Received: December 25, 2011 *Accepted: February 29, 2012*

Application of Confirmatory Factor Analysis in Construct Validity Investigation: The Case of the Grammar Sub-Test of the CEP Placement Exam

Payman Vafaee[1], Nesrine Basheer[2], Reese Heitner[3]

Abstract

An important assumption in language testing is that test items or observable variables tap the underlying latent traits hypothesized in the theoretical model or constructs governing the design of the testing instrument (e.g., Shin, 2005). Accordingly, the present study sought to investigate the extent to which scores from the grammar sub-test of the Columbia University Community English Program (CEP) placement test could be interpreted as indicators of test takers' grammatical knowledge. The authors adopted Pupura's (2004) theoretical model of grammatical knowledge, which hypothesizes that grammatical knowledge consists of two underlying traits of form and meaning. To this end, the authors conducted a confirmatory factor analysis (CFA) to investigate whether there is a match between the CEP grammar test data (n= 144) and the theoretical model as hypothesized. Since the test items were not discrete point but were nested within one of four tasks (each with their own theme), by endorsing the interactionist view of construct definition, effects of these four themes (context) on individual items were also investigated. A multitrait-multimethod matrix (MTMM) model achieved the best possible model fit based on substantive and parsimony considerations. It included two underlying traits of grammatical form and meaning and four method (context) factors, and confirmed that the CEP test examined the grammatical knowledge and included the effect of context as a part of its construct. These findings support the interpretive argument presented for the construct validly of the CEP grammar test, and the appropriateness of the *explanation* inference made based on this test's scores. Further implications are discussed.

Keywords: *Constructs validity, Confirmatory factor analysis, Explanation inference, Grammatical knowledge*

1 Teachers College, Columbia University, New York, USA. Email: pv2203@tc.columbia.edu

2 Teachers College, Columbia University, New York, USA. Email: n.basheer@gmail.com

3 Teachers College, Columbia University, New York, USA. Email: rmh2157@columbia.edu

1. Introduction

1.1 Defining Grammatical Knowledge

Over the last few decades, grammatical knowledge has been conceptualized and defined in several different ways in a variety of language knowledge models. Traditionally, grammar has been viewed as a syntactic system by which words are arranged in sentences. According to Lado (1961), grammatical knowledge is based only on morphosyntactic form, for example, verb tense. However, recent research has challenged this narrow view of grammatical knowledge and has suggested that grammatical knowledge involves not only grammatical forms but also includes the meaning expressed through those forms. For example, for Rea Dickins (1991), grammatical knowledge has three dimensions: syntax, semantics, and pragmatics. She argues that the communicative nature of grammar needs to allow for the processing of "semantically acceptable syntactic forms, which in turn are governed by pragmatic principles" (p. 114). One shortcoming of this model is that grammar and language generally are considered to be the same entity, indistinguishable from one another.

Another influential model of grammatical knowledge was proposed by Larsen-Freeman (1991). According to this model, grammatical knowledge subsumes three interconnected dimensions of language. Grammar gives us the form or structure of language, but those forms are meaningless without a second dimension, semantics, and useless without a third dimension, pragmatics. Based on this model, grammatical knowledge is defined by three related components. Similar to Rea Dickins' (1991) model, in Larsen-Freeman's model, no distinction is made between language and grammatical knowledge.

In reaction to the shortcomings of these two models, Purpura (2004) distinguished between language knowledge and grammatical knowledge by considering grammatical knowledge and pragmatic knowledge as separate subcomponents of language knowledge. In Purpura's model, grammatical knowledge—once distinguished from pragmatic knowledge—in turn consists of two closely linked but distinct subcomponents: grammatical form and grammatical meaning. Based on this model, knowledge of grammatical form and knowledge of grammatical meaning (grammatical knowledge) are deployed in the use of language (pragmatics). In this way, in order to measure grammatical knowledge, grammar tests should be designed in a way to measure knowledge of both form and meaning as related but separate components.

Nevertheless, the testing of language knowledge—like language itself—should not occur in a vacuum, and the design of tests should reflect this fact. This is why some researchers (e.g., Chapelle, 1998) have adopted an interactionist perspective whereby the underlying test constructs governing test design explicitly incorporate relevant attributes of the testing context itself. Rather than attempting to ignore, or at least, minimize any number of testing factors as the result of the influence of context, such method factors of the context can be built into the model itself. According to Chapelle (1998, p. 43), "Trait components can no longer be defined in context-independent, absolute terms, and contextual features cannot be defined without reference to their impact on underlying characteristics".

In contrast to a trait-oriented construct definition of linguistic ability which strives to minimize the effects of context on performance by placing test items within a minimal discourse

context, the interactionist view of construct definition views performance as the result of traits, contextual features, and their interaction. In other words, interactionist view of construct definition will include dimensions of both trait and context in the definition of a theoretical construct, and the effect of context on performance is not considered as a construct irrelevant factor (Chapelle, 1998).

To summarize, in construct definition based on the interactionist view, the following should be considered: "ability- in language user – in context", (Chalhoub-Deville, 2003, p. 372). According to this view of construct definition, language ability and context features are closely linked, and it is almost impossible to disentangle them (Cronbach, 2002).

According to Young (2000), in the interactionist view of construct definition, contextual features as well as the test takers' abilities should be considered. Swain (2001) also supports a socially mediated cognitive representation of language ability.

In this way, models or definitions of the grammatical knowledge construct are incomplete—or more precisely in the context of assessment, models of the assessment of grammatical knowledge are incomplete—without the specification of method (context) factors as test constructs.

1.2 Purpose of the Study and Research Questions

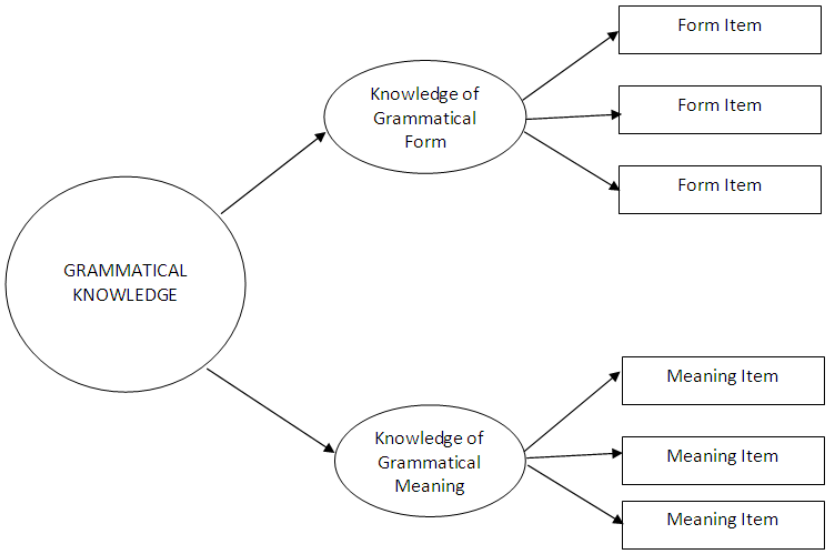
An important assumption in language testing is that test items or observable variables tap the same structural relations as those hypothesized in the theoretical model or constructs governing the design of the testing instrument (e.g., Shin, 2005). According to a prevalent view in educational measurement, a theoretical construct should serve as the basis for score interpretation of the test items (Messick, 1994). Providing justifications for the meaningfulness and usefulness of the test scores based on the underlying theoretical constructs is a piece among the whole host of evidence which is provided in a validity argument which is presented for a test's scores.

According to the current approaches toward developing validity arguments, an interpretive argument for assessment is framed (Mislevy et al., 2002, 2003). The interpretive argument consists of a chain of inferences each of which is authorized by a warrant. In such an interpretive argument, multiple types of inferences connect observations and conclusions (Kane, 1992, 2001, 2002, 2004). In the chain of inferences outlined for such an interpretive argument, Kane (1992, 2001, 2002, 2004) proposed a kind of inference (one of the inferences in the chain of the inferences in the interpretive argument) which he refers to as *explanation*. The explanation inference is made based on a theoretical construct underlying the test performance as a source of interpreting test scores as language ability of the test takers. In short, investigating the underlying theoretical construct of a test can provide justification for the appropriateness of the explanation inference which is made based on the test scores. Providing such justification for the *explanation* inference is a piece in the whole interpretive argument which is presented in the present study in the process of validating the meaningfulness and usefulness of a test's scores.

Accordingly, the purpose of the present study is to investigate the extent to which scores from the grammar sub-test of the Community English Program (CEP) placement exam

sponsored by Teachers College of Columbia University can be interpreted as indicators of test takers' grammatical knowledge as hypothesized in a theoretical construct of grammatical knowledge. To this end, we adopted Purpura's (2004) model whereby grammatical knowledge jointly consists of knowledge of grammatical form and grammatical meaning (Figure 1) (Figure 1 demonstrates three items for each of the traits of form and meaning as an example. The real number of the items depends on an actual test). In order to determine whether the CEP grammar sub-test accurately measures grammatical knowledge as hypothesized by this theoretical model, the underlying structure of the exam was investigated. In this way, we could evaluate how well the CEP grammar sub-test operationalizes Purpura's model. That is, we wish to investigate how well the CEP grammar sub-test actually tests grammatical knowledge as defined in Purpura's theoretical model. We used confirmatory factor analysis (CFA) which is a statistical technique to verify the factor structure of a set of observed variables. CFA allows the researcher to test the hypothesis that a relationship between observed variables and their underlying latent constructs exists (Kline, 2005). This section of the study can provide justification for the appropriateness of the *explanation* inference which can be made based on the CEP grammar test scores.

Figure 1. Theoretical Model of Grammatical Knowledge.



This figure illustrates the theoretical model of grammatical knowledge (Purpura, 2004) adopted for the present study with three items of form and meaning as examples of observable variables.

In addition, the CEP grammar sub-test items were not discrete point, but were nested within one of four tasks (each with their own theme). In other words, the CEP test grammar items were not standalone items independent of each other, but they were the items made based on the content of a conversation or a text each of which had its own theme. Therefore, by endorsing the interactionist view of construct definition, we also the interactionist effects of these four themes on individual items as a part of the theoretical construct; that is, we incorporated the interactionist view of construct definition in this study by taking into account the relevant context of language use in the theoretical construct to be examined. Task characteristics and context

were viewed as components of the theoretical construct based on which *explanation* inference can be made.

The four CEP grammar sub-test themes are "Visiting New York City" (Theme 1 with seven items of both grammatical form and meaning), "Business Advertisements" (Theme 2 with fourteen items of both grammatical form and meaning), "Eating Contest" (Theme 3 with five items of both grammatical form and meaning), and "Office Relationships" (Theme 4 with five items of grammatical meaning only). Therefore, in order to incorporate the effect of these themes or contexts as a construct relevant factor (according to the interactionist view of construct definition), a full latent multitrait-multimethod matrix (MTMM) model of CFA which included specification of four theme-based tasks was also evaluated.

In this way, we sought to answer the following three research questions:

1. What is the factor(ial) structure of the CEP grammar sub-test?
2. To what extent does the CEP grammar sub-test fit the hypothesized construct of grammatical knowledge?
3. To what extent do different theme-based method factors affect this model of grammatical knowledge?

2. Method

2.1 Context

The CEP courses at Teachers College of Columbia University offer English as a second language courses to adult learners from various nationalities and educational, linguistic, and socioeconomic backgrounds. Students enroll in the program for a variety of reasons. These include improving general English proficiency and oral communication with native English speakers. Students also enroll to gain a deeper understanding of the culture and environment in which they are currently residing. Before enrolling in CEP courses, students sit for a placement exam which places them into appropriate levels based on English proficiency. The exam consists of five sub-tests: listening, reading, grammar, writing, and speaking.

2.2 Participants

One hundred and forty four (144) non-native speakers of English who took the CEP speaking test in a regular administration of the test were the participants of the current study. These examinees made a diverse sample in terms of their age, native language, socio-economic status, educational background, immigration status. The majority of the examinees in this study were adult immigrants from the surrounding neighborhood or were family members of international students in the Columbia University community. In terms of their first language, a large percentage of them consisted of three languages: Japanese, Korean, and Spanish.

These participants also had different purposes for learning English, two of which were the most common ones: They either sought improvement in their everyday life communication skills, or they would like to enhance their English proficiency level to be able to continue their

education through the medium of English. Also, these participants wanted to enroll in the CEP program to gain a deeper understanding of the culture and environment in which they are currently residing.

2.3 Instrument

The instrument analyzed in this study is the grammar sub-test of the CEP placement exam including 31 four-option multiple choice items divided into four tasks and varying in the number of items tested and their associated task-based theme. Based on our theoretical model, by which grammatical knowledge is hypothesized to consist of grammatical form and meaning, the items were coded and divided into two scales of form (FR) and meaning (MG). The coding task was conducted independently by the three researchers in the present study after a complete familiarization to the theoretical model and examples presented in Purpura (2004). For the instances of disagreement between the coding of certain items, an agreement was tried to be reached through group discussion. For certain items a fourth party who was familiar with the Purpura's model was invited to resolve the disagreement. Table 1 shows an original taxonomy of these 31 test items.

Table 1. Original Item Taxonomy of the CEP Grammar Sub-Test

Scales	*Number*	*Items*
Form	19	1, 4, 5, 6, 7, 8 10, 12, 14, 15, 16, 17, 18, 19, 20, 22, 23, 25, 26
Meaning	12	2, 3, 9, 11, 13, 21, 24, 27, 28, 29, 30, 31

2.4 Procedures

The data for the present study were collected from a regular administration of the CEP exam and the test items were coded into form and meaning items, as presented in Table 1. Next, the data was organized into spreadsheets for the analysis section.

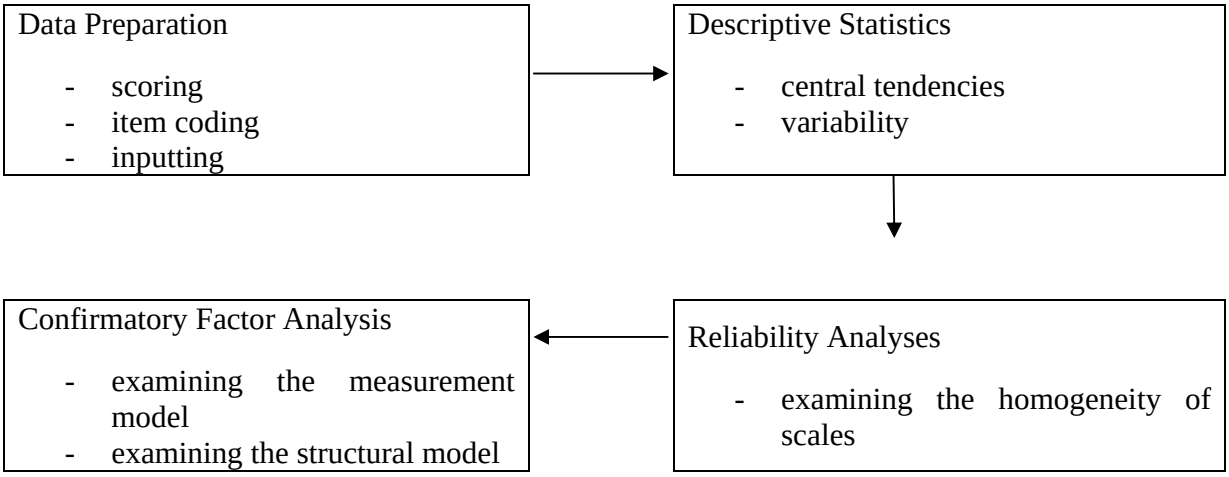
2.5 Data Analysis

The data were analyzed using SPSS Version 17.0 (SPSS Inc., 2001) and EQS Version 6.1 (Bentler & Wu, 2005). Descriptive statistics (i.e., means, standard deviations, skewness and kurtosis values) were calculated and used to investigate two important characteristics of score distribution: central tendency and variability. Reliability estimates were then calculated based on Cronbach's alpha to examine the degree of relatedness among the 31 items in the entire test, and the items within each of the two test scales.

To determine the underlying factor(ial) structure of the 31 test items, based upon (a) our hypothesized two-trait factor model of grammatical knowledge (grammatical form and grammatical meaning) , and (b) the idea of incorporating method (context) factor into the construct (involving four theme-based tasks), several CFA tests of model-fit were performed on several related CFA models to determine the extent to which these models represented the underlying trait and method (context) factors of the CEP grammar sub-test. In particular, four models of increasing model-fit were designed and tested through CFA. For each model, basic assumptions regarding model identification as well as basic data assumptions regarding univariate normality, multivariate normality, and linearity were examined as these assumptions are required for implementation of the maximum likelihood parameter estimation method utilized in CFA. In cases where these assumptions could not be satisfied, estimation was by robust maximum likelihood (Kline, 2005).

In the end, a MTMM model achieved the best possible model-fit in accordance with substantive considerations and issues of parsimony. Figure 2 represents a flow chart the procedures of the study, starting with data preparation and ending with CFA.

Figure 2. A flow chart of analyses



3. Findings

3.1 Descriptive Statistics

The descriptive statistics for each of the 31 items were calculated and are presented in Table 2. As the means for dichotomous items show the difficulty level of that item, it can be inferred that item difficulty ranged between .40 (FR22; the most difficult) and .99 (MG3; the easiest), and the standard deviations from.04 to (MG2) to.50 (FR8, FR12, FR 15, MG11).

In terms of skewness, 27of the 31 items were within the acceptable range of ± 2.5 (Bachman, 2004), indicating that these items were normally distributed. However, items MG3, MG30, MG29, and MG27 were negatively skewed with skewness of -8.39, -3.65, -2.74, and -2.61, respectively. Examining the means of these four items, which ranged between .90 and .99,

revealed that they were too easy and therefore were negatively skewed beyond the acceptable range.

As for kurtosis, five items fell outside the ± 2.5 limit. Items MG3, MG30, MG29, MG27, and MG16 had kurtosis values of 69.43, 11.50, 5.62, 4.92, and 3.29, respectively.

Table 2. Distributions for Grammar Items (N=144)

Item	*Mean*	*Std.dev*	*Skewness*	*Kurtosis*
Form				
FR1	0.72	0.45	-0.96	-1.08
FR4	0.61	0.48	-0.46	0.20
FR5	0.78	0.42	-1.40	-0.04
FR6	0.85	0.35	-2.02	2.14
FR7	0.79	0.40	-1.45	0.10
FR8	0.48	0.50	0.08	-2.02
FR10	0.67	0.47	-0.74	-1.46
FR12	0.47	0.50	0.14	-2.00
FR14	0.55	0.49	-0.19	-1.98
FR15	0.51	0.50	-0.02	-2.02
FR16	0.88	0.33	-2.29	3.29
FR17	0.74	0.43	-1.12	-0.74
FR18	0.74	0.44	-1.08	-0.84
FR19	0.76	0.43	-1.21	-0.54
FR20	0.58	0.49	-0.31	-1.92
FR22	0.40	0.49	0.40	-1.86
FR23	0.83	0.38	-1.74	1.04
FR25	0.42	0.49	0.34	-1.91
FR26	0.60	0.49	-0.43	-1.84

Meaning				
MG2	0.47	0.04	0.11	-2.01
MG3	0.99	0.11	-8.39	69.43
MG9	0.83	0.37	-1.80	1.28
MG11	0.51	0.50	-0.05	-2.02
MG13	0.55	0.49	-0.19	-1.98
MG21	0.72	0.45	-0.96	-1.08
MG24	0.65	0.47	-0.64	-1.60
MG27	0.90	0.30	-2.61	4.92
MG28	0.77	0.42	-1.30	-0.30
MG29	0.90	0.29	-2.74	5.62
MG30	0.94	0.24	-3.65	11.50
MG31	0.76	0.42	-1.25	-0.42

3.2 Internal Consistency Reliability

Cronbach's alpha reliability was calculated to investigate the internal consistency reliability of the two scales of form and meaning as well as for the entire test. Although the construct of grammatical knowledge in the current study is defined as the knowledge of grammatical form and grammatical meaning which can imply the absence of unidimentaionality, both of the two traits (form and meaning) are highly linked components of the grammatical knowledge. Therefore, it is expected that the items of grammatical form and meaning jointly tap into the grammatical knowledge of the test takers, and a high consistency between their performance is predicted. The reliability estimate for the entire test was high (0.86), signifying a high degree of homogeneity among the 31 items. The alphas for the form and meaning scales were .80 and .68, respectively. Although the reliability estimates of each of these two scales were not as high as the overall alpha, they exhibited internal consistency within the items of each scale.

3.3 Item Analysis

Given the ex post facto design of the present study, an item analysis was not carried out to actually revise any of the items. Instead, the goal of the item analysis was to investigate whether any of the items should be deleted based on the descriptive statistics. To this end, we focused on five items with skewness and/or kurtosis values beyond the acceptable ± 2.5 range. However, considering item difficulty, discrimination, and Cronbach's alpha value if item deleted, it was

decided to keep all 31 items, as the deletion of any item did not improve the overall Cronbach's alpha (Table 3).

Table 3. Item Analysis Ranked by Alpha if Item Deleted

Item	*Difficulty*	*Discrimination*	*Alpha if Item Deleted*	*Decision*
MG21	.85	-.02	.86	Keep
MG3	.99	.01	.86	Keep
MG9	.83	.14	.86	Keep
FR4	.61	.25	.86	Keep
MG13	.55	.25	.86	Keep
FR25	.42	.28	.85	Keep
FR6	.79	.24	.85	Keep
MG30	.94	.36	.85	Keep
MG29	.90	.32	.85	Keep
FR16	.88	.33	.85	Keep
FR8	.72	.33	.85	Keep
FR2	.47	.35	.85	Keep
FR19	.76	.38	.85	Keep
FR14	.55	.35	.85	Keep
MG27	.90	.43	.85	Keep
FR20	.58	.39	.85	Keep
FR5	.78	.42	.85	Keep
FR7	.48	.41	.85	Keep
FR17	.74	.42	.85	Keep
FR1	.72	.44	.85	Keep
FR23	.83	.47	.85	Keep

MG28	.77	.44	.85	Keep
MG31	.76	.45	.85	Keep
FR18	.74	.48	.85	Keep
FR10	.67	.48	.85	Keep
FR26	.60	.48	.85	Keep
FR12	.47	.47	.85	Keep
MG11	.51	.47	.85	Keep
FR15	.51	.49	.85	Keep
FR22	.40	.51	.85	Keep
MG24	.65	.58	.85	Keep

3.4 Confirmatory Factor Analysis

The next step in the analysis of the data issuing from the CEP grammar sub-test was to conduct CFA on the theoretical model hypothesized in the current study. The results of the MTMM design were also used to evaluate the role of various theme-based tasks on the assessment of grammatical knowledge.

Conducting CFA entails a number of steps. These steps are described in linear sequence below, but they are actually iterative as problems at one step require returning to earlier steps. Indeed, many models were designed in an attempt to establish a baseline model before including theme-based tasks in a full latent model. However, only three accumulative models are reported here: (a) a two-factor model with 23 variables, (b) a model also incorporating three additional cross-loading paths, and finally (c) a model also incorporating four theme-based method factors. This six-factor, 23-variable, cross-loading MTMM model produced the best model-fit. The fit criteria suggested by Hu and Bentler (1999) and used in the present study were, as follows: a ratio of chi-square to degrees of freedom (χ^2/df) less than 2.0; a comparative fit index (CFI) greater than or equal to 0.95; a standardized root mean-square residual (SRMR) less than or equal to 0.05; and a root mean-square error of approximation (RMSEA) less than or equal to 0.06.

3.4.1 Establishing a Baseline Model

We started with a model that included all 31 items of the grammar sub-test as the indicators of the two latent variables of form and meaning. The goodness-of-fit statistics for this model are as follows: χ^2/df = 1.4; CFI = 0.77; SRMR = 0.07; RMSEA = 0.05. Although the χ^2/df and the RMSEA were within the acceptable values, the CFI and the SRMR suggested poor model-fit. Therefore, we started a model improvement process by examining univariate and multivariate

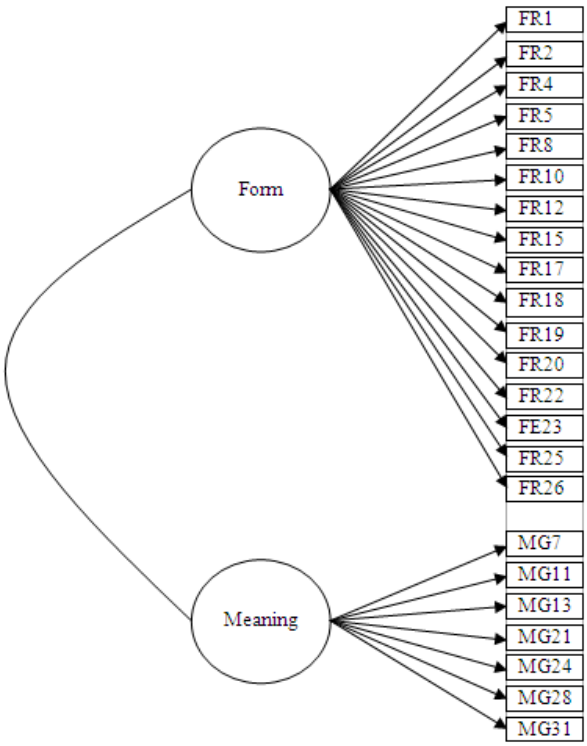
kurtosis values as well as the standardized factor loading (regression weights). This process yielded two measurement models (Model A and Model B), the better of which was then used to create the full latent model.

3.4.2. Model A

This model hypothesizes that two latent traits underlie grammatical knowledge and are predicted by 23 observable variables. Compared to the original 31-item model, Model A had the following goodness-of-fit statistics: $\chi^2/df = 0.93$; CFI = 0.90; SRMR = 0.06; RMSEA = 0.04.

Model A was designed in two steps. First, we considered the model's multivariate statistics and tentatively eliminated two kurtotic cases (cases 45 and 134). However, elimination of these outliers did not improve the model, nor did performing model estimation through robust maximum likelihood as suggested by Kline (2005). By examining univariate statistics, we decided to eliminate five items (FR16, MG3, MG27, MG29, and MG30) with kurtosis values beyond the ± 2.5 range. In addition to deleting these five kurtotic items, three items with low standardized factor loadings were also deleted. This led to a total reduction in the number of observable variables from 31 to 23 items. Figure 3 shows Model A.

Figure 3. Model A

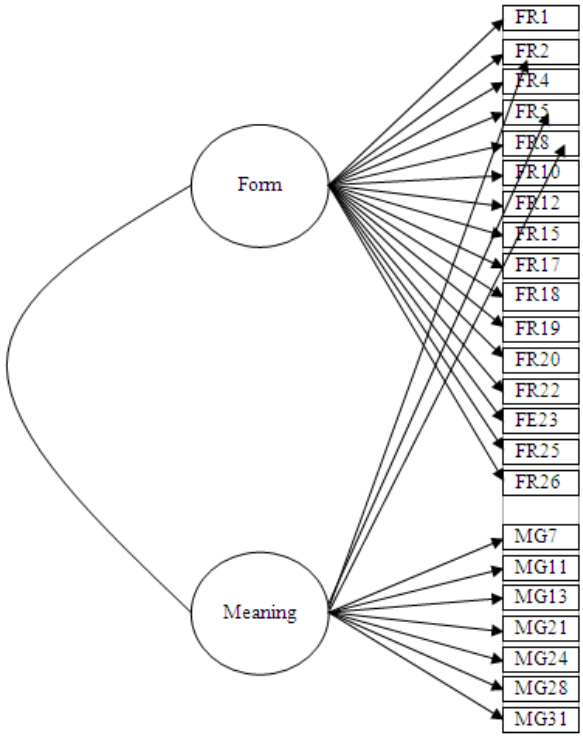


This model illustrates two latent traits of grammatical form and meaning with 23 observable variables (Error terms are not displayed in the model).

3.4.3 Model B

In addition to being statistically unacceptable in terms of CFI, Model A was theoretically unsatisfying. Based on an item coding review and substantive reasoning, we believed that some test items were likely measuring both factors. When examining all 23 items, there was a substantive reason to draw double paths between five items and both factors. However, after adding five paths, only three of these cross-loadings were kept as two of them actually decreased model-fit values. In fact, the addition of only three cross-loading paths did not improve the fit of the model (χ^2/df = 0.93; CFI = 0.90; SRMR = 0.06; RMSEA = 0.04); the addition only improved the model from a substantive perspective. Figure 4 shows the resulting Model B, which was in turn the model on which the full latent model was built, Model C.

Figure 4. Model B



This model illustrates two latent traits of grammatical form and meaning with 23 observable variables, three of which are cross loading (Error terms are not displayed in the model).

3.4.4 Model C

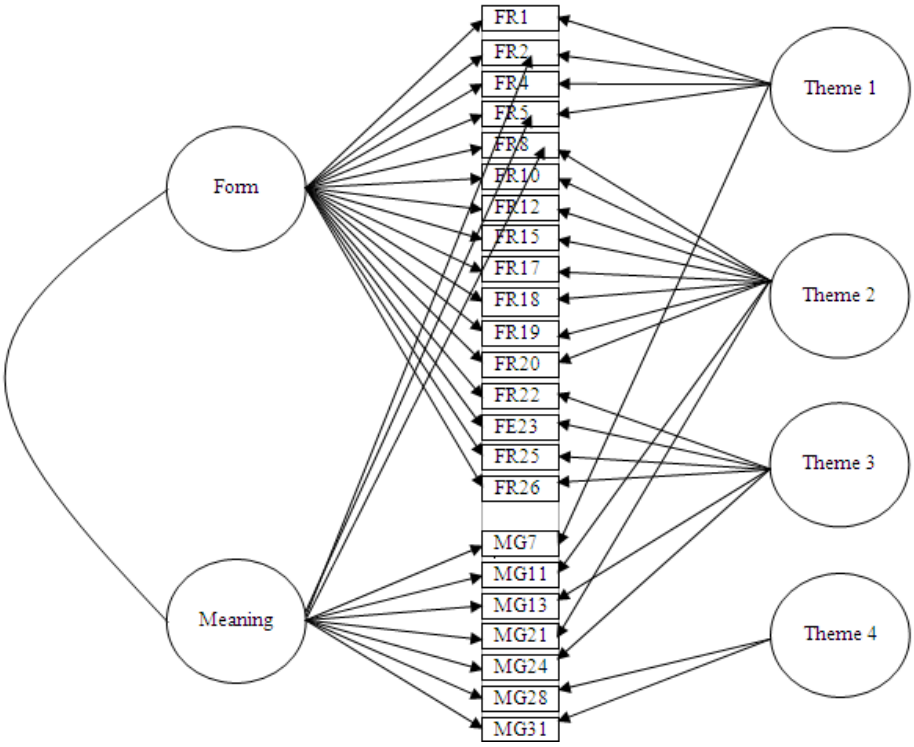
As mentioned previously, the test items were not discrete point but were nested in four themes. For this reason, we decided to design a full latent model by including four theme-based tasks as four additional method factors. This step resulted in a significant improvement of all criteria of fit, suggesting a good fit between Model C and the data (χ^2/df = 1.12; CFI = 0.95; SRMR = 0.05;

RMSEA = 0.03). Table 6 compares Models A, B and C, Figure 5 shows Model C, and Appendix A provides standardized solutions for each observed variable.

Table 4. Comparison between the Fit of Models A, B and C

Fit Criteria	Model A	Model B	Model C
ratio of chi-square to degrees of freedom (χ^2/df)	0.93	0.93	1.12
comparative fit index (CFI)	0.90	0.90	0.95
standardized root mean-square residual (SRMR)	0.06	0.06	0.05
root mean-square error of approximation (RMSEA)	0.04	0.04	0.03

Figure 5. Model C



This model illustrates two latent traits of grammatical form and meaning with 23 observable variables, three of which are cross loading. Four methods (theme or context) factors are included (Error terms are not displayed in the model).

5. Discussion and Conclusion

The present study examined the underlying factor(ial) structure of the CEP grammar sub-test. The analysis ultimately resulted in a full latent model which included a total of six factors: two underlying traits related to grammatical form and grammatical meaning, and four method (context) factors related to four theme-based tasks. In addition, the two underlying traits were correlated with one another. Although this MTMM model is premised upon a basic distinction between grammatical form and grammatical meaning, the correlation between form and meaning proved to be high—a value of .95. However, two considerations are relevant to addressing this high correlation. First, knowledge of grammatical meaning and grammatical form are no doubt both symptomatic of an overall linguistic capacity to produce and understand language. These considerations suggest moderate to high correlations, but perhaps not as high as .95. However, and second, the CEP grammar sub-test assesses both grammatical form and grammatical meaning exclusively through selected response multiple choice items which provide little opportunity—unlike ,for example, limited production items with partial credit—to more clearly distinguish between ill-formed but meaning-correct answers and vice versa.

With respect to the theme-based tasks, the full latent variable MTMM model fit the test data very well, as evidenced by the high fit indices. Moreover, the factor loadings on the two underlying traits were generally higher than the factor loadings on each of the four theme-based method factors. This observation indicates that the traits were stronger indicators than the tasks. Even though the CEP grammar sub-test specifically incorporates an over-arching theme of "cooperation and competition" (superimposed on all four tasks in an attempt to increase both the coherence and authenticity of the exam), this task-based theme did not over-influence the factor loadings for each test item. In this way, the items were more a test of grammatical meaning and grammatical form than a test of grammatical meaning and grammatical form merely through the prism of cooperation and competition.

Nevertheless, the effect of method (context) was not negligible. Given this interaction between traits and methods (context), the results of the current study can be interpreted as supporting an interactionist (Chapelle, 1998) perspective of construct definition in which knowledge of language is determined in terms of both test taker knowledge (traits) and test task characteristics (methods or context). Indeed, this MTMM model allows our three research questions to be answered.

First, the CEP grammar sub-test was found to have a two-factor(ial) trait structure of grammatical form and meaning. Second, the CEP grammar sub-test (excluding eight items) was found to fit this hypothesized two-factor model of grammatical knowledge to a high degree, but with a CFI of still less than.90. Interestingly, selectively permitting a few items to cross-load on both factors—though not increasing model-fit—did not necessarily decrease model-fit either. Though statistically neutral with regard to model-fit, cross-loading was viewed as substantively convincing, and was incorporated into the model. Finally, the MTMM model revealed that four theme-based method factors were found to affect this two-factor model of grammatical knowledge. The addition of four theme-based method factors improved the overall fit of the model.

Although the current study can contribute to recent discussions concerning the importance of both construct definitions and test task characteristics in L2 performance

assessments, it has a number of limitations. Some of these limitations issue from CFA analyses generally, and others are specific the nature of the CEP exam, its data and the present model. With respect to CFA limitations, it is possible to over-interpret goodness-of-fit statistics. This is because goodness-of-fit statistics could be the result of a few alternative possibilities—in addition to the model accurately reflecting reality: (a) the model is merely CFA equivalent to a correct model, but is itself not substantively correct; (b) the model is so complex and has so many parameters that it can hardly fail to fit any data from any data set; and (c) the model fits the sample data well but the sample data is itself unrepresentative of the population at large (Kline, 2005).

With respect to the first two CFA concerns, the current model is directly informed by previous research suggesting a substantive distinction between grammatical form and grammatical meaning (e.g., Purpura 2004) and the current model is simple relative to the complexity of other research models proposed within the field of second language (L2) assessment. With respect to the third concern, sampling concern, the number of participants (n =144) does limit the generalizability of the results. Indeed, because CFA is a data-specific statistical tool, the results of the current CFA do not necessarily generalize to other CEP grammar data with different participants. In this way, other CEP grammar data sets might be accounted for by different models with different factors, paths, and variances. Only repeated analyses of large sample sizes across multiple CEP administrations can yield a model which can be said—at least tentatively—to represent the underlying structure of the CEP grammar sub-test, and therefore, a valid instrument for basing inferences about participants' grammatical knowledge.

In sum, by accounting for all the present study's limitations, the results of this study can be cautiously used as a piece of supporting evidence in the validity argument presented for the CEP grammar test scores. Findings contribute to the appropriateness of the *explanation* inference which can be made based on the test scores. The *explanation* inference, as a piece in the chain of inferences in an interpretive argument for the validity of a test's scores, helps justifying the meaningfulness and usefulness of a test's scores by linking them to an underlying theoretical model. As the current investigation indicates, the *explanation* inference links the CEP grammar test's scores to a theoretical model which accounts for the influence of context in the construct of grammatical knowledge which is in concert with the interactionist view of construct definition. Finally, given the importance and critical role of context in the process of test development and test score validation (Chalhoub-Deville, 2001, 2003), further empirical studies are required to elaborate how elements of context can be incorporated in a theoretical construct and the process of test design. In addition, the implications of such view towards construct definition and test design on the process of test score validation need further investigation.

Acknowledgement

We would like to thank Professor James E. Purpura for what we learned from him, all his insightful comments, constructive advice and sincere support. We are also thankful to the anonymous reviewer of the current work whose detailed comments helped us greatly to learn a lot and improve our text.

References

Bachman, L. F. (2004) *Statistical analyses for language assessment*. Cambridge: Cambridge University Press.

Bentler, P. & Wu, E. (2005). *EQS 6.1 for windows user's guide.* Encino, CA: Multivariate Software, Inc.

Chalhoub-Deville, M. (2001). Task-based assessment: a link to second language instruction. In Bygate, M., Skehan, p. And Swain, M., (Eds.), *Researching pedagogic tasks: second language learning, teaching and testing* (pp. 210-28). Harlow, England: London.

Chalhoub-Deville, M. (2003). Second language interaction: current perspectives and future trends. *Language Testing*, *20*(4), 369-383.

Chapelle, C. (1998). Construct definition and validity inquiry in SLA research. In L. F. Bachman & A. D. Cohen (Eds.), *Interfaces between second language acquisition andlanguage testing research* (pp. 32-70). Cambridge: Cambridge University Press.

Cronbach, L.J. (2002). *Remaking the concept of aptitude: Extending the legacy of Richard E. Snow*. Mathwah, NJ: Lawrence Erlbaum.

Hu, L. & Bentler, P. (1999). Cutoff criteria for fit indexes covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*, 1-55.

Kane, M. (1992). An argument-based approach to validity. *Psychological Bulletin*,

112, 527–535.

Kane, M. (2001). Current concerns in validity theory. *Journal of Educational*

Measurement, 38(4), 319–342.

Kane, M. (2002). Validating high-stakes testing programs. *Educational Measurement:*

Issues and Practice, 21(1), 31–41.

Kane, M. (2004). *Certification testing as an illustration of argument-based validation.*

Measurement: Interdisciplinary Research and Perspectives, 2(3), 135–170.

Kline, R. (2005). *Principles and practice of structural equation modeling*. New York: The Guilford Press.

Lado, R. (1961). *Language testing*. New York: McGraw-Hill.

Larsen-Freeman, D. (1991). Teaching grammar. In M. Celce-Muria (Ed.), *Teaching English as a second or foreign language*(pp. 279-296). Boston: Heinle and Heinle Publishers.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2002). Design and analysis in task-based language assessment. *Language Testing*, *19*, 477-496.

Mislevy, R.J., Steinberg, L.S., & Almond, R.G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, *1*, 3-67.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. Educational Researcher, *23*(2), 13-23.

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

Rea-Dickins, P. (1991).What makes a grammar test communicative? In J. C. Alderson & B. North (Eds.), *Language testing in the 1990s: The communicative legacy* (pp. 12-35). New York: HarperCollins.

Shin, S. (2005). Did they take the same test? Examinee language proficiency and the structure of language tests. *Language Testing*, *22*, 31-57.

Swain, M. (2001). Examining dialogue: another approach to content specification and to validating inferences drawn from test scores. *Language Testing, 18*, 275-302.

SPSS Inc. (2001). SPSS Base 12.0 for Windows[Computer Software]. Chicago: SPSS Inc.

Young, R.F. (2000, March). Interactional competence: challenges for validity. Paper presented at the Language Testing Research Colloquium, Vancouver, Canada.

Appendix A

Standardized solution

STANDARDIZED SOLUTION:					R-SQUARED
FR1 =V16 =	.475*F1	+ .455*F3	+ .753 E16		.432
FR2 =V17 =	1.520*F1	- 1.156*F2	- .051*F3	+ .890 E17	.207
FR4 =V18 =	.280*F1	- .101*F3	+ .955 E18		.089
FR5 =V19 =	2.328*F1	- 1.897*F2	+ .259*F3	+ .753 E19	.433
MG7 =V20 =	.228*F2	+ .393*F3	+ .891 E20		.206
FR8 =V21 =	.392*F1	+ .059*F2	+ .091*F4	+ .889 E21	.210
FR10 =V22 =	.473*F1	+ .512*F4	+ .717 E22		.486
MG11 =V23 =	.536*F2	+ .060*F4	+ .842 E23		.291
FR12 =V24 =	.470*F1	+ .309*F4	+ .827 E24		.316
MG13 =V25 =	.264*F2	+ .508*F5	+ .820 E25		.328
FR15 =V27 =	.523*F1	+ .016*F4	+ .852 E27		.274
FR17 =V28 =	.454*F1	+ .050*F4	+ .890 E28		.208
FR18 =V29 =	.541*F1	- .121*F4	+ .832 E29		.307
FR19 =V30 =	.349*F1	+ .586*F4	+ .731 E30		.465
FR20 =V31 =	.396*F1	- .011*F4	+ .918 E31		.157
MG21 =V32 =	.372*F2	+ .137*F4	+ .918 E32		.157
FR22 =V33 =	.574 F1	+ .205*F5	+ .793 E33		.371
FR23 =V34 =	.559*F1	- .242*F5	+ .793 E34		.371
MG24 =V35 =	.634 F2	+ .034*F5	+ .773 E35		.403
FR25 =V36 =	.347*F1	- .000*F5	+ .938 E36		.121
FR26 =V37 =	.510*F1	+ .244*F5	+ .825 E37		.320
MG28 =V38 =	.475*F2	+ .277*F6	+ .835 E38		.302
MG31 =V39 =	.489*F2	+ .455*F6	+ .744 E39		.446