Sims and Kunnan *Language Testing in Asia* (2016) 6:1
DOI 10.1186/s40468-016-0024-x

Language Testing in Asia
a SpringerOpen Journal

RESEARCH Open Access

Developing evidence for a validity argument for an English placement exam from multi-year test performance data

James M. Sims[1*] and Antony John Kunnan[2]

* Correspondence: sims@thu.edu.tw
[1]Tunghai University, Taichung, Taiwan
Full list of author information is available at the end of the article

Abstract

Background: This study investigated the factor structure and factorial invariance of an English Placement Exam (EPE) from 1998 to 2011 to provide evidence for both the appropriateness of the test scores interpretations and for a validity argument.

Methods: Test performance data collected from 38,632 freshmen non-English majors from a university in central Taiwan from 13 years (1998–2001, and 2003–2011) was examined using both exploratory factor analysis (EFA) and confirmatory factor analysis (CFA). EFA was performed on 6 years of data (2006–2011) to establish a baseline structure, which was then further tested using CFA on data from all the years.

Results: Results from EFA supported a three-factor oblique (correlated) solution and CFA determined that a three-factor model as the best fit for the data. This model reflected the test structure posited by the test designers (grammar, reading, and listening sections) and remained factorially invariant and factorially distinct across all years, with insignificant variation in factor loadings and model fit indices.

Conclusion: The study concluded that the factor structure of the EPE provided evidence for a construct validity argument for the placement exam based on multi-year performance data, thus supporting the inferences regarding test scores interpretations and the soundness of score comparisons made across years.

Keywords: Validity argument, Factor structure, Factor analysis, Multidimensionality

Background

Research studies focused on collecting evidence for the validity argument of test score interpretations and decisions have been the subject of empirical investigations for decades in the field of language assessment. Such studies have included examinations of the trait structure and the constructs of tests by examining test performance data. The dominant method for these examinations have been exploratory and confirmatory factor analysis of test performance data as seen through a number of studies (for example, Bachman and Palmer 1981; Bachman 1982; Blais and Laurier 1995; Carr 2000; Kunnan 1994; Pomplun and Omar 2003; Stricker 2004; and others). Exploratory factor analysis (EFA) is used to determine the best interpretable factor structure of a dataset and confirmatory factor analysis (CFA) is effective in testing hypothesized factor structures and the fit of tested models. In addition, CFA can be used to investigate the impact upon

the factor structure of a test by variables outside the scope of the exam such as test taker characteristics, cultural background, previous exposure to English, and language background (for example, Abedi 2002; Boldt 1998; Kunnan 1994; Oltman et al. 1988; Shin 2005).

Similar to previous studies that employed EFA and CFA to examine the factor structure of tests (for example, Carr 2006; Shin 2005), this study employed both EFA and CFA on the English Placement Exam (EPE) by using data from 13 years. This longitudinal comparison examines factorial invariance between test taker groups over the 13 administrations in terms of the factor structure of the EPE. While most studies of factorial invariance focused on test taker characteristics and proficiency, research on factorial invariance across many years is sparse. This study aimed to fill this gap in the literature.

Previous relevant studies

Sims and Liu (2013) provided support for the content validity and concurrent validity of the EPE. Following Hughes' (2003) recommendations for a content validity study, they reported that teachers who were trained in language teaching and testing, but who were not directly involved in the production of the EPE, compared test specifications and test content (see Sims 2006 for test specifications). These teachers concluded that the content of EPE was a valid measure of the desired test specifications for grammar, reading, and listening.

In addition to providing support for the content validity, Sims and Liu (2013) also reported a strong correlation between the EPE and the High-Intermediate General English Proficiency Test (GEPT). The GEPT is designed on the criteria the Common European Framework of Reference for Languages (CEFR), while EPE was not specially designed on this framework. However, both the GEPT and EPE are "general" proficiency exams. It is worthy to note that in a correlation study between the EPE and the Intermediate GEPT, 158 out of 165 students (95.8 %) who scored 60 points or higher on the EPE passed the first stage (listening and reading) of the Intermediate GEPT. According to the LTTC website, the first stage of the Intermediate GEPT corresponds to the B1 level of the CEFR. This suggests that students who score 60 and above on the EPE may be at least at the B1 level of the CEFR for listening and reading. Also, since both exams are "general" proficiency exams, the strong correlation to the GEPT suggests that the EPE may have concurrent validity similar to that of the GEPT.

Sims and Liu (2013) also pointed out that there have been numerous changes in English language learning over the last two decades in Taiwan, and as a result there have been changes in the language performances of incoming university freshmen. Their study investigated changes in the English ability of freshmen at a Taiwanese university from 1998 to 2010 by conducing year-by-year comparisons of scores for grammar, reading, and listening from the EPE. Their results indicated that students' total scores during the 12 years remained relatively consistent from year to year, while from 1998 to 2005 students' grammar and reading performances declined. These declines in students' grammar and reading scores tapered off during the last 5 years of this study. However, students' listening scores never declined, but increased significantly from year to year across the 12 years.

Sims and Liu (2013) provided support for the content validity and concurrent validity of the EPE, but the construct validity of the EPE must also be established in order for credible score comparisons across the years. It must be shown that differences in performances are not the result of changes in the construct measured. Since the EPE content remained unchanged from year to year, yet different students took the EPE each year, evidence of the lack of factorial invariance of EPE between the years would allow for meaningful comparisons of scores from the EPE from different years. Changes in these scores can be used to identify changes in the characteristics of test takers from year to year.

Also of interest is whether language ability is comprised of a unitary construct or composed of multiple components. The multidimensionality of language has been widely debated for decades, with some favoring the notion that language is a single, undifferentiated construct and others viewing it as having multiple components (Oller and Hinofotis 1980; Oller 1983; Farhady 1983; Vollmer and Sang 1983, and Kunnan 1994, 1998). This study took a novel look at this debate by using a multi-year data approach to examine the factor structure of a test. The findings could be used to support the claim of stability of the EPE over time.

Research questions

This study attempted to answer the following three research questions:

RQ1: What is the factor structure of the EPE? Does the factor structure reflect the test design?
RQ2: Does the structure of the EPE remain factorially invariant across years?
RQ3: Are the three subsections (grammar, reading, and listening) of the EPE factorially distinct?

Methods

Participants

The subjects for this study were comprised of the incoming freshman at a university in central Taiwan from 1998 to 2011 (excluding 2002 because a different procedure was used to place students that year). The freshman population at this university was extremely homogeneous: 99.9 % were Chinese (citizens of the Republic of China), the average age was 18, and there were approximately equal numbers of males and females in each year. The freshmen who entered this university each year reflected the same portion of the total freshmen population of Taiwan for each year (see Sims and Liu 2013). The number of test takers for each year is listed in Table 1.

Instruments

The EPE assess three constructs: grammar, reading, and listening. It is an entirely multiple-choice response-format test exam composed of 60 test items. The grammar section (20 %) is composed of two cloze paragraphs (G1 and G2) with ten items each for a total of 20 points. The reading section (40 %) is composed of two passages (R1 and R2) with ten items per passage for a total of 40 points. The listening section (40 %) is composed of three parts: story (LS) with seven items, dialogue (LD) with seven items,

Table 1 Test takers by year

Year	Test takers	Male	Female
2011	3223	1515	1708
2010	3002	1349	1653
2009	3083	1516	1567
2008	3108	1546	1562
2007	3009	1415	1594
2006	2946	1356	1590
2005	3053	1486	1567
2004	2828	1396	1432
2003	2966	1458	1508
2001	2835	1361	1474
2000	2796	1461	1335
1999	2906	1552	1354
1998	2877	1465	1412

and appropriate response (LA) with six items. Detailed descriptions of test specifications, test construction, item analysis, content validity studies, correlation studies, and reliability studies of the EPE can be found in Sims (2006) and Sims and Liu (2013).

Test committees composed of five to seven experienced Freshman English teachers developed each section of the exam using the desired specifications as a blueprint. Each component had been used on midterm or final exams in previous years. Cloze tests, reading passages, scripts, and questions were selected and modified based on the results of test analysis from these previous administrations. The following four guidelines were used for designing the multiple-choice items: 1) each item measured a specific objective; 2) both the question and distractors were stated simply and directly; 3) the intended answer was the only correct answer; and 4) items were accepted, discarded or revised based on item difficulty, item discrimination, and distractor analysis.

One version of the EPE was used from 1998 to 2005, while another version was used from 2006 to 2011. Both versions of the EPE were identical in terms of structure and have equal numbers of corresponding items and questions. Content validity studies, correlation studies, and item analysis comparisons verify that the two versions of the EPE used in this study were equivalent in terms of test content, item difficulty and item discrimination (sees Sims and Liu 2013).

In short, the intended design of the EPE was to assess three constructs: grammar (G), reading (R), and listening (L) with the sub-structure of the EPE being further divided into two grammar tasks (G1 and G2), two reading tasks (R1 and R2) and three listening tasks (LS, LD, LA).

Data collection

During freshman orientation all incoming freshmen at the university took the EPE. The main purpose of the placement exam was to divide students into three levels of Freshman English for Non-Majors (FENM). Students were placed into FENM from a continuum of scores with the top third placed into high-level FENM (EPE scores were usually 75 or higher), the middle third into mid-level FENM (usually ranging between

74 and 55), and the bottom third into low-level FENM (usually 54 or below). Students scoring below 30 points on the listening section were required to take an additional listening lab.

The students in this study were informed of the purpose of the exam and were administered the placement exam approximately 1 year apart from each other at the same time of day (1:30 pm - 3:30 pm) and in the same locations. For all groups, the results were obtained in-house, using the same equipment and software for each of the administrations.

Data analysis

Data analysis consisted of a three-step process. First, descriptive statistics were examined for normality and linearity for all data. EFA was then performed on data from 2006 to 2011 for the purpose of determining the best factor structure for the EPE based on both interpretability and parsimony and orthogonal (uncorrelated latent variables or constructs) and oblique (correlated latent variables or constructs) solutions. Next, the best solution from EFA was tested with CFA on data from all years. Finally, factorial invariance was assessed by using Comparative Fit Index (CFI), Goodness of Fit Index (GFI), and the root square error approximation (RMSEA) to ascertain any deviations from the derived model or significant changes in indicators of model fit.

The reasoning behind not using all datasets for EFA was to cross-validate the factor structure with additional data from the remaining years. Using identical data for both EFA and CFA would have been confirming a structure with the same data that had been used to derive it. Using additional datasets made the results from CFA more trustworthy.

Results

Descriptives

Descriptive statistics and measures of central tendency and distribution showed no signs of skewness or kurtosis for any of the different variables across the 13 years, meaning that all distributions were considered to be normal for the sake of analysis. Furthermore, the datasets were considered appropriate for EFA based on Bartlett tests for homogeneity of variances of data across the years.

Exploratory Factor Analysis (EFA)

Test performance data from 6 years (2006–2011) were used to conduct EFA to establish a baseline structure. Factors were identified from EFA by first eliminating factors with eigenvalues of less than one, and then under-factoring and over-factoring to determine the best factor structure.

The initial extraction from EFA supported a two-factor solution because only two factors had eigenvalues greater than one and because the slope on the curve of the scree plot decreases significantly after the second factor. The factor loadings for a two-factor solution are given in Tables 2 and 3. The primary loadings are shown in bold. In the two-factor solution, the grammar tasks loaded into a single factor, with the second factor being comprised of the reading and listening tasks. The only anomalies were for R1 in 2007 and 2010 which loaded more closely with the grammar tasks. The solutions

Table 2 EFA two-factor loadings 2006–2008

	2006		2007		2008	
	Factor		Factor		Factor	
	1	2	1	2	1	2
G1	**.649**	*.323*	**.630**	*.354*	**.556**	*.455*
G2	**.851**	.017	**.837**	-.022	**.902**	.034
R1	*.416*	**.523**	**.511**	*.458*	*.406*	**.564**
R2	*.389*	**.592**	*.487*	**.570**	*.375*	**.651**
LD	.164	**.837**	.213	**.820**	.154	**.833**
LS	.152	**.832**	.201	**.828**	.166	**.828**
LA	.097	**.780**	.094	**.789**	.072	**.770**

*Bolded numbers represent high loadings

were not simple as there were many secondary loadings above 0.30 (in italics). The two-factor models were able to explain approximately 61, 62, 63, 63, 60, and 62 % of the variance in each of the variables from 2006 to 2011 respectively.

As recommended by Stevens (1992), since less than 70 % of the variance was accounted for by the two-factor solution, under-factoring and over-factoring were attempted. The results of the one-factor solution were not interpretable in this context. Tables 4 and 5 give the results of the three-factor solution. In the three-factor solution, loadings were unambiguous and clearly interpretable for all 6 years with each type of task loading into separate factors. There were fewer secondary loadings above 0.30 (in italics). The three factors were able to explain approximately 72, 73, 73, 72, 70, and 72 % of the variance from 2006 to 2011 respectively. Further, orthogonal and oblique solutions were examined and this indicated that there were substantial correlations in the solutions among the latent variables or constructs to prefer oblique solutions.

Confirmatory Factor Analysis (CFA)

In comparison to the two-factor solution, the three-factor solution was selected as the baseline model for CFA because it accounted for a higher level of variance, lacked the anomalies found in 2007 and 2010, and comprised higher loadings. The model tested

Table 3 EFA two-factor loadings 2009–2011

	2009		2010		2011	
	Factor		Factor		Factor	
	1	2	1	2	1	2
G1	**.583**	*.413*	**.604**	*.363*	**.610**	*.383*
G2	**.888**	.014	**.841**	-.039	**.862**	-.011
R1	*.462*	**.530**	**.562**	*.403*	**.513**	*.564*
R2	*.383*	**.644**	*.451*	**.531**	*.455*	**.609**
LD	.178	**.825**	.185	**.820**	.200	**.819**
LS	.174	**.809**	.191	**.812**	.200	**.810**
LA	.078	**.793**	.127	**.768**	.096	**.785**

*Bolded numbers represent high loadings

Table 4 EFA three-factor loadings 2006–2008

	2006			2007			2008		
	Factor			Factor			Factor		
	1	2	3	1	2	3	1	2	3
G1	**.593**	*.424*	.152	**.673**	.284	.196	**.666**	.219	.293
G2	**.945**	.105	.128	**.963**	.160	.120	**.963**	.166	.122
R1	.049	**.783**	.223	-.022	**.833**	.190	.212	**.826**	.028
R2	.053	**.741**	*.320*	*.392*	**.677**	.113	*.404*	**.678**	.100
LD	.094	*.314*	**.804**	.112	.290	**.808**	.095	*.320*	**.802**
LS	.085	*.307*	**.800**	.084	*.306*	**.805**	.098	*.338*	**.788**
LA	.126	.133	**.830**	.075	.149	**.818**	.094	.159	**.815**

*Bolded numbers represent high loadings

for the CFA is displayed in Fig. 1 (the error variances associated with each variable are not shown).

Each factor (latent variable or construct) was represented as an oval, which in turn correlates with scores for each of the tasks. Due to the high correlations between factors found in the EFA from the oblique solutions, each latent variable was correlated with the others, as represented by the double-headed arrows. This model was tested with data from all years to determine the overall fit of the model to the data as well as any variance in that fit. Three types of results were analyzed to determine this: the loadings on each path leading from a latent variable to an observed variable, the total variance of each observed variable explained by the corresponding latent variable (the r^2), and the model fit indicators, specifically the goodness-of-fit index (GFI), comparative-fit index (CFI), and the root mean square error of approximation (RMSEA) (Byrne 1994).

The three-factor model path loadings and r^2 are reported in Tables 6, 7, 8, and 9. Figure 2 shows the model applied to 1998 (dataset not part of EFA) and Fig. 3 shows the model applied to 2011 (dataset part of EFA) with standardized estimates.

The correlations between the factors and the observed variables (path loadings) range between .53 and .83 and are consistent across the years for all datasets. The path loadings for G1 (.67 - .73) are higher than those for G2 (.53 -.69). The reading variables (R1: .62 - .66; R2: .68 - .76) tend to have higher path loadings than those for the

Table 5 EFA three-factor loadings 2009–2011

	2009			2010			2011		
	Factor			Factor			Factor		
	1	2	3	1	2	3	1	2	3
G1	**.734**	.122	.180	**.570**	*.344*	.256	**.673**	.273	.219
G2	**.969**	.183	.122	**.960**	.143	.105	**.964**	.177	.118
R1	.045	**.795**	.247	.098	**.813**	.159	.024	**.833**	.184
R2	.099	**.610**	.466	.024	**.740**	*.310*	.126	**.635**	.449
LD	.088	.287	**.810**	.096	.258	**.815**	.089	*.302*	**.796**
LS	.108	*.304*	**.780**	.084	.281	**.796**	.100	.286	**.794**
LA	.069	.166	**.819**	.092	.165	**.792**	.079	.145	**.816**

*Bolded numbers represent high loadings

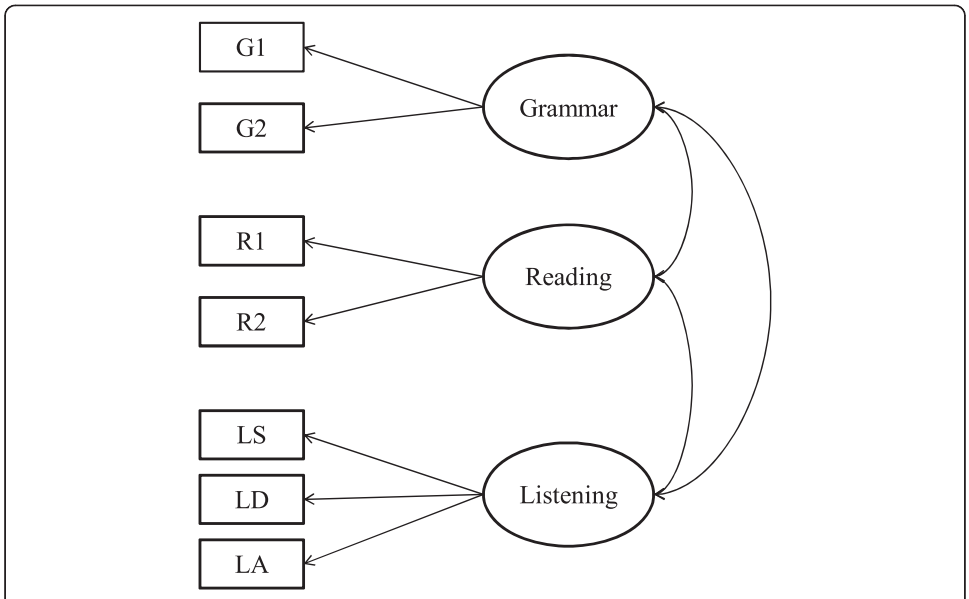

Fig. 1 Structural model used for confirmatory factor analysis

grammar variables. The listening tasks generally have the highest path loadings (LS: .74 - .83; LD: .71 - .83; LA: .66 - .76). The consistencies between the results for the datasets not used in EFA and datasets used in EFA cross-validate reconfirm the derived model. Between 19 and 78 % of the variance within the observed variables are explained by the factors.

Indices of model fit are given in Table 10. The GFI and CFI indicators are all above .992 (.90 being the threshold) while the RMSEA is below .03 (.05 being the threshold) for all years (Byrne 1994). The results of these three fit indices indicate that the three-factor structure model is a good fit for all datasets.

The correlations between grammar (G), reading (R), and listening (L) are presented in Table 11. The correlations between grammar and reading range from .77 to .87 and tended to have the highest correlations of the three latent variables each year. The correlations between reading and listening range from .73 to .82 are stronger each year than those for grammar and listening which range from .68 to .81.

Table 6 CFA three-factor model path loadings and r^2, 1998–2001

	1998		1999		2000		2001	
	Loading	r^2	Loading	r^2	Loading	r^2	Loading	r^2
G1	.68	.47	.68	.47	.72	.52	.69	.51
G2	.67	.49	.65	.28	.62	.33	.65	.39
R1	.65	.42	.65	.43	.66	.44	.65	.46
R2	.72	.52	.74	.55	.72	.51	.73	.57
LS	.74	.63	.79	.63	.76	.66	.78	.66
LD	.77	.60	.77	.59	.80	.64	.78	.63
LA	.71	.51	.72	.52	.73	.52	.72	.54

Table 7 CFA three-factor model path loadings and r^2, 2003–2005

	2003		2004		2005	
	Loading	r^2	Loading	r^2	Loading	r^2
G1	.69	.48	.68	.47	.73	.54
G2	.62	.38	.61	.37	.69	.47
R1	.65	.42	.66	.44	.66	.44
R2	.73	.53	.72	.52	.68	.46
LS	.78	.60	.76	.58	.80	.64
LD	.71	.50	.71	.51	.76	.58
LA	.73	.53	.76	.58	.78	.60

Discussion

This study investigated the factor structure and factorial invariance of the EPE from 1998 to 2011 to provide evidence for both the appropriateness of the test scores interpretations and for a validity argument. This section discusses the results in relation to the three research questions.

RQ1: What is the factor structure of the EPE? Does the factor structure reflect the test design?

Based on the eigenvalues and the scree plot, the initial EFA derived a two-factor solution. Over-factoring revealed an interpretable three-factor solution that explained more of variance in the total task scores than the two-factor solution. The type of two-factor solution grouped the grammar tasks as a single factor and grouped the reading and listening tasks as another factor, while the three-factor solution loaded each task into a distinct factor. The two-factor solution is supported by previous research (Romhild 2008; Tomblin and Zhang 2007) with language performance test data that found the separability of grammar from both reading and listening. However, the grouping of reading and listening together conflicts with the widely accepted notion of the distinctiveness of reading from listening (Bae and Bachman 1998; Hale et al. 1989; Shin 2005; Song 2008). In addition, a panel of teachers who conducted a content validity study of the EPE determined that each section of test were individually distinct (Sims 2006). In short, the higher explained variance, previous research on the separability of language tasks, and the content validity study of the EPE, all make the three-factor solution more interpretable than the two-factor solution.

Table 8 Three-factor model path loadings and r^2, 2006–2008

	2006		2007		2008	
	Loading	r^2	Loading	r^2	Loading	r^2
G1	.67	.45	.68	.47	.70	.49
G2	.54	.21	.53	.22	.54	.19
R1	.64	.41	.64	.42	.66	.44
R2	.71	.50	.76	.58	.75	.56
LS	.81	.65	.82	.68	.83	.68
LD	.83	.69	.82	.67	.83	.69
LA	.69	.48	.67	.45	.67	.44

Table 9 CFA three-factor model path loadings and r^2, 2009–2011

	2009		2010		2011	
	Loading	r^2	Loading	r^2	Loading	r^2
G1	.69	.47	.68	.46	.69	.48
G2	.55	.20	.53	.22	.55	.20
R1	.65	.43	.63	.39	.62	.38
R2	.74	.54	.69	.47	.75	.57
LS	.79	.62	.79	.62	.80	.63
LD	.82	.68	.80	.65	.81	.66
LA	.69	.48	.66	.44	.67	.45

CFA confirmed that the EPE measures three constructs: grammar proficiency, reading proficiency, and listening proficiency. Given the high inter-factor correlations found in EFA and correlations between latent variables from CFA, these three constructs may be in turn governed by a higher order factor. This higher order model would posit that the three factors of grammar proficiency, reading proficiency, and listening proficiency are components of a fourth factor (which could be labeled as a general language ability factor). Other studies (Bachman, Davidson, Ryan, and Choi 1995; Sawaki et al. 2008; Shin 2005) label this higher order factor as general English proficiency. Since the EPE does not have writing or speaking components and students are not required to produce extended discourse, a better term for the EPE context might be *English language comprehension*. The term "English proficiency" may be too encompassing for a test that only assesses students' knowledge of grammatical rules, ability to read, and understanding spoken English.

The factor structure also indicated that constructs are measured by the corresponding tasks within the EPE: the two grammar passages measure grammar proficiency; the reading passages measure reading proficiency, and so on. This solution supported the

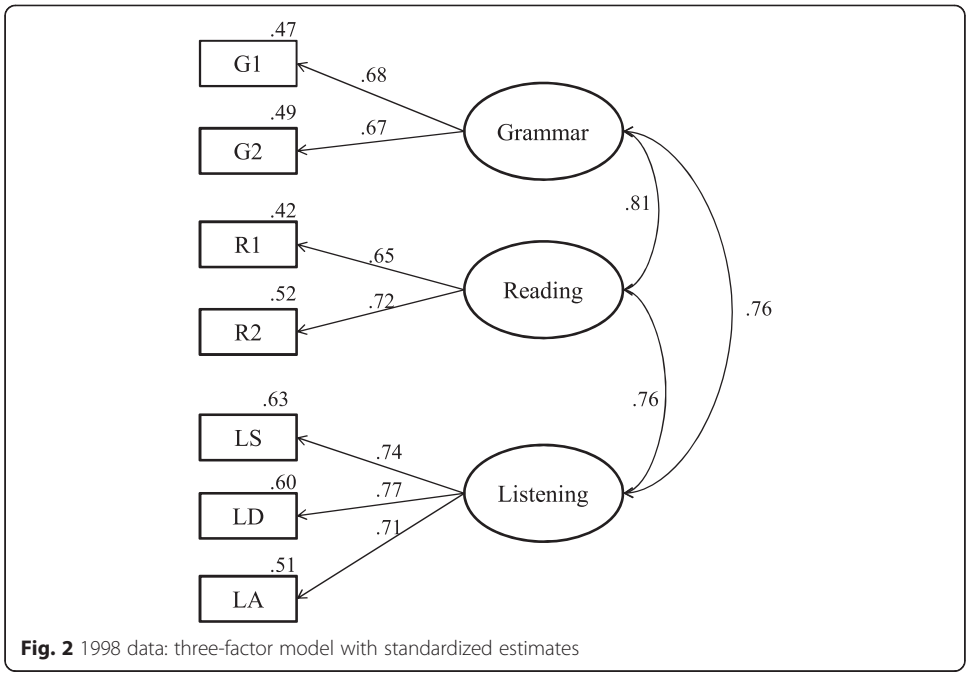

Fig. 2 1998 data: three-factor model with standardized estimates

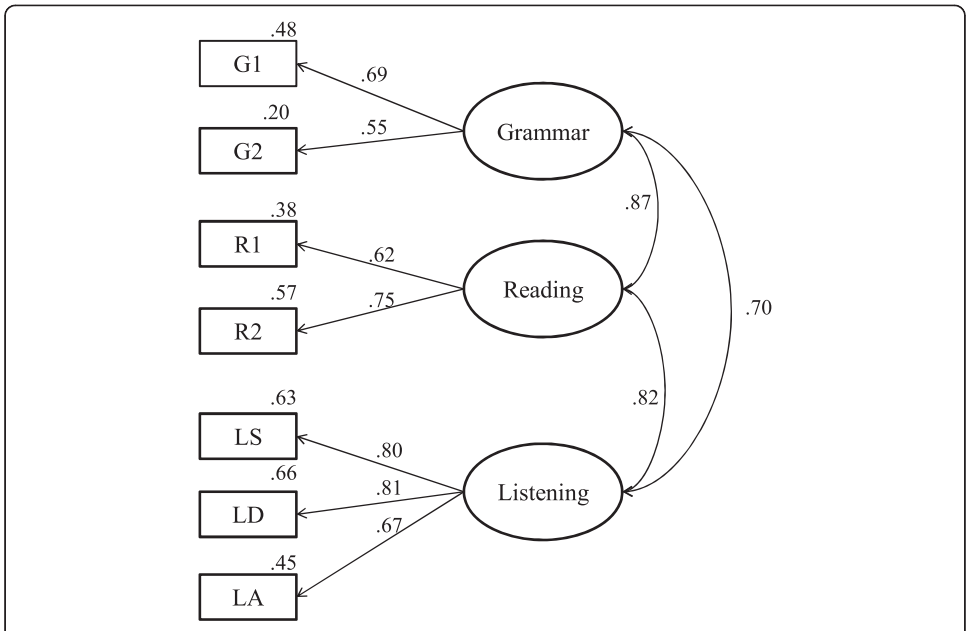

Fig. 3 2011data: three-factor model with standardized estimates

test structure proposed by the test designers, and served to validate the inferences made based on scores on this exam.

RQ2: Does the structure of the EPE remain factorially invariant across years?

The results not only found that the test structure derived via factor analysis reflected that intended by the designers, but also that this structure remained constant across 13 distinct groups of test takers. The stability, or lack of variation, of the factor structure over time indicated that students from subsequent years continue to perceive the test content in a consistent manner. As a result, meaningful comparisons of test scores can be made across years because the EPE measured the same constructs from year to year.

Table 10 CFA three-factor model fit indices, 1998–2011

	GFI	CFI	RMSEA
1998	.997	.996	.023
1999	.996	.996	.028
2000	.996	.995	.031
2001	.999	.998	.005
2003	.994	.992	.038
2004	.995	.994	.034
2005	.993	.991	.045
2006	.999	1.000	.002
2007	.998	.999	.015
2008	.999	1.000	.008
2009	.997	.998	.022
2010	.998	.998	.017
2011	.999	1.000	.007

Table 11 CFA three-factor model correlations, 1998–2011

	1998	1999	2000	2001	2003	2004	2005
G - R	.81	.81	.81	.82	.84	.78	.81
R - L	.76	.77	.75	.81	.75	.73	.80
G - L	.76	.69	.71	.76	.73	.73	.81
	2006	2007	2008	2009	2010	2011	
G - R	.77	.83	.83	.84	.83	.87	
R - L	.78	.77	.82	.84	.77	.82	
G - L	.64	.69	.71	.71	.68	.70	

The factorial invariance of the EPE over time means that changes in the performance on each of the three sections of the EPE found by Sims and Liu (2013) were not due to shifts in the interpretation of those tasks by successive groups of students. This adds credence to Sims and Liu (2013) assertion of changes in the language performance of incoming freshmen in Taiwan.

RQ3: Are the three subsections (grammar, reading, and listening) of the EPE factorially distinct?

The results of this study found that the three sections of the EPE reflected constructs that were factorially distinct. While the separablity of listening from reading and grammar is widely accepted (Bae and Bachman 1998; Hale et al. 1989; Shin 2005; Song 2008), the distinctness of grammar from reading is more controversial. For example, Tomblin and Zhang (2007)) found the two construct were distinct, whereas Romhild (2008) found grammar and reading grouping together. An explanation for the separation of reading from grammar found in this study may be found in an examination of the content of EPE. The reading items focused on main ideas, specific details, and vocabulary in context, and did not require a deep understanding of the syntax or grammar. On the other hand, the grammar items mostly dealt with appropriate verb tense, subject-verb agreement, count versus non-count nouns, possessive pronouns, conjunctions, and passive voice, and did not require a deep understanding of the content of the passage.

To summarize, the first research question addresses the factor structure of the EPE. Both EFA and CFA confirm that the EPE measured three constructs: grammar proficiency, reading proficiency, and listening proficiency. The second question regards whether this structure is invariant across all years. The lack of variance of the factor loadings and strong indicators of model fit supported the notion that factor structure of the EPE remained invariant from 1998 to 2011. The final research question concerns whether the three section of the EPE (grammar, reading, and listening) are factorially distinct. The three-factor model confirmed by CFA found these three tasks to be distinct factors for all years.

Conclusion

Firstly, from a test development perspective, this study confirmed that the structure of the EPE conforms to the structure posited by the test designers and that the constructs being measured are those that the designers intended to have measured. There was

overall agreement between the intended test structure and that derived by this study. Secondly, the study established that the factor structure remained constant across all years and sound comparisons of performances can be made across years. In other words, differences in performance cannot be due to different constructs being measured from year to year. Finally, the distinctiveness of the three constructs and the constancy of that distinctness over time, provided evidence for the concept of language as a multidimensional construct.

This study was limited as it investigated test takers as a homogenous group, who differed only across years, and did not take in to account ability level. It was possible that the factor structure may not remain constant over time for all levels. Future studies could investigate factorial invariance with respect to student performance and proficiency level, similar to that of Romhild (2008). Another limitation of this study is that it only analyzed section score totals. Future research could conduct factor analysis on item-level data rather than just section score totals. Finally, a detailed analysis of the EPE's content at the item-level with detailed specifications could shed light on how the structure of the test should be interpreted and whether or not the test items adequately capture the TLU (Target Language Use) domain (Bachman and Palmer 1996).

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

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

Author details
[1]Tunghai University, Taichung, Taiwan. [2]Guangdong University of Foreign Studies, Guangzhou, China.

Received: 11 October 2015 Accepted: 12 January 2016
Published online: 21 January 2016

References

Abedi, J. (2002). Standardized achievement tests and English language learners: Psychometrics issues. *Educational Assessment, 8*, 231–257.
Bachman, L. F. (1982). The trait structure of cloze test scores. *TESOL Quarterly, 16*, 61–70.
Bachman, L. F., Davidson, F. G., Ryan, K., & Choi, I.-C. (1995). *An investigation into the comparability of two tests of English as a foreign language: The Cambridge-TOEFL comparability study*. Cambridge, England: UCLES.
Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford, England: Oxford University Press.
Bachman, L. F., & Palmer, A. S. (1981). The construct validation of the FSI oral interview. *Language Learning, 31*, 67–86.
Bae, J., & Bachman, L. F. (1998). A latent variable approach to listening and reading: testing factorial invariance across two groups of children in the Korean/English Two-Way Immersion Program. *Language Testing, 15*, 380–414.
Blais, J. G., & Laurier, M. D. (1995). The dimensionality of a placement test from several analytical perspectives. *Language Testing, 12*, 72–98.
Boldt, R. F. (1998). *Latent structure analysis of the Test of English as a Foreign Language. TOEFL Research Report 28*. Princeton, NJ: Educational Testing Service.
Byrne, B. M. (1994). *Structural equation modeling with EQS and EQS/Windows*. Thousand Oaks, CA: Sage Publications.
Carr, N. (2000). A comparison of analytic and holistic rating scale types in the context of composition tests. *Issues in Applied Linguistics, 11*, 207–241.
Carr, N. (2006). The factor structure of test task characteristics and examinee performance. *Language Testing, 2006*(23), 269–289.
Farhady, H. (1983). On the plausibility of the unitary language proficiency factor. In J. W. Oller (Ed.), *Issues in language testing research* (pp. 11–28). Rowley, MA: Newbury House.
Hale, G. A., Rock, D. A., & Jirele, T. (1989). *Confirmatory factor analysis of the TOEFL.TOEFL Research Report 32*. Princeton, NJ: Educational Testing Service.
Kunnan, A. J. (1994). Modeling relationships among some test-taker characteristics and performance on EFL tests: an approach to construct validity. *Language Testing, 11*, 225–252.
Kunnan, A. J. (1998). Approaches to validation in language assessment. In A. Kunnan (Ed.), *Validation in language assessment* (pp. 1–8). Mahwah, NJ: Lawrence Erlbaum Associates.
Oller, J. W., & Hinofotis, F. (1980). Two mutually exclusive hypotheses about second language ability: Indivisible or partially divisible competence. In J. W. Oller & K. Perkins (Eds.), *Research in language testing* (pp. 13–23). Rowley, MA: Newbury House.
Oller, J. W. (1983). Evidence for a general language proficiency factor: An expectancy grammar. In J. W. Oller (Ed.), *Issues in language testing research* (pp. 3–10). Rowley, MA: Newbury House.

Oltman, P. K., Stricker, L. J., & Barrows, T. (1988). *Native language, English proficiency and the structure of the TOEFL (TOEFL Research Report 27)*. Princeton, NJ: Educational Testing Service.
Pomplun, M., & Omar, M. (2003). Do minority representative reading passages provide factorially invariant scores for all students? *Structural Equation Modeling, 10*, 276–288.
Romhild, A. (2008). Investigating the factor structure of the ECPE across different proficiency levels. *Spaan Fellow Working Papers in Second of Foreign Language Assessment, 6*, 29–55.
Sawaki, Y., Stricker, S., & Oranje, A. (2008). *Factor structure of the TOEFL Internet-Based Test (iBT): Exploration in a field trial sample. TOEFL iBT Research Report 4*. Princeton, NJ: Educational Testing Service.
Shin, S. (2005). Did they take the same test? Examinee language proficiency and the structure of language tests. *Language Testing, 22*, 31–57.
Sims, J. (2006). The creation of a valid and reliable university proficiency exam. *Tunghai Journal of Humanities, 47*, 325–344.
Sims, J., & Liu, J. (2013). Two decades of changes in the English ability of freshmen at a university in Taiwan. *Hwa Kang English Journal, 19*, 23–51.
Song, M. Y. (2008). Do divisible subskills exist in second language (L2) comprehension? A structural equation modeling approach. *Language Testing, 25*, 435–464.
Stevens, J. (1992). *Applied multivariate statistics for social science* (2nd ed.). Hillsdale, NJ: Erlbaum.
Stricker, L. J. (2004). The performance of native speakers of English and ESL speakers on the computer-based TOEFL and GRE General Test. *Language Testing, 21*, 146–173.
Tomblin, J. B., & Zhang, X. (2007). The dimensionality of language ability in school-age children. *Journal of Speech, Language, and Hearing Research, 49*, 1193–1208.
Vollmer, H., & Sang, F. (1983). Competing hypotheses about second language ability. A plea for caution. In J. W. Oller (Ed.), *Issues in language testing research* (pp. 29–79). Rowley, MA: Newbury House.