The Factor Structure of Multiple-Choice Items of the English Subtest of the General Scholastic Ability Test

Wen-Ying Lin[1*]　Yu-Ling Liu[2]　Ching-Yun Yu[3]

[1*]Associate Professor, Department of English Instruction, University of Taipei

[2]English Teacher, Qing Xi Elementary School, Taoyuan City

[3]Associate Professor, Department of Psychology and Counseling, University of Taipei

Abstract

Administered annually in January by the College Entrance Examination Center (CEEC) in Taiwan, the General Scholastic Ability Test (GSAT) is a high-stakes college entrance test for high school seniors. This study aimed at validating the multiple-choice (MC) items of its English subtest (GSAT-ES) using confirmatory factor analysis (CFA). The GSAT-ES contains a total of 56 MC items grouped into three sections: vocabulary (15 items), cloze items (25 items), and reading comprehension items (16 items). In this study, two data sets (one for 2015 and one for 2016) were provided by CEEC, with each set containing 5,500 randomly-selected test-takers' responses to the 56 MC items. Performed by a panel of five raters, the vocabulary and cloze items were classified into four language components developed by Purpura (2004), and the reading comprehension items were classified into two reading sub-skills developed by Purpura (1999). Then CFAs were applied to the two sets of data. For both years, the CFA results showed that the raters' item classifications failed to fit the test-takers' responses. Instead, based on three common measures, the single-factor model best captured the characteristics of the data, suggesting that the three MC sections together appeared to tap into the general English reading ability rather than a number of divisible reading sub-traits. Finally, based on the results of the study, some pedagogical and practical implications can be drawn for English teachers and test constructors.

Keywords: confirmatory factor analysis, construct validity, Purpura's model of grammatical competence, general reading ability

*Corresponding author: Department of English Instruction, University of Taipei, No. 1, Ai-Guo West Road, Taipei 10048, Taiwan
Tel: +886-2-23113040 ext. 4901
E-mail: wylin@utaipei.edu.tw

I. Introduction

The General Scholastic Ability Test (GSAT) is a high-stakes college entrance test for high school seniors held annually in January by the College Entrance Examination Center (CEEC) in Taiwan. GSAT test scores affect critically which universities they are qualified to be admitted to. Those with high scores have a great likelihood of getting into their preferred universities, while those with low scores have to settle for less preferred universities. What is worse, extremely poor test scores may result in very undesirable outcomes: they will have to take either the other high-stakes college entrance test (i.e., Advanced Subjects Test) in July or the GSAT again next year. Hence, students, parents, and teachers alike all place high emphasis on the GSAT because it influences considerably students' future career path.

1. Multiple-Choice Items in English Subtest of the GSAT

As English is a key subject in the high school curriculum in Taiwan, the GSAT is designed to contain an English subtest (GSAT-ES), which is divided into two parts: multiple-choice (MC) questions and constructive-response questions. The first part, which is the focus of this study, contains three sections: vocabulary (15 items), cloze (15 rational and 10 banked cloze items), and reading comprehension (16 items), all of which are claimed to assess test-takers' general reading ability by CEEC (2016). For the vocabulary section, each of the 15 items contains one or two sentences, where one word is deleted and replaced with a blank. Each item is then followed by four options, from which test-takers choose one for the answer. For the rational cloze items, test-takers are given three short passages, where 15 words/phrases are removed and replaced with 15 blanks, each of which is provided with four options, from which they choose one to restore the deleted word or phrase. For the banked cloze items, one passage is given, followed with 10 blanks. For each blank, test-takers choose one from a pool of 10 options. For the reading comprehension section, test-takers are given usually four reading passages, each of which is followed by three to five MC questions.

2. Construct Validity of GSAT-ES MC Items

Given the importance of the GSAT-ES to high school students, the construct validity of its three MC sections deserves to be explored. That is, a legitimate and crucial question is whether their scores on the three sections reflect the level of their general English reading ability. In fact, although there have been several studies attempting to address the validity of the MC items, most of them have focused on just a certain item type rather than the entire three MC sections. In addition, most of the studies have tried to examine their validity by collecting only the content-related evidence through content analysis approach. For instance, Chou (2009) conducted a qualitative analysis using only data on the 25 MC cloze items. Similarly, several other studies (Lan, 2007; Liu, 2009; Lu, 2002) have also been done from the perspective of content validity to identify the reading subskills using only data on the 16 reading comprehension items. Surprisingly, no study has been conducted to examine the construct validity of all the items of the three GSAT-ES MC sections. Considered as the "superordinate form" of the various types of validity (Alderson, Clapham & Wall, 1995), construct validity refers to the extent to which an assessment measures a theoretical construct that it is supposed to measure (Anastasi & Urbina, 1997). As pointed out by Hughes (2003), content validity is not the only source of validity evidence for a measure. Specifically, he states that "content validation of the test might confirm these sub-skills were well

represented in the test … But one could still not be sure that the items in the test were 'really' measuring the sub-skills listed in the specifications" (pp. 31-32). Instead, with construct-related validity evidence obtained through construction validation, one can better understand and assure what sub-skills listed in the specification are measured by the items of the test. As such, this study was aimed to probe into the construct validity of all the items of the three MC sections, with two goals in mind. The first, of theoretical interest, was to find out the underlying language trait structure of all the MC items in the GSAT-ES. The second, of practical interest, was to obtain results that will be informative to English teachers and test constructors.

3. Research Question

Divided into two stages, this study aimed at delving into the construct validity of all the 56 MC items administered from 2015 to 2016. In the first stage, each of the items was examined and classified by five English content-expert raters into various categories based on Purpura's (2004) model of grammatical knowledge for the vocabulary and cloze items, and based on Purpura's (1999) classification for the reading comprehension items. In the second stage, a series of confirmatory factor analyses (CFAs) were carried out to determine the fit between raters' item categorization/classification and test-takers' item-by-item, dichotomously scored responses to the 56 MC items. In particular, the present study attempted to address the following research question: What is the underlying trait or factor structure of the MC items as a whole?

II. Literature Review

Over the years, there has been no unanimous agreement among researchers on the nature of general language ability. On one side are some researchers who believe that language ability is composed of a set of divisible subskills. Based on the structuralist school of linguistic, Lado (1961) proposed a "skills-and-elements" model of language proficiency. According to Lado, language proficiency consists of language skills and language elements. His belief in isolated language components led to the development of the discrete-point approach to language testing, which posits that each language component or element can be measured separately. Aside from the theoretical descriptions of the components of language ability, an extensive body of research (e.g., Carroll, 1975; Gardner & Lambert, 1965; Hosley & Meredith, 1979; Lofgren, 1969; Pimsleur, Stockwell & Comrey, 1962) has been carried out on the divisible traits or multi-dimensionalities of language ability. On the other side are researchers who hold the view that language ability consists of only one general factor (e.g., Oller, 1976, 1979, 1983; Scholz, Henricks, Spurling, Johnson & Vandenburg, 1980). For example, Oller (1976) claimed that language ability is unitary in that it is an indivisible set of interacting abilities which cannot be broken down into separate components. Specifically, Oller (1979) proposed the unitary trait hypothesis, which says that a significant portion of reliable variance in scores of the language test can be accounted for by a single trait or factor — pragmatic expectancy grammar. His firm belief is that pragmatic expectancy grammar constitutes a single, unitary language ability, which can be measured as a whole by integrative and pragmatic procedures such as cloze tests. Still, some other researchers take the view that language ability is best represented by a higher-order secondary factor and several other specific factors (e.g., Bachman, 1982; Bachman & Palmer, 1981, 1982). In particular, this view, based mainly on empirical studies using CFAs, posits that test-takers' performance on language tests is

governed by separate but correlated traits, which are in turn influenced by a single high-order factor.

Despite the differing views about the nature of overall language ability, the field of language assessment has used a common practice to assess overall reading ability — constructing a reading test that includes a few subsections and summing up the scores on the subsections to give an indication of the overall language ability. For example, the reading section of the Test of English for International Communication (TOEIC) is made up of 100 items, which in turn are grouped into various subsections, such as subsections for vocabulary, cloze, and reading comprehension. Likewise, the 56 MC items of the GSAT-ES are grouped into subsections for vocabulary, cloze, and reading comprehension. The unspoken assumption is that each of the subsections aims at testing somewhat different abilities and that they together determine the overall English reading ability of test-takers. Unfortunately, little has been known about the tenability of this assumption. In fact, not a single study along this line has been done in Taiwan. Hence, a question remains unanswered regarding what reading ability or abilities the MC items of the GSAT-ES are intended to measure. Given the high-stakes nature of the GSAT-ES, it is clearly necessary to gain a better understanding about the construct validity or the factor structure of its three MC sections.

Construct validity can be evaluated through different forms of factor analysis. Of numerous factor analysis procedures, confirmatory factor analysis (CFA) — an approach that formulates specific hypotheses of what a test measures and then examines whether or not test-takers' item-by-item response patterns agree with the a priori hypothesis — has been widely recognized as a powerful tool for extracting empirical evidence confirming hypothesized factorial structures and supporting the construct validity of a test (e.g., Dimitrov, 2010; DiStefano & Hess, 2005). As pointed out by Strauss & Smith (2009), a major strength of CFA in construct validation research is its theory-testing availability of direct comparison among various alternative models of construct relationships. In addition, by allowing correlated errors of measurement, underlying latent constructs resulting from CFA hypotheses tend to be less confounded by measurement errors than observed variables or indicators.

Despite its popularity in construct validation research, no studies to date have employed CFA to investigate the construct validity of the MC items of the GSAT-ES in Taiwan. That said, this study aimed to probe into the underlying factor structure of the vocabulary, cloze, and reading comprehension sections of the GSAT-ES, using the powerful CFA as the statistical tool. See Section III below for a brief explanation of CFA.

III. Methodology

1. Test Items

For this study, two datasets — one for 2015 and one for 2016 — were obtained from CEEC. Each dataset contained 5,500 randomly selected test-takers' item-by-tem dichotomously scored responses to the MC items of the GSAT- ES. For each of the two years, there are 56 MC items in all, including 15 vocabulary items, 15 rational cloze items, 10 banked cloze items, and 16 reading comprehension items.

2. The Instrument

What components of language ability should be considered in a language test? Purpura (2004) proposed a total of 12 components/categories to be considered for assessing English grammatical knowledge. However, an assessment of the 80 (2 x 40) vocabulary and cloze MC items showed that only five appeared to be relevant: lexical meaning (LM),

morphosyntactic form (MF), morphosyntactic meaning (MM), cohesive form (CF), and cohesive meaning (CM). In addition, given that CF and CM are closely related (see Purpura, 2004), the two were combined into the CFCM category. That said, this study used LM, MF, MM, and CFCM as the four categories for classifying the 80 items from the vocabulary and cloze sections. At this point, a brief description of these four categories is fitting. 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 enables them to comprehend the morphological and syntactic forms of the language (e.g., *-ed* affix, -talked). Knowledge of MM allows them to interpret and express meanings from inflections such as time, meanings from derivations such as negation and agency, and meanings from syntax such as those used to express attitudes or show focus or contrast. Finally, knowledge of CFCM permits them 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*).

In respect of the 32 reading comprehension MC items, this study employed the Purpura (1999) classification, which included the reading subskill for explicit information (RSEI) and that 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. Confirmatory Factor Analysis

Simply put, in confirmatory factor analysis (CFA), a researcher posits an a priori theoretical measurement model to describe or explain the relationship/connection between the underlying unobserved constructs/factors and the empirical evidence. Then he/she employs statistical fit measures to evaluate the extent to which the sample data are consistent with the hypothesized model. That is, to determine whether the sample data fit the hypothesized model. In the present study, the hypothesized model is the raters' item classifications and the data are the test-takers' responses to the 56 MC items of the GSAT-ES. The statistical fit measures used are the three measures that will be explained below. For more details on CFA, see Brown (2006), DiStefano & Hess (2005), and Harrington (2009).

4. Raters and Item Classification

Five English teachers were invited to serve as raters to examine and classify the 112 (56 for 2015 and 56 for 2016) MC items in this study. Among them, four raters have not only a PhD degree in Linguistics, Teaching English as a Foreign Language (TEFL), or English Literature, but also more than 10 years of tertiary-level English-teaching experience. Although the fifth rater has only a master's degree in English Instruction, he has taught English in high school for two years and in college for three years. Also, one of the four raters with a PhD degree taught English in junior and senior high schools for many years.

A practice session was arranged for the five raters. During the session, the 56 MC items from the

2011 GSAT-ES were used for practice. The raters classified each of the 56 items based on instructions in a practice sheet. If there were any disagreements in item classification, a discussion was arranged in order to reach a unanimous consensus. During the actual classification, all the 112 items were classified independently by the five 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 0.89 was obtained, exceeding the recommended value of 0.80 (Landis & Koch, 1977). Hence, the five raters in general were quite consistent in their item classifications.

5. Three Measures for Data Analysis

To determine if the test-takers' responses fitted the classifications of the five raters, CFA was applied to each of the two datasets using Mplus, a statistical package which contains a procedure dealing specially with dichotomously scored data. For both years, CFA was conducted first on the vocabulary and cloze items, then on the reading comprehension items, and finally on all the items. Three measures were used to assess the model fit: (1) the values of selected global model fit indices, (2) the values of some selected psychometric property indicators, and (3) the appropriateness and interpretability of individual parameter estimates.

For measure (1), this study employed the following global model fit indices that are used commonly for model evaluation and selection: the χ^2 (chi-square) test of significance and various goodness-of-fit indices, such as 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 represent better model fit), values less than 0.05 suggest a close fit and values as high as 0.08 an acceptable fit (see Burns & Patterson, 2000; Joreskog & Sorbom, 1993).

For measure (2), 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 trait'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 (see Fornell & Larcker, 1981). The minimum value for CR is 0.60 (see Bagozzi & Yi, 1988) and that for AVE is 0.40 (see Diamantopoulos & Siguaw, 2000).

For measure (3), 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 computed and examined for their theoretical appropriateness and interpretability. The SFL of an item is basically considered as its correlation with its underlying trait. Similarly, the R-Squared (R^2) of each item, which is the squared value of SFL, reflects the amount of the variance in each item that can be explained by its specified factor. The minimum value for SFL is 0.3 (Kline, 1994) and the minimum value for R^2 is 0.2 (Bentler & Wu, 1993; Joreskog & Sorbom, 1993). According to MacKenzie, Podsakoff & Javis (2005), a value of 0.71 or less for the correlation coefficient is necessary for any two components to be distinct. By applying the three measures, the underlying factor/trait structure of the 56 MC items of the GSAT-ES can basically be established.

Table 1 Classification of 2015 MC Items

Category	No.	2015
LM	16	3, 4, 7, 8, 9, 10, 13, 14, 15, 16, 17, 24, 30, 34, 38, 40
MF	3	18, 23, 26
MM	2	27, 29
CFCM	19	1, 2, 5, 6, 11, 12, 19, 20, 21, 22, 25, 28, 31, 32, 33, 35, 36, 37, 39
RSEI	9	41, 42, 44, 46, 49, 50, 51, 52, 54
RSII	7	43, 45, 47, 48, 53, 55, 56

Table 2 Classification of 2016 MC Items

Category	No.	2016
LM	12	1, 3, 4, 5, 13, 15, 22, 23, 27, 29, 30, 32
MF	2	19, 26
MM	1	24
CFCM	25	2, 6, 7, 8, 9, 10, 11, 12, 14, 16, 17, 18, 20, 21, 25, 28, 31, 33, 34, 35, 36, 37, 38, 39, 40
RSEI	9	43, 44, 45, 46, 47, 51, 52, 54, 55
RSII	7	41, 42, 48, 49, 50, 53, 56

Table 3 Fit Indices for Several Appropriate Models of 2015-2016 GSAT-ES

Year	Subsections	No of items	Model	χ^2	df	CFI	TLI	RMSEA
2015	V + C	35	1-component	5900.64	560	0.97	0.97	0.04
	V + C	34	1-component	5847.66	527	0.97	0.97	0.04
	RC	16	2-component	625.18	103	0.98	0.98	0.03
	RC	16	1-component	627.54	104	0.98	0.98	0.03
	V + C + RC	50	2-factor	8101.65	1174	0.97	0.97	0.03
	V + C + RC	50	1-factor	8440.34	1175	0.97	0.97	0.03
2016	V + C	37	2-component	4962.93	628	0.98	0.98	0.04
	V + C	37	1-component	5007.05	629	0.98	0.98	0.04
	RC	16	1-component	578.29	104	0.98	0.98	0.03
	RC	14	1-component	532.63	77	0.98	0.98	0.03
	V + C + RC	51	2-factor	7905.54	1223	0.98	0.98	0.03
	V + C + RC	51	1-factor	7940.81	1224	0.98	0.98	0.03

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

IV. Results

1. Item Classifications by the Five Raters

The classification of the vocabulary and cloze items is listed in Table 1 for 2015 and in Table 2 for 2016. For 2015, 16 items were classified into LM, three into MF, two into MM, and 19 into CFCM. For 2016, 12 items were classified into LM, two into MF, one into MM, and 25 into CFCM. As the number of items classified into MF or MM for each year was too small and thus they were not representative of the two categories, the items (i.e., items 18, 23, 26, 27 and 29 for 2015; items 19, 24 and 26 for 2016) under MF and MM were excluded from this study. That is, only items (35 for 2015 and 37 for 2016) under LM and CFCM were tested. The classification of the reading comprehension items into RSEI or RSII is also listed in Table 1 for 2015 and in Table 2 for 2016.

For each of the two years, nine were classified into RSEI and seven into RSII.

2. Result for Vocabulary and Cloze

Given the classifications of the remaining 72 vocabulary and cloze items, two-component and one-component models were used to assess the degree of fit to test-takers' responses. For 2015, the two-component model tested 16 items from LM and 19 items from CFCM as two separate components. For 2016, the two-component model tested 12 items from LM and 25 items from CFCM as two separate components.

Unexpectedly, for 2015, Mplus yielded a warning message for the two-component model that a non-positive definite matrix was involved, indicating the possibility of collinearity between LM and CFCM (Gignac, 2005). Therefore, for 2015, the items from LM and CFCM were merged to form a one-component model with 35 (16 + 19) items. That is, LM+CFCM component.

Shown in Table 3, the results for 2015 for the one-component model were quite satisfactory by measure (1), with $\chi 2$ = 5900.64, df = 560, p < 0.0001, CFI = 0.97, TLI = 0.97, and RMSE*A* = 0.04. The SFL values for all the 35 vocabulary and cloze items were statistically significant with p < 0.0001. Further, except for item 15, all other SFL values for the vocabulary and cloze items exceeded 0.3, the minimum acceptable value recommended by Kline (1994). Given that item 15 had a SFL value less than 0.3, another run of the CFA using a one-component model was conducted without item 15. Shown in Table 3, the new one-component model, with 34 items, generated similarly satisfactory goodness-of-fit results by measure (1), with $\chi 2$ = 5847.66, df = 527, p < 0.0001, CFI = 0.97, TLI = 0.97, and RMSEA = 0.04. In addition, the model produced a CR value of 0.97 and an AVE value of 0.48, both of which suggested its satisfactory psychometric properties by measure (2). Further, the SFLs for the remaining 34 items ranged from 0.3 to 0.84, with a mean of 0.69 and a standard deviation of 0.12. Although there were two items (i.e., items 13 and 30) with R^2 values less than the minimum value of 0.2, all the SFL values and most of the R^2 values were satisfactory by measure (3). To sum up, the above results were clear evidence supporting the one-component model (LM+CFCM component) for the 2015 vocabulary and cloze MC items.

For 2016, CFA was performed with a two-component model testing 12 items from LM and 25 items from CFCM. As shown in Table 3, the two-component model indicated a good overall fit by measure (1), with $\chi 2$ = 4962.93, df = 628, p < 0.0001, CFI = 0.98, TLI = 0.98, and RMSEA = 0.04. Further, the model rendered a CR value of 0.9 and an AVE value of 0.44 for LM, and a CR value of 0.97 and an AVE value of 0.54 for CFCM. All these CR and AVE values were signs of satisfactory psychometric properties by measure (2). In addition, all the SFL values and most of the R^2 values, generated by the two-component model, exceeded their respective minimum values suggested by Kline (1994).

However, a correlation as large as 0.98 was obtained between LM and CFCM, suggesting that the two-component model failed to satisfy measure (3). That said, another CFA was performed using a one-component model, where items from LM and CFCM were combined. In Table 3, the resulting goodness-of-fit of the one-component model was satisfactory by measure (1) with $\chi 2$ = 5007.05, df = 629, p < 0.0001, CFI = 0.98, TLI = 0.98, and RMSEA = 0.04. By measure (2), the one-component model produced a CR value of 0.97 and an AVE value of 0.5, both of which suggested satisfactory psychometric properties. Furthermore, the SFLs for the 37 vocabulary and cloze items were statistically

significant with $p < 0.0001$. Specifically, the SFLs ranged from 0.37 to 0.91, with a mean of 0.70 and a standard deviation of 0.13. With respect to the R^2 values, all but one item (item 30) had values greater than the minimum acceptable value of 0.2. That is, the SFLs and R^2 values generated by the one-component were satisfactory by measure (3). Taken together, similar to those for 2015, the results for 2016 indicated satisfactory evidence in support of the one-component model (with LM+CFCM) for the vocabulary and cloze MC items.

3. Result for Reading Comprehension

For the reading comprehension items, two-component and one-component models were in turn used to determine the degree of fit to the test-takers' responses using CFA. For each of the two years, the two-component model tested nine items from RSEI and seven items from RSII as two separate components.

For 2015, the CFA results in Table 3 for the two-component model indicated good overall fit by measure (1), with $\chi 2 = 625.18$, $df = 103$, $p < 0.0001$, CFI = 0.98, TLI = 0.98, RMSEA = 0.03. For RSEI, the model produced a CR value of 0.86 and an AVE value of 0.41. For RSII, it produced a CR value of 0.78 but a slightly unsatisfactory AVE value of 0.37, suggesting its partial failure to satisfy measure (2). Most undesirably, a very high correlation of 0.99 was obtained between RSEI and RSII, which not only substantiated Purpura's (1999) claim that the two components seem to be inextricably related in the reading process but also suggested that the two-component model's failure to meet measure (3). That said, another CFA test was performed using a one-component model, which lumped together all the items from RSEI and RSII. In Table 3, the goodness-of-fit indices of the one-component were quite satisfactory with $\chi 2 = 627.54$, $df = 104$, $p < 0.0001$, CFI = 0.98, TLI = 0.98, and RMSE*A* = 0.03. In addition, the one-component model produced a satisfactory CR value (0.90) but a slightly low AVE value (0.38), which suggested its partial failure to satisfy measure (2). However, according to Bettencourt (2004), models with slightly lower AVE values can still be considered acceptable if the CR values and the overall model fit indices are fairly good. Further, although there were four items (items 52, 54, 55, 56) with R^2 values less than 0.2 by measure (3), the SFL values for all 16 reading comprehension items exceeded 0.3 and were statistically significant. Hence, the one-component model — or perhaps the overall reading comprehension component — appeared appropriate for the reading comprehension items of the 2015 GSAT-ES.

Similarly, the CFA results for the 2016 reading comprehension items also seemed to be in favor of the one-component model. In fact, the two-component (i.e., RSEI and RSII) model resulted in a warning message from Mplus that a non-positive definite matrix was involved, indicating the possibility of collinearity between RSEI and RSII (Gignac, 2005). Hence, the items from the two components were lumped together to form a one-component model. As shown in Table 3, the goodness-of-fit results for the one-component model were quite satisfactory with $\chi 2 = 578.29$, $df = 104$, $p < 0.0001$, CFI = 0.98, TLI = 0.98, RMSE*A* = 0.03. However, items 41 and 51 had SFL values lower than the minimum acceptable value of 0.3. Hence, another CFA test without these two items was performed. In Table 3, the new one-component model generated similarly satisfactory goodness-of-fit results by measure (1) with $\chi 2 = 532.63$, $df = 77$, $p < 0.0001$, CFI = 0.98, TLI = 0.98, and RMSEA = 0.03. Similar to the results of the one-component model for 2015, an acceptable CR value of 0.89 but a slightly unacceptable AVE value of 0.37 were obtained.

Further, the resulting SFL values for all the remaining 14 items were statistically significant with $p < 0.0001$. Specifically, all the SFL values exceeded 0.3, ranging from 0.45 to 0.76 with a mean of 0.6 and a standard deviation of 0.1. Likewise, all the R^2 values were satisfactory by measure (3). In general, the CFA results seemed to be in support of the model fit between this one-component model and test-takers' responses to the reading comprehension MC items for 2016.

4. Result 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 2015 and the 2016 GSAT-ES — several CFAs were further performed. Surprisingly, it turned out that the one-factor model appeared to be the best based on all three MC sections for both years!

Initially, a two-factor[1] model was tested for 2015, where the 34 vocabulary and cloze items were used to assess the first factor (LM+CFCM) and the 16 reading comprehension items to assess the second factor (RSEI+RSII). The CFA results in Table 3 showed that the two-factor model seemed to provide a good fit, with $\chi 2 = 8101.65$, $df = 1174$, $p < 0.0001$, CFI = 0.97, TLI = 0.97, and RMSEA = 0.03. But the correlation between the two factors was 0.95, which was larger than the recommended value of 0.71 between two distinct factors. This finding was similar to that of Purpura's (1999) study, where a correlation coefficient of 0.98 was obtained between his two factors (lexico-grammatical factor and reading comprehension factor). Hence, the two factors (LM+CFCM and RSEI+RSII) were merged to form a one-factor model. That is, the one-factor model posited that all the MC items measures simply one factor (LM+CFCM+RSEI+RSII) or perhaps the general reading ability. In Table 3, the CFA results of this one-factor model were nearly as good as the two-factor model, with $\chi 2 = 8440.34$, $df = 1175$, $p < 0.0001$, CFI = 0.97, TLI = 0.97, and RMSEA = 0.03. In addition, the CR and the AVE values for this one-factor model was respectively 0.97 and 0.46, both of which were satisfactory by measure (2). As shown in Table 4, although there were five items (i.e., items 13, 52, 54, 55, 56) with R^2 value less than 0.2, all the 50 SFL values exceeded 0.3 and were statistically significant. In short, the one-factor model appeared most appropriate for fitting the test-takers' responses to all the MC items of the 2015 GSAT-ES.

In a similar fashion, a two-factor model was tested for 2016, where the 37 vocabulary and cloze items were used to assess the first factor (LM+CFCM) and the 14 reading comprehension items to assess the second factor (RSEI+RSII). In Table 3, the CFA results showed that the two-factor model seemed to provide a good fit to the test-takers' responses, with $\chi 2 = 7905.54$, $df = 1223$, $p < 0.0001$, CFI = 0.98, TLI = 0.98, and RMSEA = 0.03. However, a correlation of 0.95 was obtained between the two factors, which was larger than the recommended value of 0.71 between two distinct factors. Hence, the two factors were merged to form a one-factor model. In other words, the one-factor model posited that all the MC items of the GSAT-ES measures simply one factor (LM+CFCM+RSEI+RSII) or the general reading ability. In Table 3, the CFA results of this one-factor model were almost as good as the two-factor model, with $\chi 2 = 7940.81$, $df = 1224$, $p < 0.0001$, CFI = 0.98, TLI = 0.98, and RMSEA = 0.03. Furthermore, the CR and the AVE values for this model were respectively 0.98 and 0.46, both of which were satisfactory by measure (2). As shown in Table 5, although there were three items (items 30, 53, and 54) with R^2 value less than 0.2, all the 51 SFL values exceeded 0.3 and

[1] The usage of the word "factor" here is somewhat different from the usage of the word "component" above.

were statistically significant. Hence, the one-factor model — the overall reading ability — seemed to best portray the test-takers' responses to the MC items of the 2016 GSAT-ES.

Table 4 One-Factor Model of 2015 GSAT-ES

Item	Cat	SFL	R^2	Item	Cat	SFL	R^2
Vocabulary				Reading Comprehension			
1	CFCM	0.81	0.66	41	RSEI	0.69	0.48
2	CFCM	0.70	0.49	42	RSEI	0.73	0.54
3	LM	0.80	0.64	43	RSII	0.46	0.22
4	LM	0.69	0.48	44	RSEI	0.69	0.47
5	CFCM	0.59	0.35	45	RSII	0.69	0.48
6	CFCM	0.61	0.37	46	RSEI	0.52	0.27
7	LM	0.79	0.62	47	RSII	0.51	0.26
8	LM	0.70	0.49	48	RSII	0.78	0.61
9	LM	0.54	0.29	49	RSEI	0.77	0.59
10	LM	0.70	0.50	50	RSEI	0.73	0.53
11	CFCM	0.59	0.35	51	RSEI	0.65	0.43
12	CFCM	0.57	0.33	52	RSEI	0.34	0.12
13	LM	0.32	0.10	53	RSII	0.67	0.45
14	LM	0.64	0.41	54	RSEI	0.31	0.10
15	LM			55	RSII	0.39	0.15
				56	RSII	0.34	0.12
Cloze							
16	LM	0.72	0.52				
17	LM	0.57	0.32				
18	MF						
19	CFCM	0.72	0.51				
20	CFCM	0.75	0.57				
21	CFCM	0.55	0.31				
22	CFCM	0.56	0.31				
23	MF						
24	LM	0.72	0.51				
25	CFCM	0.78	0.60				
26	MF						
27	MM						
28	CFCM	0.77	0.60				
29	MM						
30	LM	0.40	0.16				
31	CFCM	0.83	0.68				
32	CFCM	0.71	0.56				
33	CFCM	0.79	0.61				
34	LM	0.76	0.57				
35	CFCM	0.65	0.43				
36	CFCM	0.75	0.57				
37	CFCM	0.76	0.57				
38	LM	0.78	0.61				
39	CFCM	0.82	0.68				
40	LM	0.81	0.65				

Notes: Blank row refers to item excluded from this study. All SFL values have a *p*-value that is less than 0.0001.

Table 5 One-Factor Model of 2016 GSAT-ES

Item	Cat	SFL	R^2	Item	Cat	SFL	R^2
Vocabulary				Reading Comprehension			
1	LM	0.65	0.42	41	RSII		
2	CFCM	0.62	0.38	42	RSII	0.62	0.39
3	LM	0.76	0.58	43	RSEI	0.60	0.36
4	LM	0.58	0.34	44	RSEI	0.78	0.61
5	LM	0.81	0.66	45	RSEI	0.53	0.29
6	CFCM	0.73	0.53	46	RSEI	0.46	0.22
7	CFCM	0.65	0.43	47	RSEI	0.59	0.35
8	CFCM	0.72	0.51	48	RSII	0.65	0.42
9	CFCM	0.68	0.47	49	RSII	0.73	0.53
10	CFCM	0.80	0.64	50	RSII	0.56	0.31
11	CFCM	0.69	0.47	51	RSEI		
12	CFCM	0.60	0.36	52	RSEI	0.70	0.50
13	LM	0.78	0.61	53	RSII	0.44	0.19
14	CFCM	0.55	0.31	54	RSEI	0.42	0.18
15	LM	0.81	0.65	55	RSEI	0.48	0.23
				56	RSII	0.53	0.29
Cloze							
16	CFCM	0.70	0.49				
17	CFCM	0.55	0.30				
18	CFCM	0.69	0.48				
19	MF						
20	CFCM	0.71	0.50				
21	CFCM	0.58	0.34				
22	LM	0.53	0.28				
23	LM	0.52	0.27				
24	MM						
25	CFCM	0.61	0.38				
26	MF						
27	LM	0.54	0.30				
28	CFCM	0.62	0.39				
29	LM	0.57	0.32				
30	LM	0.38	0.14				
31	CFCM	0.87	0.76				
32	LM	0.76	0.57				
33	CFCM	0.85	0.72				
34	CFCM	0.87	0.76				
35	CFCM	0.78	0.60				
36	CFCM	0.91	0.82				
37	CFCM	0.88	0.78				
38	CFCM	0.83	0.68				

39	CFCM	0.81	0.66
40	CFCM	0.75	0.56

Notes: Blank row refers to item excluded from this study. All SFL values have a *p*-value that is less than 0.0001.

V. Discussion and Conclusion

Based on the results of this study, it was concluded that, for both years, item classification by the raters did not fit the test-takers' responses to the MC items of the GSAT-ES. Instead, the results indicated that the three MC sections for 2015 and 2016 — the vocabulary, the cloze, and the reading comprehension — together appeared to measure a single overall factor, namely the general reading ability.

The finding of a single overall factor in this study can serve as empirical evidence in support of Oller's view (1979) that language ability can be accounted for by a single global trait, which explains sufficiently all of the common variance in language tests. That is, the present study's finding seemed to corroborate his unitary trait hypothesis. More specifically, he contends that language proficiency consists of a single overall trait rather than several distinct traits, and that the cognitive processing of the single overall trait will essentially determine test-takers' language test performance. More importantly, this finding lent support to the claim by CEEC (2016) that the three MC sections of the GSAT-ES are designed to evaluate test-takers' general reading ability.

The fact that this study started out with several components identified by the raters but ended up with a single-factor model is, in some way, consistent with Henning's (1992) claim that the skills that are theoretically regarded as being psychologically distinct may not empirically be proved to be psychometrically distinguishable from one another. To some degree, the finding of a single general factor in this study confirmed the longstanding contention that the sub-skills that some researchers have long strived to tap may not actually exist or may not be engaged by test-takers during reading test process. As argued by Buck (2001), it is likely that the sub-skills are simply "useful ways of describing what we do when we comprehend language" (p. 257) rather than something that actually exists 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 GSAT-ES MC items.

As a pioneering CFA research to delve into the construct validity of the MC items of the GSAT-ES, at least one pedagogical implication can be drawn for English teachers when they strive to improve their students' performance on the GSAT-ES MC items. The CFA results of this study revealed that the correlations between the components originally identified by the raters were substantially high, suggesting that the MC items seemed to tap the same language trait. That is, the identified components were so inextricably intertwined and inseparable that the three MC sections appeared to measure simply a single overall reading ability. This finding implied that test-takers' performance on the three sections was primarily determined by the level of their overall reading ability. That said, instead of paying too much attention to teaching different reading sub-skills, English teachers should expose their students as much as possible to meaningful and interesting English reading materials to improve their reading skill.

Given the finding of a single overall factor (i.e., the general reading ability) in this study, a practical implication can be drawn with respect to the construction of the MC items of future GSAT-ES. Many scholars (e.g., Bachman & Palmer, 1996;

Hughes, 2003) hold the views that the desirability of a test's validity 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 one is preferred. These views, in fact, make the construction of the MC items easier in terms of achieving the claim of CEEC, which is to assess test-takers' general reading ability. Given the finding of this study, GSAT-ES test constructors should strive to improve the quality of the MC items, instead of their quantity, making sure that each and every MC item measures exactly what it is supposed to measure — the general reading ability.

References

1. Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language Test Construction and Valuation*. New York: Cambridge University Press.
2. Anastasi, A., & Urbina, S. (1997). *Psychological Testing*. New Jersey: Prentice Hall.
3. Bachman, L. F. (1982). The traits structure of cloze test scores. *TESOL Quarterly, 16*(1), 61–70.
4. 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. M. Groot, & G. A. Trosper (Eds.), *The Construct Validation of Tests of Communicative Competence*, (pp. 149-165). Washington: TESOL Publications.
5. Bachman, L. F., & Palmer, A. S. (1982). The construct validation of some components of communicative proficiency. *TESOL Quarterly*, *16*(4), 449-465.
6. Bachman, L. F., & Palmer, A. S. (1996). *Language Testing in Practice: Designing and Developing Useful Language Tests*. Oxford: Oxford University Press.
7. Bagozzi, R. P., & Yi, Y. (1988). On the evaluation of structural equation models. *Academic of Marketing Science*, *16*(1), 76-94.
8. Bentler, P. M., & Wu, E. J. C. (1993). *EQS Windows User's Guide*. Los Angeles: BMDP Statistical Software.
9. Bettencourt, L. A. (2004). Change-oriented organizational citizenship behaviors: The direct and moderating influence of goal orientation. *Journal of Retailing*, *80*(3), 165-180.
10. Brown, T. A. (2006). *Confirmatory Factor Analysis for Applied Research*. New York: Guilford.
11. Buck, G. (2001). *Assessing Listening*. Cambridge: Cambridge University Press.
12. Burns, G. L., & Patterson, D. R. (2000). Factor structure of the Eyberg child behavior inventory: A rating scale of oppositional defiant behavior toward adults, inattentive behavior, and conduct problem behavior. *Journal of Clinical Child Psychology*, *29*(4), 569-577.
13. Carroll, J. B. (1975). *The Teaching of French as a Foreign Language in Eight Countries*. New York: Wiley Center for Applied Linguistics.
14. CEEC. (2016). *Test Manual of the English Subtest of the General Scholastic Ability Test*. Retrieved from http://www.ceec.edu.tw/107 學測英文考試說明定稿.
15. Chou, S. Y. (2009). A study of cloze test items in Scholastic Aptitude English Test and Required English Test (unpublished master's thesis). National Chung Cheng University, Chiayi, Taiwan.
16. Diamantopoulos, A., & Siguaw, J. A. (2000). *Introducing LISREL: A Guide for the Uninitiated*. London: Sage.
17. Dimitrov, D. (2010). Testing for factorial invariance in context of construct validation. *Measurement and Evaluation in Counseling and Development*, *43*, 121-149.
18. DiStefano, C., & Hess, B. (2005). Using confirmatory factor analysis for construct validation: An empirical review. *Journal of Psychoeducational Assessment*, *23*, 225-241.
19. Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research*, *18*(1), 39-50.
20. Gardner, R. C., & Lambert, W. E. (1965). Language aptitude, intelligence, and second language achievement. *Journal of Educational Psychology*, *56*(4), 191–199.
21. Gignac, G. E. (2005). Evaluating the MSCEIT

V2.0 via CFA: Corrections to Mayer et al. (2003). *Emotion*, *5*, 233-235.

22. Harrington, D. (2009). *Confirmatory Factor Aanalysis*. New York: Oxford University Press.
23. Henning, S. D. (1992). Assessing literary interpretation skills. *Foreign Language Annals, 25*, 339–344.
24. Hosley, D., & Meredith, K. (1979). Inter- and intra-test correlates of the TOEFL. *TESOL Quarterly, 13*(2), 209–217.
25. 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.
26. Hughes, A. (2003). *Testing for Language Teachers* (2nd ed.). Cambridge: Cambridge University Press.
27. Joreskog, K., & Sorbom, D. (1993). *LISREL 8: Structural Equation Modeling with the SIMPLIS Command Language*. Hillsdale: Lawrence Erlbaum Associates, Inc.
28. Kline, P. (1994). *An Easy Guide to Factor Analysis.* New York: Routledge.
29. Lado, R. (1961). *Language Testing: The Construction and Use of Foreign Language Tests*. London: Longman.
30. Lan, W. H. (2007). An analysis of reading comprehension questions on the SAET and DRET using revised Bloom's taxonomy (unpublished master's thesis). National Taiwan University, Taipei, Taiwan.
31. Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159-174.
32. Liu, C. N. (2009). Evaluating the reading comprehension questions of the SAET and the DRET (unpublished master's thesis). University of Taipei, Taipei, Taiwan.
33. Lofgren, H. (1969). *Measuring Proficiency in the German Language: A study of Pupils in Grade 7* (Didakometry No. 25). Malmo Sweden: School of Education.
34. Lu, J. Y. (2002). An analysis of the reading comprehension test in the English Subject Ability Test in Taiwan and its pedagogical implications (unpublished master's thesis). National Chengchi University, Taipei, Taiwan.
35. MacKenzie, S. B., Podsakoff, P. M., & Jarvis, C. B. (2005). The problem of measurement model misspecification in behavioral and organizational research and some solutions. *Journal of Applied Psychology*, *90*(4), 710-730.
36. Oller, J. W. (1976). Evidence for a general language proficiency factor: An expectancy grammar. *Die Neueren Sprachen*, *75*, 165-174.
37. Oller, J. W. (1979). *Language Tests at School*. London: Longman.
38. Oller, J. W. (1983). Evidence for a general language proficiency factor and expectancy grammar. In J. W. Oller, (Ed.), *Issues in Language Testing Research* (pp. 3–28). Rowley: Newbury House.
39. Pimsleur, P., Stockwell, R., & Comrey, A. (1962). Foreign language learning ability. *Journal of Educational Psychology*, *53*, 15–26.
40. Purpura, J. E. (1999). *Strategy Use and Second Language Test Performance: A Structural Equation Modeling Approach*. Cambridge: Cambridge University Press.
41. Purpura, J. E. (2004). *Assessing Grammar.* Cambridge: Cambridge University Press.
42. Scholz, G., Hendricks, D., Spurling, R., Johnson, M., & Vandenburg, L. (1980). Is language ability divisible or unitary? A factor analysis of twenty-two English proficiency tests. In J. W. Oller, & K. Perkins (Eds.), *Research in Language Testing* (pp. 24–33). Rowley: Newbury House.

43. Strauss, M. E., & Smith, G. T. (2009). Construct validity: Advances in theory and methodology. *Annual Review of Clinical Psychology, 5*, 1-25.
44. Yu, C. Y. (2002). Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes (unpublished doctoral thesis). University of California, Los Angeles, CA.

大學學科能力測驗英語科測驗選擇題效度的驗證

林文鶯 [1*]　劉玉玲 [2]　游錦雲 [3]

[1*]臺北市立大學英語教學系 副教授
[2]桃園市青溪國小英語老師
[3]臺北市立大學心理與諮商學系 副教授

摘　要

本研究目的乃是藉由驗證性因素分析方法，探討台灣大學學科能力測驗英語科測驗中選擇題的構念效度。具體主要研究問題是: 2015 及 2016 年學科能力測驗英語科選擇題 (包括字彙題、克漏字、及閱讀理解題等三大類)所測量的潛在特質因素結構為何?為回答此研究問題，本研究向大學考試入學中心申請 2015 及 2016 年隨機抽樣各 5,500 位學生每題選擇題答題得分記錄。本研究透過 Mplus 軟體進行驗證性因素分析，藉以確認專家將題目分類後產生的測量模型是否與考生的實徵資料相互配適，並期能更進一步找出實徵資料的最佳適配模型。研究結果顯示，專家將兩年的字彙題及克漏字題目根據 Purpura's (2004) 分類成三 或四個主要測量語言特質，也將閱讀理解題題目根據 Purpura's (1999)的研究分類成兩個主要閱讀能力特質。然而，驗證性因素分析結果顯示，專家的題目分配模型與實徵資料的適配度不是最佳的，最佳的適配度模型反而是單一整體特質模型。也就是說，本研究結果顯示，大學學科能力測驗英語科測驗選擇題乃是在測量整體英語閱讀能力。最後，本研究也根據研究結果，針對高中英語教師與大學考試入學中心編製測驗的相關人員，提出建議。

關鍵詞：驗證性因素分析、構念效度、Purpura 文法能力分類模式、整體英語閱讀能力

*聯繫作者：臺北市立大學英語教學系，臺北市中正區愛國西路一號。
Tel: +886-2-23113040 分機 4901
E-mail: wylin@utaipei.edu.tw