learners only through NCT reading, we can quite confidently surmise their English proficiency.

Nevertheless, this paper was written without any intention to propose that only reading should be assessed for university entrance examinations in Japan. Such a practice will naturally lead English classes to heavily focus on reading only, possibly via the Grammar Translation Method. In order to foster communicative skills in a balanced manner among students, including productive skills, tasks in which students utilize all four skills should be implemented. For that sake, the washback effect imposed by such a high-stake test as NCT is useful, if not indispensable. It is sincerely hoped that, through implementing the system to require all students to take the tests of four skills for university entrance examination, English education in Japan will cultivate all four skills among students more than ever.

Endnotes

[1] Readers may wonder why a defunct test was chosen and claim that a test still in use should be targeted instead. The abolishment of the test was announced after a large portion of data were already collected, all of a sudden by ETS. So this incident unpredictably happened.

[2] The deadline was postponed until the end of March 2017 for some exceptional cases, including the present study.

Acknowledgements
First of all, I would like to express my deepest gratitude to Atsushi Mizumoto and Yo In'nami who gave me valuable comments on statistical procedures and also to Neal Snape for proofreading the paper. Second, this study was not feasible without the technical help of Kazuma Shigeno for his technical help and the help of the five staffers at GC&T, Sota Hori, Kotaro Kurihara, Naoko Ohno, Hiromi Ando, and especially Mikihiko Takawa; so I greatly appreciate their continuous support. Lastly, I thank all of the participants and the high school principals who kindly allowed me to recruit students at their schools and, not to mention, the editor and anonymous reviewers for their constructive feedback and patience.

Funding
This study was supported by KAKENHI (JP15K02789).

Authors' contributions
NK singly designed the study, collected the data, performed the statistical analysis, drafted the manuscript, and approved the final manuscript.

Competing interests
The author declares that he has no competing interests.

Publisher's Note

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

Received: 10 July 2017 Accepted: 8 September 2017
Published online: 23 October 2017

References

Alderson, C. J. (1991). Language testing in the 1990s: how far have we come? How much further have we to go? In S. Anivan (Ed.), *Current developments in language testing* (pp. 1–26). Singapore: SEAMEO Regional Language Centre.
Bachman, L. F., & Palmer, A. S. (1981). The construct validation of the FSI oral interview. *Language Learning, 31*(1), 67–86.
Bachman, L. F., & Palmer, A. S. (1982). The construct validation of some components of communicative proficiency. *TESOL Quarterly, 16*(4), 449. https://doi.org/10.2307/3586464.
Benesse. (1997a-2017a). GTEC. Retrieved from http://www.benesse.co.jp/gtec/
Benesse. (1997b-2017b). GTEC CTE. Retrieved from http://www.benesse.co.jp/gtec/cte_company/index.html
Benesse. (2000–2017). *GETC CBT*. Retrieved from http://www.benesse-gtec.com/cbt/.
Benesse. (n.d.-a). GTEC daigakuban 4-ginouka start no oshirase *flier*. Tokyo, Japan.
Benesse. (n.d.-b). GTEC Junior. Retrieved from http://www.benesse-gtec.com/jr/
British Council, IDP: IELTS Australia, & Cambridge English Language Assessment. (2017). IELTS. Retrieved from https://www.ielts.org/
Byrne, B. M. (2016). *Structural equation modeling with AMOS: basic concepts, applications, and programming* (3rd ed.). New York: Routledge.
Cambridge English Language Assessment. (2017). Cambridge English. Retrieved from http://www.cambridgeenglish.org/
Educational Testing Service. (2017a). TOEFL iBT. Retrieved from https://www.ets.org/toefl
Educational Testing Service. (2017b). TOEFL Junior. Retrieved from https://www.ets.org/toefl_junior/
Educational Testing Service. (2017c). TOEIC. Retrieved from https://www.ets.org/toeic/
Eigo yonginou shikaku kentei kondankai. (2016). Eigo yonginou shiken jyouhou site. Retrieved from http://4skills.jp/qualification/comparison.html
Eiken Foundation of Japan. (2015). *Daigakunyushi centersiken tono soukanchousa: Jitsuyoueigo ginoukentei to TEAP de jissi*. Retrieved from Tokyo, Japan: https://www.eiken.or.jp/teap/info/2015/pdf/20151007_pressrelease_testresearch.pdf
Eiken Foundation of Japan. (n.d.-a). BULATS. Retrieved from http://www.eiken.or.jp/bulats/
Eiken Foundation of Japan. (n.d.-b). Eiken. Retrieved from http://www.eiken.or.jp/eiken/
Eiken Foundation of Japan. (n.d.-c). TEAP. Retrieved from http://www.eiken.or.jp/teap/
Eiken Foundation of Japan. (n.d.-d). TEAP CBT. Retrieved from http://www.eiken.or.jp/teap/
Fouly, K. A., Bachman, L. F., & Cziko, G. A. (1990). The divisibility of language competence: a confirmatory approach. *Language Learning, 40*(1), 1–21. dx.doi.org/10.1111/j.1467-1770.1990.tb00952.x.
GC&T. (n.d.). TOEFL Junior Comprehensive Test naiyou. Retrieved from https://gc-t.jp/toefljunior/comprehensive/test/detail/
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.
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*(1). https://doi.org/10.1186/s40468-016-0025-9.
In'nami, Y., & Koizumi, R. (2012). Factor structure of the revised TOEIC test: a multiple-sample analysis. *Language Testing, 29*(1), 131–152. https://doi.org/10.1177/0265532211413444.
Llosa, L. (2007). Validating a standards-based classroom assessment of English proficiency: a multitrait-multimethod approach. *Language Testing, 24*(4), 489–515. https://doi.org/10.1177/0265532207080770.

Ministry of Education, C., Sports, Science and Technology. (2010). Koutou gakkou gakushuu shidou yoryo eiyakuban (kariyaku). Retrieved from http://www.mext.go.jp/a_menu/shotou/new-cs/youryou/eiyaku/__icsFiles/afieldfile/2011/04/11/1298353_9.pdf

Mizumoto, A. (2014). Sokuteino datouseito shinraisei -yuidatano hissujyouken towa. In O. Takeuchi & A. Mizumoto (Eds.), *Gaikokugo kyouiku kenkyu handbook -kaiteiban* (pp. 17–31). Tokyo: Shohakusha.

National Center for University Entrance Examinations. (2015). National Center for University Entrance Examinations Annual Report. Retrieved from Tokyo, Japan: http://www.dnc.ac.jp/albums/abm.php?f=abm00006725.pdf&n=2015大学入試センター英版.pdf

National Center for University Entrance Examinations. (2016). Heisei 28 nendo daigaku nyushi center shiken jisshi kekka no gaiyou. Retrieved from http://www.dnc.ac.jp/albums/abm.php?f=abm00006898.pdf&n=別添2:[試験情報]平成28年度大学入試センター試験実施結果の概要.pdf

National Center for University Entrance Examinations. (2017). Heisei 29 nendo daigaku nyushi center shiken jisshi kekka no gaiyou. Retrieved from http://www.dnc.ac.jp/albums/abm.php?f=abm00009105.pdf&n=別添2:【修正　試験情報】平成29年度大学入試センター試験実施結果の概要%2B%2B-%2Bコピー.pdf

National Center for University Entrance Examinations. (n.d.-a). Center shiken sanka daigaku jyouhou. Retrieved from http://www.dnc.ac.jp/center/daigaku_jouhou.html

National Center for University Entrance Examinations. (n.d.-b). Shiganshasuu jyukenshasuu touno suii. Retrieved from http://www.dnc.ac.jp/data/suii/suii.html

Oller, J. W. J. (1979). *Language tests at school: A pragmatic approach*. London: Longman.

Oller, J. W. J. (1983). Evidence for a general language factor: an expectancy grammar. In J. W. J. Oller (Ed.), *Issues in language testing research* (pp. 3–10). Rowley: Newbury House.

Otsu, T. (2013). Monitor chousasha jyukensha ni okeru eigohyoujyunka test no seiseki. *Heisei 24 nendo listening test no jisshikekka ya seikatou wo kensyoushi sono kaizenwo hakarutameno chousakenkyuu ni kannsuru houkokusho*, 89–95.

Otsu, T. (2014). Hyoujyunka eigoshiken to center shiken eigokamoku tokuten tono kankei bunseki. Retrieved from Tokyo, Japan: http://antlers.rd.dnc.ac.jp/~otsu/doc/jart2014_p66-69.pdf

Papageorgiou, S., Xi, X., Morgan, R., & So, Y. (2015). Developing and validating band levels and descriptors for reporting overall examinee performance. *Language Assessment Quarterly, 12*(2), 153–177. https://doi.org/10.1080/15434303.2015.1008480.

Pearson. (2014). PTE Academic. Retrieved from http://pearsonpte.com/

Pearson. (n.d.). Progress. Retrieved from http://product.pearsonelt.com/progress/

Sasaki, M. (1996). *Second language proficiency, foreign language aptitude, and intelligence: quantitative and qualitative analyses* (Vol. 6): Peter Lang Pub Incorporated.

Sawaki, Y., Stricker, L. J., & Oranje, A. H. (2008). Factor structure of the TOEFL Internet-based test (iBT): exploration in a field trial sample. Retrieved from Princeton, NJ: http://onlinelibrary.wiley.com/doi/10.1002/j.2333-8504.2008.tb02095.x/epdf

Sawaki, Y., Stricker, L. J., & Oranje, A. H. (2009). Factor structure of the TOEFL Internet-based test. *Language Testing, 26*(1), 005–030. https://doi.org/10.1177/0265532208097335.

Shin, S.-K. (2005). Did they take the same test? Examinee language proficiency and the structure of language tests. *Language Testing, 22*(1), 31–57. https://doi.org/10.1191/0265532205lt296oa.

So, Y., Wolf, M. K., Hauck, M. C., Mollaun, P., Rybinski, P., Tumposky, D., & Wang, L. (2015). *TOEFL Junior design framework* (RR-15-13). Retrieved from Princeton, NJ: http://onlinelibrary.wiley.com/doi/10.1002/ets2.12058/epdf

Stricker, L. J., & Rock, D. A. (2008). Factor structure of the TOEFL Internet-based test across subgroups. Retrieved from Princeton, NJ: https://www.ets.org/Media/Research/pdf/RR-08-66.pdf

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Needham Heights: Allyn and Bacon.

Tannenbaum, R. J., & Baron, P. A. (2015). Mapping scores from the TOEFL Junior Comprehensive test onto the Common European Framework of Reference (CEFR). Retrieved from Princeton, NJ: https://www.ets.org/Media/Research/pdf/RM-15-13.pdf

United Nations Associations of Japan. (n.d.). Kokuren Eiken. Retrieved from http://www.kokureneiken.jp/

Yamanishi, T. (2001). Eiken syutokukyuu to daigaku nyuushi center shiken eigokamoku no tensuu tono soukankankei. *STEP Bulletin, 13*, 26–42.