http://e-flt.nus.edu.sg/

Electronic Journal of Foreign Language Teaching
2019, Vol. 16, No. 1, pp. 80–94
© Centre for Language Studies
National University of Singapore

Construct Validation of the Multiple-Choice Items of the English Subtest of the Advanced Subjects Test in Taiwan

Wen-Ying Lin
(wylin@utaipei.edu.tw)
University of Taipei, Taiwan

Yi-Ju Chang
(g10209011@go.utaipei.edu.tw)
University of Taipei, Taiwan

Abstract

Administered in each July in Taiwan, the Advanced Subjects Test (AST) is a high-stakes college entrance test, of which an English subtest (AST-E) is an integral part. The AST-E score affects considerably which universities test takers are qualified to be admitted to. To them, the major concern is: what English ability does the AST-E really measure? This study was intended to investigate the construct validity of the multiple-choice (MC) items of the AST-E through confirmatory factor analysis (CFA) using data on test takers' responses to the MC items from 2015 to 2016. In the first part of this study, the MC items were classified into several language components by five experienced raters based on the two classification frameworks of Purpura (1999, 2004). The chosen components represented an incipient model for describing the relationship between the unobserved components and the data. In the second part, a series of CFAs were performed, first on the incipient model and then on some alternative models, to look for a model that offers a relatively good fit to the data. The CFA results showed that, although the incipient model involved multiple components/factors, a one-factor model was found to best portray the characteristics of test takers' responses to the MC items for 2015 and 2016, suggesting that the MC items appeared to tap simply their general English reading ability rather than a set of their divisible English reading skills. Finally, this study concluded with some pedagogical and practical implications for Taiwan high school English teachers and AST-E test constructors.

1 Introduction

What does a language test measure? Or more specifically, what language ability does a language test measure? In the 1960s and 1970s, two major propositions, among some others, were proposed one after the other to explain and measure language ability. Both coined by Oller (1979), one is referred to as the divisible competence hypothesis (DCH) and the other the unitary competence hypothesis (UCH). The DCH says that language ability can be broken up into different components and skills. The components are lexicon, morphology, phonology, and syntax; and the skills are listening, reading, speaking, and writing. According to the DCH, language ability can be measured through assessing these components and skills separately. The UCH, proposed by Oller (1976), says that language ability is unitary in that it is an inseparable set of interacting abilities which cannot be split up into isolated components. According to the UCH, language ability is a combination of many language-related skills and thus should be evaluated and measured in its entirety. These two conflicting propositions on language ability have crucial implication for language testing. Roughly speaking, if language ability is divisible, several tests should be administered with each test evaluating a specific part of an individual's proficiency in a language; in contrast, if language ability is unitary, a comprehensive test should be administered to test an individual's overall proficiency in a language.

In Taiwan, English is regarded as a major foreign language in the high school curriculum. In each July, tens of thousands of high school seniors take the Advanced Subjects Test (AST), of which an English subtest (AST-E) is an integral part. Their AST-E score affects considerably which universities they are qualified to be admitted to and, therefore, has a huge impact on their future career trajectory. As always, preparation is the best way to achieve good results in a test. But how high school seniors prepare for the AST-E depends significantly on what English ability it is claimed to measure. In the present context, does the AST-E measure several aspects/components of their English ability or simply their overall English ability? This study was intended to address such a question and, in particular, focus on the multiple-choice (MC) items of the AST-E. At this juncture, a brief description of the MC items is fitting.

The AST-E consists of 51 MC items, two translation items, and one guided composition. The 51 MC items are grouped into three sections: vocabulary (10 items), cloze (10 rational, 10 banked, and 5 gap-filling cloze items), and reading comprehension (16 items), all of which are claimed to evaluate test takers' general English reading ability by the College Entrance Examination Center (CEEC). For the vocabulary items, each question contains one or two sentences, where one word is deleted and replaced with a blank. Each question is followed by four options, from which test takers have to pick one for the correct answer. For the rational cloze items, two short passages are given, where some words or phrases are removed and replaced with blanks. For each blank, test takers have to pick one of the four given options to restore the deleted word or phrase. For the banked cloze items, one passage is given with ten blanks. The deleted words or phrases for the ten blanks normally involve cohesive devices, lexical cohesion, collocation, repetition, synonymy, or hyponym. For each blank, test takers have to choose one from a pool of 12 options to restore the deleted word or phrase. For the gap-filling cloze items, one passage is given with five blanks. For each blank, test takers have to choose one from a pool of six options to restore the deleted word or phrase. For the reading comprehension items, four reading passages are usually given, with each followed by three to five MC questions.

As mentioned above, all three MC sections are claimed by CEEC to evaluate test takers' general English reading ability. In actuality, do the three sections of the AST-E, administered over the years, live up to the claim of CEEC? That is, do these MC items have construct validity[1]? There have been a couple of related studies (e.g. Chen, 2009; Lan & Chern, 2010) examining the MC items of the AST-E through qualitative analysis, by which the items were classified subjectively by a few raters into different categories of English reading skills. Based on the classifications[2], these studies made recommendations for test preparation of the MC items. However, given the subjective way the items were classified in these studies, it is likely that some categories chosen were highly related or correlated. In other words, two or more categories may simply measure the same construct or factor associated with test takers' English ability. In this situation, factor analysis is a proper statistical tool to reduce the number of categories to a few constructs/factors which can equally well explain the factor structure of the MC items. That said, this study was intended to investigate the construct validity of the three MC sections through confirmatory factor analysis (CFA), a special form of factor analysis – with two objectives in mind. The first was to determine the underlying factor structure of the three sections of the AST-E. The second was to offer some recommendations, based on the results of this study, to Taiwan high school English teachers and AST-E test constructors.

Accordingly, the research hypothesis, phrased in terms of a question, is: What do the MC items of the AST-E measure – a set of their divisible English reading skills or simply their general English reading ability? As a pioneering CFA research in Taiwan probing into the construct validity of the MC items of the AST-E, this study was carried out in two parts using data on test takers' responses to the MC items administered from 2015 to 2016. In the first part, the MC items were classified into several language components/categories by five experienced raters based on the two classification frameworks of Purpura (1999, 2004). The chosen components represented an incipient model for describing the relationship between the unobserved components and the data. In the second part, a series of CFAs were performed, first on the incipient model and then on some alternative models, to determine what English ability the MC items of the AST-E measure.

2 Literature review

2.1 Studies related to the DCH and the UCH

Over the past half a century, there has been no unanimous agreement among researchers on the nature of language ability. In the 1960s and 1970s, a large body of theoretical discussions (Davies, 2007; Fries, 1945; Fulcher, 2015; Lado, 1961, 1964; Oller, 1976, 1979; Spolsky, 1973; Weir, Vidakovic, & Galaczi, 2013) and empirical investigations (Ajideh & Esfandiari, 2009; Carroll, 1975; Gardner & Lambert, 1965; Hosley & Meredity, 1979; Lofgren, 1969; Oller & Hinofotis, 1980; Pimsleur, Stockwell, & Comrey, 1962) revolved around the dimensionality of language ability. As mentioned at the outset, one is the DCH and the other the UCH. The DCH, also known as the skills-and-elements approach, holds that language ability can be broken up into distinct skills and distinct elements. Influenced by the structuralist school of linguistics, Lado (1961) suggested that language ability can be represented by a four-by-three matrix, where the four rows of the matrix are the four language skills (listening, reading, speaking, and writing) and the three columns are the three language elements (lexicon, phonology, and structure). According to the DCH, these 12 combinations of skill and element can be taught and measured separately.

On the other hand, inspired by the pioneering work of Charles Spearman (1904), who identified a single general factor to explain human intelligence, Oller (1976) proposed the UCH in which an individual's language ability can be explained in terms of his/her "pragmatic expectancy grammar." According to Oller (1979), pragmatic expectancy grammar (see also Beresova, 2014; Purpura, 2004) is a psychologically real system that "causes the learner to process sequences of elements in a language that conform to the normal contextual constraints of that language, and … requires the learner to relate sequences of linguistic elements via pragmatic mappings to the extralinguistic context" (p. 38). Oller claimed that pragmatic expectancy grammar constitutes a single unitary language ability, which can be measured in its entirety by integrative and pragmatic tests, such as cloze tests or dictation tests. For example, an important study by Oller & Hinofotis (1980), in which three different English tests[3] were administered to 159 Iranian students, provided strong support for the UCH.

In the 1980s, some other researchers took the position that language ability can best be represented by a general factor and several other specific factors (e.g. Bachman, 1982; Bachman & Palmer, 1981, 1982). In particular, this view, based mainly on empirical findings of studies using confirmatory factor analysis (CFA), claims that an individual's performance on language test is governed by separate but correlated traits/factors, which are in turn influenced by a single general factor. For example, using a series of CFAs, Bachman (1982) found that the model with a general factor and three specific factors best explained his data.

2.2 Studies related to the AST-E

There have been a few related studies (e.g. Chen, 2009; Lan & Chern, 2010) examining the MC items of the AST-E through qualitative analysis. Simply put, these studies involve roughly two steps: first, a few raters assembled examine the MC items and classify each of them into one of several categories of English reading skills based on a classification system or taxonomy; second, some relevant descriptive statistics are computed to describe quantitatively these raters' classifications.

Chen (2009) investigated the reading comprehension (R-C) items of the AST-E from 2002 to 2007 based on a modified version of Nuttall's taxonomy (2005), which classifies reading skills into eight categories: (1) understanding syntax; (2) recognizing and interpreting cohesive devices; (3) identifying and interpreting discourse markers; (4) recognizing functional value; (5) recognizing text organization; (6) recognizing the presuppositions underlying the text; (7) recognizing implications and making inferences; and (8) recognizing theme of the text. Her study found that the R-C items exhibited, to different degrees, nearly all the eight categories of reading skills. Category (3) was the most measured skill (accounting for 55.4% of the R-C items) and category (5) was the least measured skill (accounting for almost 0% of the R-C items).

Lan & Chern (2010) examined the R-C items of the AST-E from 2002 to 2006 based on the revised Bloom taxonomy (see Anderson & Krathwohl, 2001; Bloom, Engelhart, Furst, Hill, & Krathwohl, 1956; Krathwohl, 2002). Their study found that the R-C items were classified into "recognizing" under the "remember" category, into "interpreting", "classifying", "summarizing", "inferring", and "explaining" under the "understand" category, into "executing" under the "apply" category, and into "attributing" under the "analyze" category. In terms of category, the "understand" category contained the most R-C items (52.4%). In terms of subcategory, the "recognizing" subcategory contained the most R-C items (36.5%).

2.3 CFA for construct validation

The construct validity of the three MC sections can be evaluated through different forms of factor analysis. Of various factor analysis procedures, confirmatory factor analysis (CFA) – an approach that sets up specific hypothesis of what is measured by a test and then examines whether or not test takers' item-by-item responses agree with the a priori hypothesis – has been widely regarded as a powerful tool for extracting empirical evidence supporting the construct validity of a test (e.g. Dimitrov, 2010; DiStefano & Hess, 2005; Strauss & Smith, 2009). As pointed out by Strauss & Smith (2009), a major strength of CFA in construct validation is its theory-testing availability of direct comparison between various alternative models of construct relationship. Despite its popularity in construct validation research in recent decades, no study in Taiwan has employed CFA to investigate the construct validity of the MC items of the AST-E. That said, this study was intended to probe into the factor structure of the vocabulary, cloze, and reading comprehension sections of the AST-E using CFA as the research tool.

3 Methodology

3.1 Test items

Two datasets – one for 2015 and one for 2016 – were obtained from CEEC. Each dataset[4] contained 5,000 randomly selected test takers' item-by-item dichotomously scored responses to the MC items of the AST-E[5]. For each year, there are 51 MC items in all, including ten vocabulary items, ten rational cloze items, ten banked cloze items, five sentence gap-filling cloze items, and 16 reading comprehension items.

3.2 The instrument

The instrument used was the coding sheet for the raters to classify the MC items of the AST-E. Following Saito's study (2003), this study adopted a simplified version of Purpura's (2004) model[6] of grammatical ability for classification of the vocabulary and cloze MC items. Grammatical ability described in Purpura's full model consists of 12 components. However, a preliminary examination of the 70 (35 for 2015 and 35 for 2016) vocabulary and cloze items indicated that only five of the 12 components seemed relevant: lexical meaning (LM), morphosyntactic form (MF), morphosyntactic meaning (MM), cohesive form (CF), and cohesive meaning (CM). As pointed out by Purpura

(2004) that CF and CM are closely related to each other through cohesive devices within linguistic environment, the two were combined into the CFCM component. That said, this study employed LM, MF, MM, and CFCM as the four components for classifying the vocabulary and cloze MC items. The following is a brief description of the four components. Knowledge of LM enables test takers to understand and use a word's literal meaning. It encompasses the literal meaning of fixed or lexicalized expressions (e.g. How are you?). Knowledge of MF requires test takers to comprehend and produce the morphological and syntactic forms of the language (e.g. *-ed* affix, -talked). Knowledge of MM allows test takers to interpret and express meanings from inflections such as aspect and time, meanings from derivations such as negation and agency, and meanings from syntax such as those used to express attitudes or show emphasis or contrast. Finally, knowledge of CFCM permits test takers to adopt the lexical and morphosyntactic features of the language for understanding cohesion on sentence or discourse levels through cohesive devices (e.g., she, that, there), which can make a direct connection between cohesive forms and their meaning within the context (e.g. *the girl* linked to *she*).

For the reading comprehension MC items, this study adopted the classification framework of Purpura (1999), which includes two categories: reading subskill for explicit information (RSEI) and reading subskill for inferential information (RSII). According to Purpura (1999), RSEI involves a lower-level or bottom-up process of reading, where test takers are required to decode input at the lexical or syntactic level. That is, they are required to answer questions about specific information that is explicitly stated in the text and to understand synonymous words or sentences. RSII requires test takers to infer meaning from the information that is implicitly stated in the text. That is, RSII involves a higher-level, top-down or interactive process of reading by engaging test takers in processing input at the semantic and discourse levels and relating it to prior knowledge schemata.

3.3 Item classification

Although there are no precise guidelines in related literature as to the optimal number of raters for performing item classification, a minimum of three raters[7] is often recommended (e.g. Lynn, 1986; Rubio, Berg-Weger, Tebb, Lee, & Rauch, 2003). From a statistical perspective, using a larger number of raters is likely to produce more reliable results. This study required the following three qualities of a rater: (1) at least three years of English teaching experience in either senior high school or college; (2) at least a master's degree in Linguistics, English Literature, or Teaching English as a Foreign Language (TEFL); and (3) plenty of experience constructing and/or evaluating English tests. Based on these qualities, a panel of five raters[8] was assembled. Among them, four raters each have not only a PhD degree in TEFL, Linguistics, or English Literature, but also more than ten years of experience in teaching general English reading and writing at the university level. Although the fifth rater has only a master's degree in English Instruction, he has taught English in senior high school for two years and in college for three years. All five raters have some experience with item construction and classification.

A practice session was arranged on 21 July 2016 for the raters. During the session, the 51 MC items from the 2011 AST-E were used for practice. In case of any disagreement in item classification, a discussion among them was arranged to reach a consensus. A week after the practice, the actual classification took place, where all the 102 MC items from the 2015-2016 AST-E were classified independently by them[9]. The classification results were as follows: The five raters agreed on the classification of 87 items. For the 15 unresolved items, nine were allocated to the categories chosen unanimously by four raters and the other six to the categories chosen unanimously by three raters. To determine the extent to which they were consistent in their classifications, Cohen's kappa was computed to measure the inter-rater reliability. A value of 0.90 was obtained, exceeding the recommended value of 0.80 (Landis & Koch, 1977). Hence, the five raters were consistent in their item classifications.

In Table 1, the classifications of the vocabulary and cloze MC items are listed for 2015 and 2016. For 2015, ten items were classified into LM, three into MF, none into MM, and 22 into CFCM. For

2016, 13 items were classified into LM, one into MF, one into MM, and 20 into CFCM. Since the number of items classified into MF or MM for each year was too small and thus these items were not representative of the two categories, those items (i.e., items 11, 15, and 19 for 2015; items 13 and 18 for 2016) under MF and MM were excluded from this study. That is, only MC items (32 for 2015 and 33 for 2016) under LM and CFCM were tested. The classifications of the reading comprehension items into RSEI and RSII are also listed in Table 1. For 2015, 12 items were classified into RSEI and four into RSII. For 2016, ten items were classified into RSEI and six into RSII.

Table 1. Classifications of the multiple-choice items of the 2015-2016 AST-E

	2015		**2016**	
Category	**No. of items**	**Item number**	**No. of items**	**Item number**
LM	**10**	1, 4, 9, 12, 16, 22, 23, 24, 25, 30	**13**	1, 2, 3, 5, 8, 9, 10, 14, 16, 19, 22, 28, 30
MF	**3**	11,15,19	**1**	13
MM	**0**		**1**	28
CFCM	**22**	2, 3, 5, 6, 7, 8, 10, 13, 14, 17, 18, 20, 21, 26, 27, 28, 29, 31, 32, 33, 34, 35	**20**	4, 6, 7, 11, 12, 15, 17, 20, 21, 23, 24, 25, 26, 27, 29, 31, 32, 33, 34, 35
RSEI	**12**	36, 37, 38, 39, 40, 41, 43, 45, 46, 48, 49, 50	**10**	36, 37, 40, 41, 42, 44, 45, 46, 49, 50
RSII	**4**	42, 44, 47, 51	**6**	38, 39, 43, 47, 48, 51

Notes: LM = lexical meaning, MF = morphosyntactic form, MM = morphosyntactic meaning, CFCM = cohesive form and cohesive meaning, RSEI = reading subskill for explicit information, RSII = reading subskill for inferential information.

3.4 Three criteria for data analysis

To determine if the test takers' responses fitted the classifications of the five raters, a series of CFAs were applied to the two datasets using Mplus, a statistical package with a built-in function dealing specially with dichotomously scored data. For each year, CFAs were conducted first on the vocabulary and cloze sections, then on the reading comprehension section, and finally on all three sections. Three common criteria were employed to determine the model fit: (1) the values of selected global model fit indices; (2) the values of two selected psychometric property indicators; and (3) the appropriateness and interpretability of individual parameter estimates.

For criterion (1), this study used the following global model fit indices that are commonly used for model evaluation and selection: the χ2 (chi-square) test of significance and three goodness-of-fit indices – the comparative fit index (CFI), the Tucker-Lewis index (TLI), and the root mean square error of approximation (RMSEA). In respect of CFI and TLI, values greater than 0.95 represent a good fit between the data and the hypothesized model (see Hu & Bentler, 1999; Yu, 2002). In respect of RMSEA (where smaller values indicate better model fit), values less than 0.05 suggest a close fit and values between 0.05 and 0.08 an acceptable fit (see Burns & Patterson, 2000; Joreskog & Sorbom, 1993).

For criterion (2), the values of two psychometric property indicators – the composite reliability (CR) and the average variance extracted (AVE) – for each of the components identified by the raters were determined. CR serves as an overall measure of each latent factor's reliability and AVE serves to explain the amount of variance that is captured by its indicators relative to the amount due to measurement error (Fornell & Larcker, 1981). The minimum value for CR is 0.60 (Bagozzi & Yi, 1988) and that for AVE is 0.40 (Diamantopoulos & Siguaw, 2000).

For criterion (3), the values for the standardized factor loading (SFL) of each item, the R-Squared (R^2) of each item, and the correlation coefficients among the components were determined and examined for their theoretical appropriateness and interpretability. The SFL of an item is basically considered as its correlation with its underlying factor. The R-Squared (R^2) of an item, which is the squared value of SFL, reflects the amount of the variance in the item that can be explained by its specified factor. The minimum value for SFL is 0.30 (Kline, 1994) and that for R^2 is 0.20 (Bentler & Wu, 1993; Joreskog & Sorbom, 1993). According to MacKenzie, Podsakoff and Jarvis (2005), a correlation coefficient of 0.71 or less is necessary for any two components to be distinct. By employing these three criteria, it was hoped that the factor structure of the MC items of the AST-E can be established.

4 Results

4.1 Results for vocabulary and cloze MC sections

Given the classifications of the vocabulary and cloze items in Table 1, two-component and one-component models were employed one after the other to determine the degree of fit between raters' classifications and test takers' responses through CFAs. For 2015, the two-component model tested ten items from LM and 22 items from CFCM, where LM and CFCM were treated as two separate components. In Table 2, the CFA results indicate a good overall model fit by criterion (1), with $\chi 2$= 4210.35, *df* = 463, $p < 0.0001$, CFI = 0.99, TLI = 0.98, and RMSEA = 0.04. For LM, the model produced a CR value of 0.90 and an AVE value of 0.49, suggesting LM's satisfactory psychometric properties by criterion (2). For CFCM, although a fairly satisfactory value of 0.92 was obtained for CR, a slightly unsatisfactory value of 0.38 was determined for AVE. More undesirably, a correlation value as high as 0.98 was obtained between LM and CFCM, suggesting that the two-component model failed to satisfy criterion (3).

Table 2. Fit indices for the models for the multiple-choice items of the 2015-2016 AST-E

Year	Sections	No. of items	Model	$\chi 2$	Df	CFI	TLI	RMSEA
2015	V+C	32	2-component	4210.35	463	0.99	0.98	0.04
	V+C	32	1-component	4243.95	464	0.99	0.98	0.04
	V+C	30	1-component	4065.19	405	0.99	0.98	0.04
	RC	16	2-component	528.73	103	0.99	0.99	0.03
	RC	16	1-component	562.21	104	0.99	0.99	0.03
	V+C+RC	46	2-factor	5913.62	988	0.99	0.99	0.03
	V+C+RC	46	1-factor	6265.15	989	0.99	0.99	0.03
2016	V+C	33	2-component	4157.32	494	0.99	0.99	0.04
	V+C	33	1-component	4159.80	495	0.99	0.99	0.04
	V+C	31	1-component	3938.92	434	0.99	0.99	0.04
	RC	16	1-component	363.48	104	0.99	0.99	0.02
	V+C+RC	47	2-factor	5566.48	1033	0.99	0.99	0.03
	V+C+RC	57	2- factor	5972.88	1034	0.99	0.99	0.03

Notes: V = vocabulary items, C = cloze items, RC = reading comprehension items.

Given the inappropriateness of the two-component model, another CFA run was performed with a one-component model, where the 32 items from LM and CFCM were combined together. Table 2 shows that the resulting goodness-of-fit indices were satisfactory by criterion (1), with $\chi 2$= 4243.95, *df* = 464, $p < 0.0001$, CFI = 0.99, TLI = 0.98, and RMSEA = 0.04. The model produced a CR value of 0.95 and an AVE value of 0.41, both of which suggested its satisfactory psychometric properties by criterion (2). However, the SFL values for items 3 and 7 failed to attain the minimum value of 0.30 by criterion (3). Hence, one more CFA run was conducted using a one-component model without the two items. In Table 2, the new one-component model produced overall satisfactory goodness-of-fit indices with $\chi 2$= 4065.19, *df* = 405, CFI = 0.99, TLI = 0.98, and RMSEA = 0.04. In addition, the model produced a CR value of 0.96 and an AVE value of 0.44, suggesting its satisfactory psychometric properties by criterion (2). Further, the SFL values ranged from 0.33 to 0.85, with a mean of 0.64 and a standard deviation of 0.15. Although there were three items (i.e., items 9, 13, 18) with R^2 values less than the minimum value of 0.20, the SFL and the R^2 values for the most of the MC items were satisfactory by criterion (3). In sum, the above results appeared to support the one-component model (i.e., the LM + CFCM component) for the 2015 vocabulary and cloze items.

For 2016, the two-component model tested 13 items from LM and 20 items from CFCM. In Table 2, the CFA results show a good overall model fit by criterion (1), with $\chi2= 4157.32$, $df = 494$, $p < 0.0001$, CFI = 0.99, TLI = 0.99, and RMSEA = 0.04. For LM, the model produced a CR value of 0.95 and an AVE value of 0.49, suggesting its satisfactory psychometric properties by criterion (2). For CFCM, although a fairly satisfactory value of 0.87 was obtained for CR, a slightly unsatisfactory value of 0.37 was determined for AVE. Most critically, a correlation value of 0.99 was obtained between LM and CFCM, indicating that the two-component model failed to satisfy criterion (3).

Given the unsatisfactory results for the two-component model, another run of CFA was conducted based on a one-component model, where the 33 items from LM and CFCM were merged together. Table 2 shows that the resulting overall goodness-of-fit indices were quite satisfactory by criterion (1), with $\chi2= 4159.80$, $df = 495$, $p < 0.0001$, CFI = 0.99, TLI = 0.99, and RMSEA = 0.04. The model produced a CR value of 0.96 and an AVE value of 0.44, both of which suggested its satisfactory psychometric properties by criterion (2). However, by criterion (3), the SFL values of items 5 and 10 were less than the minimum value of 0.30. Hence, one more run of CFA was performed using a one-component model with the two items excluded. In Table 2, the new one-component model produced overall satisfactory goodness-of-fit indices with $\chi2= 3938.92$, $df = 434$, $p <$ 0.0001, CFI = 0.99, TLI = 0.99, and RMSEA = 0.04. In addition, the model produced a CR value of 0.96 and an AVE value of 0.47, suggesting its satisfactory psychometric properties by criterion (2). Further, the SFL values for the remaining 31 items ranged from 0.38 to 0.88, with a mean of 0.67 and a standard deviation of 0.14. Although there were three items (i.e. items 7, 16, 19) with R^2 values less than the minimum value of 0.20, all the SFL values and most of the R^2 values were satisfactory by criterion (3). In sum, the above results seemed to lend support to the one-component model (i.e. the LM + CFCM component) for the 2016 vocabulary and cloze items.

4.2 Results for reading comprehension MC sections

Given the classifications of the reading comprehension items in Table 1, two-component and one-component models were employed one after the other to evaluate the degree of fit between raters' classifications and test takers' responses through CFAs. For 2015, the two-component model tested 12 items from RSEI and 4 items from RSII. In Table 2, the CFA results for the two-component model indicate a good overall model fit by criterion (1), with $\chi2= 528.73$, $df = 103$, $p < 0.0001$, CFI = 0.99, TLI = 0.99, and RMSEA = 0.03. The model produced a CR value of 0.88 for RSEI and a CR value of 0.68 for RSII. However, in respect of AVE, a value of 0.38 was produced for RSEI and a value of 0.36 for RSII, suggesting its partial failure to meet criterion (2). More undesirably, a rather high correlation value of 0.92 was found between RSEI and RSII, indicating that the model failed to satisfy criterion (3).

Given the inappropriateness of the two-component model, another run of CFA was performed with a one-component (RSEI + RSII) model, where the 16 items from RSEI and RSII were combined. Table 2 shows that the resulting overall goodness-of-fit indices were quite satisfactory by criterion (1), with $\chi2= 565.21$, $df = 104$, $p < 0.0001$, CFI = 0.99, TLI = 0.99, and RMSEA = 0.03. The model produced a satisfactory CR value of 0.90 but a slightly smaller AVE value of 0.36, which suggested its partial failure to satisfy criterion (2). However, according to Bettencourt (2004), models with slightly smaller AVE values can still be considered acceptable if the CR values and the overall model fit indices are reasonably good. Although there were two items (i.e. items 44 and 50) with R^2 values less than the minimum value of 0.20, all the 16 SFL values exceeded 0.30 and were statistically significant. Hence, the one-component model appeared appropriate for the reading comprehension items of the 2015 AST-E.

For 2016, the two-component model tested ten items from RSEI and six items from RSII. Unexpectedly, the CFA results yielded a warning message that a non-positive definite matrix was involved, indicating the possibility of collinearity between RSEI and RSII (Gignac, 2005). Hence, the items from these two components were merged to form a one-component (RSEI + RSII) model with 16 items. As shown in Table 2, the overall goodness-of-fit results for the one-component model turned out to be quite satisfactory by criterion (1), with $\chi 2 = 363.48$, $df = 104$, $p < 0.0001$, CFI = 0.99, TLI = 0.99, and RMSEA = 0.02. In addition, the model produced a CR value of 0.88 and an AVE value of 0.49, both of which suggested its satisfactory psychometric properties by criterion (2). Further, the SFL values for all the 16 items were statistically significant with $p < 0.0001$. Although there were two items (i.e., items 38 and 49) with R^2 values less than the minimum value of 0.20, all the SFL values and most of the R^2 values were satisfactory by criterion (3). Hence, the CFA results seemed to be in support of the fit between the one-component model and the test takers' responses to the reading comprehension items of the 2016 AST-E.

4.3 Results for all three MC sections

To probe deeper into the research question of this study – what is the factor structure underlying the three MC sections of the AST-E for 2015 and 2016 – a series of CFAs were further conducted on the two datasets. Surprisingly, for both years, it turned out that the one-factor (LM + CFCM + RSEI + RSII) model appeared to be the best based on all three MC sections. Note that the usage of the word "factor" henceforth is somewhat different from the usage of the word "component" above.

For the 2015 AST-E, a two-factor model was initially tested, where the 30 vocabulary and cloze items were used to represent the first factor (LM + CFCM) and the 16 reading comprehension items to represent the second factor (RSEI + RSII). In Table 2, the CFA results show that the two-factor model seemed to provide a good fit by criterion (1), with $\chi 2 = 5913.62$, $df = 988$, $p < 0.0001$, CFI = 0.99, TLI = 0.99, and RMSEA = 0.03. For the first factor LM + CFCM, the model produced a CR value of 0.96 and an AVE value of 0.44, both of which were satisfactory by criterion (2). For the second factor RSEI + RSII, the model produced a satisfactory CR value of 0.90 but a slightly unsatisfactory AVE value of 0.36. What is worse, the correlation value between the two factors was 0.95, which far exceeded the maximum value of 0.71 by criterion (3). Hence, the two factors were combined to form a one-factor (LM + CFCM + RSEI + RSII) model. That is, this one-factor model posits that all the MC items of the AST-E measure simply one single factor or perhaps the general reading ability. In Table 2, the CFA results show that this one-factor model was basically as good as the two-factor model, with $\chi 2 = 6265.15$, $df = 989$, $p < 0.0001$, CFI = 0.99, TLI = 0.99, and RMSEA = 0.03. In addition, the CR and AVE values for this model were respectively 0.97 and 0.40, both of which were satisfactory by criterion (2). As shown in Table 3, although there were six items (i.e. items 9, 13, 18, 44, 48, 50) with R^2 values less than 0.2, all the 51 SFL values exceeded 0.30 and were statistically significant with $p < 0.0001$. The SFL mean was 0.62 and the SFL range was from 0.32 to 0.86. Taken together, the one-factor model – preferably phrased as the overall reading ability – appeared most appropriate for fitting the test takers' responses to all the MC items of the 2015 AST-E.

Similarly, a two-factor model was tested for the 2016 AST-E, where the 31 vocabulary and cloze items were used to represent the first factor (LM + CFCM) and the 16 reading comprehension items to represent the second factor (RSEI + RSII). In Table 2, the CFA results show that the two-factor model seemed to provide a good fit to the test takers' responses by criterion (1), with $\chi 2 = 5566.48$, $df = 1033$, $p < 0.0001$, CFI = 0.99, TLI = 0.99, and RMSEA = 0.03. The model produced a CR value of 0.96 and an AVE value of 0.47 for the first factor and a CR value of 0.88 and an AVE value of 0.48 for the second factor, both of which suggested satisfactory psychometric properties by criterion (2). However, a correlation value of 0.93 between the two factors was obtained, which was much larger than the maximum value of 0.71 by criterion (3). Hence, the two factors were merged to form a one-factor (LM + CFCM + RSEI + RSII) model. That is, the one-factor model posits that all the MC items of the AST-E measure simply one single factor or the general reading ability. In Table 2, the CFA results show that this one-factor model was almost as good as the two-factor model, with $\chi2= 5972.88$, $df = 1034$, $p < 0.0001$, CFI = 0.99, TLI = 0.99, and RMSEA = 0.03. Further, the CR and AVE values for this model were respectively 0.97 and 0.41, both of which were satisfactory by criterion (2). As shown in Table 4, although there were five items (i.e. items 7, 16, 19, 38, 49) with R^2 value less than 0.2, all the 51 SFL values exceeded 0.3 and were statistically significant with $p < 0.0001$. The SFL mean was 0.63 and the SFL range was from 0.37 to 0.87. Taken together, the one-factor model – or the overall reading ability – appeared to best portray the test takers' responses to all the MC items of the 2016 AST-E.

Table 3. SFLs and R^2s of the one-factor model for the 51 multiple-choice items of the 2015 AST-E

Vocabulary items				Cloze items				Reading comprehension items			
Item	**Cat.**	**SFL**	**R^2**	**Item**	**Cat.**	**SFL**	**R^2**	**Item**	**Cat.**	**SFL**	**R^2**
1	LM	0.61	0.37	11	MF			36	RSEI	0.77	0.59
2	CFCM	0.66	0.44	12	LM	0.54	0.29	37	RSEI	0.69	0.48
3	CFCM			13	CFCM	0.37	0.14	38	RSEI	0.53	0.28
4	LM	0.70	0.50	14	CFCM	0.45	0.20	39	RSEI	0.56	0.31
5	CFCM	0.55	0.30	15	MF			40	RSEI	0.60	0.36
6	CFCM	0.57	0.32	16	LM	0.64	0.41	41	RSEI	0.65	0.42
7	CFCM			17	CFCM	0.60	0.36	42	RSII	0.65	0.42
8	CFCM	0.59	0.35	18	CFCM	0.34	0.12	43	RSEI	0.52	0.27
9	LM	0.34	0.12	19	MF			44	RSII	0.32	0.10
10	CFCM	0.51	0.26	20	CFCM	0.70	0.49	45	RSEI	0.70	0.48
				21	CFCM	0.79	0.63	46	RSEI	0.68	0.46
				22	LM	0.79	0.62	47	RSII	0.66	0.44
				23	LM	0.70	0.49	48	RSEI	0.43	0.19
				24	LM	0.80	0.64	49	RSEI	0.48	0.23
				25	LM	0.80	0.65	50	RSEI	0.40	0.16
				26	CFCM	0.86	0.73	52	RSII	0.50	0.25
				27	CFCM	0.82	0.67				
				28	CFCM	0.55	0.30				
				29	CFCM	0.80	0.64				
				30	LM	0.85	0.72				
				31	CFCM	0.74	0.54				
				32	CFCM	0.52	0.28				
				33	CFCM	0.59	0.35				
				34	CFCM	0.66	0.43				
				35	CFCM	0.81	0.65				

Notes: LM = lexical meaning, MF = morphosyntactic form, CFCM = cohesive form and cohesive meaning, RSEI = reading subskill for explicit information, RSII = reading subskill for inferential information.
The blank row refers to item excluded from this study. All SFLs have a p-value that is less than 0.0001.

Table 4. SFLs and R^2s of the one-factor model for the 51 multiple-choice items of the 2016 AST-E

Vocabulary items				Cloze items				Reading comprehension items			
Item	Cat.	SFL	R^2	Item	Cat.	SFL	R^2	Item	Cat.	SFL	R^2
1	LM	0.71	0.50	11	CFCM	0.80	0.63	36	RSEI	0.48	0.23
2	LM	0.63	0.40	12	CFCM	0.48	0.23	37	RSEI	0.51	0.26
3	LM	0.69	0.47	13	MF			38	RSII	0.37	0.13
4	CFCM	0.65	0.43	14	LM	0.59	0.35	39	RSEI	0.68	0.47
5	LM			15	CFCM	0.46	0.21	40	RSEI	0.48	0.23
6	CFCM	0.63	0.40	16	LM	0.40	0.16	41	RSEI	0.62	0.39
7	CFCM	0.39	0.15	17	CFCM	0.70	0.49	42	RSEI	0.58	0.34
8	LM	0.75	0.57	18	MM			43	RSII	0.68	0.47
9	LM	0.59	0.35	19	LM	0.40	0.16	44	RSEI	0.51	0.26
10	LM			20	CFCM	0.64	0.41	45	RSEI	0.52	0.27
				21	CFCM	0.87	0.75	46	RSEI	0.45	0.20
				22	LM	0.68	0.46	47	RSII	0.65	0.43
				23	CFCM	0.75	0.56	48	RSII	0.56	0.32
				24	CFCM	0.82	0.67	49	RSEI	0.39	0.15
				25	CFCM	0.82	0.68	50	RSEI	0.60	0.36
				26	CFCM	0.71	0.50	52	RSII	0.54	0.29
				27	CFCM	0.76	0.57				
				28	LM	0.85	0.72				
				29	CFCM	0.82	0.68				
				30	LM	0.79	0.62				
				31	CFCM	0.79	0.62				
				32	CFCM	0.64	0.41				
				33	CFCM	0.63	0.40				
				34	CFCM	0.67	0.44				
				35	CFCM	0.68	0.47				

Notes: LM = lexical meaning, MF = morphosyntactic form, CFCM = cohesive form and cohesive meaning, RSEI = reading subskill for explicit information, RSII = reading subskill for inferential information. The blank row refers to item excluded from this study. All SFLs have a p-value that is less than 0.0001.

5 Discussion and conclusion

Based on the results of this study, it was concluded that item classifications by the five raters did not fit the test takers' responses to the MC items of the AST-E for 2015 and 2016. Instead, the results indicated that the three MC sections – vocabulary, cloze, and reading comprehension – together appeared to measure a single factor, namely the general English reading ability.

The finding of a single factor in this study can serve as empirical evidence in support of Oller's proposition (1976) that language ability can be accounted for by a single global factor. That is, the present study's finding seems to corroborate his unitary competence hypothesis. More specifically, he contends that language ability is basically determined by a single global factor rather than several divisible factors, and that the cognitive processing of the single global factor determines basically test takers' language test performance. More importantly, this finding lends support to the claim of CEEC that the three MC sections of the AST-E are designed to assess test takers' general reading ability.

The fact that this study started out with several components/categories identified by the raters but ended up with a single-factor model is, in a way, consistent with Henning's (1992) claim that the skills that are theoretically conceptualized as being psychologically distinct may not empirically be proved to be psychometrically distinguishable from one another. To some extent, the finding of a single factor in this study confirms the longstanding contention that divisible reading skills that some researchers have long strived to identify may not actually exist or may not be engaged by test takers during reading tests. As argued by Buck (2001), it is likely that “divisible” reading skills are simply “useful ways of describing what we do when we comprehend language” (p. 257) rather than some separate entities that actually exist within us. In fact, his claim appears to have been empirically substantiated by this study’s failure to obtain a good fit between raters’ item classifications and test takers’ responses to the MC items of the AST-E.

As a pioneering CFA research in Taiwan probing into the construct validity of the MC items of the AST-E, at least one pedagogical implication can be drawn for Taiwan high school English teachers when they attempt to improve their students’ performance on the MC items of the AST-E. The CFA results of this study showed that the correlations between the components identified by the raters were substantially high, suggesting that the MC items seemed to tap the same reading construct. That is, the identified components were so inextricably intertwined and inseparable that the three MC sections appeared to measure simply a single general reading ability. This finding implies that test takers’ performance on the three MC sections is primarily determined by the level of their general reading ability. That said, to prepare their students correctly for the MC items, English teachers are advised to focus on ways to expose them as much, and as often, as possible to meaningful and interesting English reading materials in order to improve their overall reading ability, rather than paying too much attention to teaching them different reading skills.

Given the finding of a single factor in this study, a practical implication can be drawn with respect to the construction of the MC items of future AST-Es in Taiwan. Many academics (e.g. Bachman & Palmer, 1996; Hughes, 2003) hold the views that the desirability of a test has to be balanced against practicality and that tests should be constructed so that they are as economical of time and effort as possible. That is, if two tests of different lengths measure the same language ability with about equal degree of validity, then the shorter is preferred. Given the high-stakes nature of the AST-E, test takers should be given adequate time to complete the test so as to maximize their performance. However, the fact of the matter is that they have only 80 minutes to complete the entire AST-E, which is made up of 51 MC items, two translation items, and one guided composition. Given the finding that the MC items measure essentially test takers’ overall English ability, a shorter version (i.e. reducing the MC items from 51 to a smaller number) of the AST-E would be more appropriate in terms of achieving, to a greater degree, what the MC items are purported to measure – the general English reading ability of test takers.

Because of limited resources, this study is subject to two limitations, both of which are related to raters. The first is that the finding depends considerably on the five raters’ classifications (Gable & Wolf, 1993; Rubio et al., 2003; Waltz, Strickland, & Lenz, 1991), where their subjectivity could have played a significant role in the results of this study. To deal with this limitation, item classification performed by a larger panel of raters would help minimize such subjectivity and should lead to more reliable results. The second is that, in respect of the MC items, there is undeniably a divergence in cognition between the test takers and the raters. To deal with this limitation, future studies can be supplemented with some qualitative methods, such as “think aloud” or “retrospection” (Hughes, 2003).

Given enough resources, more reliable results can possibly be obtained by minimizing the effect of the two limitations. This study was intended to explore the factor structure of the MC items of the AST-E with two objectives in mind. At this very juncture, it is timely to take account of what have been accomplished. First, the finding of a single-factor model indicates that the three MC sections measure essentially test takers’ general reading ability, which validates the claim of CEEC. Second, the finding provides valuable information to high school English teachers, with regard to how to prepare their students for the MC items, and to AST-E test constructors, with regard to how to construct proper MC items for future AST English subtests.

Notes

[1] According to Anastasi and Urbina (1997), construct validity is the extent to which an assessment measures a theoretical construct that it is supposed to measure.

[2] Classification in the present context means that each rater, to the best of his/her judgment, decides which one

[3] The three English tests are a cloze test, a dictation, and the five parts of the TOEFL test.
[4] For each of the two years (2015 and 2016), there are 5,000 test takers' responses. Each response contains 51 answers to the 51 MC items.
[5] The AST-E requires more than a 7,000-word vocabulary.
[6] To classify the vocabulary and cloze MC items, this study adopted a simplified version of Purpura's (2004) model based on the results of the study by Saito (2003), where Purpura's model provides a better classification framework for the ECPE cloze data than another popular model proposed by Hale et al. (1988).
[7] There might be a problem if too many raters (say 15) were involved. As a case in point, the study of Alderson (1990) shows that a panel of 18 raters led to considerable disagreement among them on item classifications.
[8] This study used an odd number of raters (i.e. five) instead of an even number of raters (e.g., four or six) to eliminate the possibility that a decision on item classification would be split evenly between the raters, which simply means no decision.
[9] The raters were given two weeks (29 July 2016 – 11 August 2016) to do the item classification at their home.

References

Ajideh P., & Esfandiari, R. (2009). A close look at the relationship between multiple choice vocabulary test and integrative cloze test of lexical words in Iranian context. *English Language Teaching*, *2*(3), 163–170.

Alderson, J. C. (1990). Testing reading comprehension skills. *Reading in a Foreign Language*, *6*(2), 425–439.

Anastasi, A., & Urbina, S. (1997). *Psychological Testing.* New Jersey: Prentice Hall.

Anderson, L. W. (Ed.), & Krathwohl, D. R. (Ed.). (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's educational objectives.* New York: Longman.

Bachman, L. F. (1982). The traits structure of cloze test scores. *TESOL Quarterly*, *16*(1), 61–70.

Bachman, L. F., & Palmer, A. S. (1981). A multitrait-multimethod investigation into the construct validity of six tests of speaking and reading. In A. S. Palmer, P. J. M. Groot & G. A. Trosper (Eds.), *The construct validation of tests of communicative competence* (pp. 149–165). Washington, D.C.: TESOL Publications.

Bachman, L. F., & Palmer, A. S. (1982). The construct validation of some components of communicative proficiency. *TESOL Quarterly*, *16*(4), 449–465.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford: Oxford University Press.

Bagozzi, R. P., & Yi, Y. (1988). On the evaluation of structural equation models. *Academic of Marketing Science*, *16*(1), 76–94.

Bentler, P. M., & Wu, E. J. C. (1993). *EQS/Windows user's guide*. Los Angeles: BMDP Statistical Software.

Beresova, J. (2014). Assessing grammar and vocabulary – Yes or No? *Journal of International Scientific Publications*, *8*, 158–166.

Bettencourt, L. A. (2004). Change-oriented organizational citizenship behaviors: The direct and moderating influence of goal orientation. *Journal of Retailing*, *80*(3), 165–180.

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain*. New York: David McKay.

Buck, G. (2001). *Assessing listening*. Cambridge: Cambridge University Press.

Burns, G. L., & Patterson, D. R. (2000). Factor structure of the Eyberg child behavior inventory: A parent rating scale of oppositional defiant behavior toward adults, inattentive behavior, and conduct problem behavior. *Journal of Clinical Child Psychology*, *29*(4), 569–577.

Carroll, J. B. (1975). *The teaching of French as a foreign language in eight countries*. New York: Wiley Center for Applied Linguistics.

CEEC. (2016). *Test manual of the English Subtest of the Advanced Subjects Test*. Retrieved from http://www.ceec.edu.tw/107學測英文考試說明定稿.pdf

Chen, C. C. (2009). *An analysis of the reading skills measured in reading comprehension tests on the Scholastic Achievement English Test (SAET) and the Department Required English Test (DRET)* (Unpublished master's thesis). National Taiwan Normal University, Taipei, Taiwan.

Davies, A. (2007). *An introduction to applied linguistics: From practice to theory* (2nd ed.). Edinburgh: Edinburgh University Press.

Diamantopoulos, A., & Siguaw, J. A. (2000). *Introducing LISREL: A guide for the uninitiated*. London: Sage.

Dimitrov, D. (2010). Testing for factorial invariance in context of construct validation. *Measurement and Evaluation in Counseling and Development*, *43*, 121–149.

DiStefano, C., & Hess, B. (2005). Using confirmatory factor analysis for construct validation: An empirical review. *Journal of Psychoeducational Assessment*, *23*, 225–241.

Fornell, C., & Larcker, D. F. 1981. Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research*, *18*(1), 39–50.

Fries, C. (1945). *Teaching and learning English as a foreign language*. Ann Arbor: University of Michigan Press.

Fulcher, G. (2015). *Re-examining language testing: A philosophical and social inquiry*. London: Routledge, Taylor & Francis Group.

Gable, R. K., & Wolf, M. B. (1993). *Instrument development in the affective domain: Measuring attitudes and values in corporate and school settings* (2nd ed.). Boston: Kluwer Academic Publishers.

Gardner, R. C., & Lambert, W. E. (1965). Language aptitude, intelligence, and second language achievement. *Journal of Educational Psychology*, *56*(4), 191–199.

Gignac, G. E. (2005). Evaluating the MSCEIT V2.0 via CFA: Corrections to Mayer et al. (2003). *Emotion*, *5*, 233–235.

Hale, G. A., Stansfield, C. W., Rock, D. A., Hicks, M. M., Butler, F. A., & Oller, J. W. (1988). *Multiple-choice items and the test of English as a foreign language* (TOEFL Research Report No. 26). Princeton: Educational Testing Service.

Henning, S. D. (1992). Assessing literary interpretation skills. *Foreign Language Annals, 25*, 339–344.

Hosley, D., & Meredith, K. (1979). Inter- and intra-test correlates of the TOEFL. *TESOL Quarterly, 13*(2), 209–217.

Hu, L. T., & Bentler, P. (1999). Evaluating model fit. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 76–99). London: Sage.

Hughes, A. (2003). *Testing for Language Teachers* (2nd ed.). Cambridge: Cambridge University Press.

Joreskog, K., & Sorbom, D. (1993). *LISREL 8: Structural equation modeling with the SIMPLIS Command Language*. Hillsdale: Lawrence Erlbaum Associates, Inc.

Kline, P. (1994). *An easy guide to factor analysis.* New York: Routledge.

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory Into Practice*, *41*(4), 212–218.

Lado, R. (1961). *Language testing: The construction and use of foreign language tests*. London: Longman.

Lado, R. (1964). *Language teaching: A scientific approach*. New York: McGraw-Hill.

Lan, W. H., & Chern, C. L. (2010). Using revised Bloom's taxonomy to analyze reading comprehension questions on the SAET and the DRET. *Contemporary Educational Research Quarterly*, *18*(3), 165–206.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159–174.

Lofgren, H. (1969). *Measuring proficiency in the German Language: A study of pupils in Grade 7* (Didakometry No. 25). Malmo: School of Education.

Lynn, M. (1986). Determination and quantification of content validity. *Nursing Research*, *35*, 382–385.

MacKenzie, S. B., Podsakoff, P. M., & Jarvis, C. B. (2005). The problem of measurement model misspecification in behavioral and organizational research and some recommended solutions. *Journal of Applied Psychology*, *90*(4), 710–730.

Nuttall, C. (2005). *Teaching reading skills in a foreign language.* Oxford: Macmillan Education.

Oller, J. W. (1976). Evidence for a general language proficiency factor and expectancy grammar. *Die Neueren Sprachen*, *75*, 165–174.

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

Oller, J. W., & Hinofotis, F. B. (1980). Two mutually exclusive hypotheses about second language ability: Indivisible or partially divisible competence. In J. Oller & K. Perkins (Eds.), *Research in Language Testing* (pp. 13–23). Rowley, MA: Newbury House.

Pimsleur, P., Stockwell, R., & Comrey, A. (1962). Foreign language learning ability. *Journal of Educational Psychology*, *53*, 15–26.

Purpura, J. E. (1999). *Learner strategy use and performance on language tests: A structural equation modeling approach*. Cambridge: Cambridge University Press.

Purpura, J. E. (2004). *Assessing grammar.* Cambridge: Cambridge University Press.

Rubio, D. M., Berg-Weger, M., Tebb, S. S., Lee, E. S., & Rauch, S. (2003). Objectifying content validity: Conducting a content validity study in social work research. *Social Work Research*, *27*, 94–104.

Saito, Y. (2003). Investigating the construct validity of the cloze section in the Examination for the Certificate of Proficiency in English. *Spaan Fellow Working Papers in Second or Foreign Language Assessment, 2*, 39–82.

Spearman, C. (1904). General intelligence: Objectively determined and measured. *American Journal of Psychology*, *15*, 201–292.

Spolsky, B. (1973). What does it mean to know a language; or how do you get somebody to perform his competence? In J. W. Oller & J. R. Richards (Eds.), *Focus on the Learner* (pp. 164–176). Rowley, MA: Newbury House.

Strauss, M. E., & Smith, G. T. (2009). Construct validity: Advances in theory and methodology. *Annual Review of Clinical Psychology*, *5*, 1–25.

Waltz, C. F., Strickland, O., & Lenz, E. (1991). *Measurement in nursing research* (2nd ed.). Philadelphia: FA Davis.

Weir, C. J., Vidakovic, I., & Galaczi, E. D. (2013). *Measured constructs: A history of Cambridge english language examinations 1913–2012*. Cambridge: Cambridge University Press.

Yu, C. Y. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes* (Unpublished PhD thesis). University of California, Los Angeles, CA.