Examining the Structure of a Writing Assessment Using Confirmatory Factor Analysis

Taejoon Park (Korea Institute for Curriculum & Evaluation)


The purpose of this study was twofold: (a) to examine the factorial structure of a writing assessment administered to a group of English learners, and (b) to examine the extent to which the test takers’ performance on the writing assessment is affected by the underlying traits (i.e., the aspects of writing ability being tabbed into by the analytic rating scales) and the test methods (e.g., writing tasks and raters) being used to elicit and score test performance. A series of confirmatory factor analysis (CFA) models were tested for the purpose of seeking the best representation of the underlying trait and method factor structure for the measurement design of the writing assessment. The examination of the factor loadings of the final CFA model revealed that the method factor loadings associated with writing tasks were higher than the trait factor loadings for all but one of the observed variables, indicating the impact of strong task effects on the test takers’ performance on the writing assessment. The implications of the results are discussed in terms of the role of tasks in writing performance assessments.

1. INTRODUCTION

Performance assessment, which can be defined as the direct, systematic observation of an actual student performance and the rating of that performance according to previously established performance criteria, has attracted considerable attention in the field of language testing (e.g., Bachman, 2002; Kondo-Brown, 2002; Lumley, 2002; Lynch & McNamara, 1998; McNamara, 1996; Mislevy,
Compared to traditional multiple-choice tests, performance assessments are often regarded as a more valid form of language assessment because test takers’ performance derived from performance-based tasks of speaking or writing comes closer to performance as it would be in a non-test situation, and inferences made about a student’s language ability from the actual performance are presumed to be more trustworthy (Bachman, Lynch, & Mason, 1995; Bachman & Palmer, 1996). However, the apparent advantage of performance assessments over highly standardized multiple-choice tests comes at a price. As noted by Bachman et al. (1995), performance assessments bring with it potential variability in tasks and rater judgments, as sources of potential measurement error. That is, the complexity and multi-facetedness of language performance assessments inevitably introduce a range of factors that may influence the students’ observed score. In an essay test, for instance, a number of sources of variance other than students’ writing ability may contribute to the variance in test scores. As Weigle (2004) notes, two main sources of variance in an essay test are raters (e.g., Kondo-Brown, 2002; Lumley, 2002; O’Loughlin, 1992; Weigle, 1994, 1998) and tasks (e.g., Bridgeman, Morgan, & Wang, 1997; Engelhard, Gordon, & Gabrielson, 1992; Powers & Fowles, 1999). To better understand the complex nature of writing performance assessment, the sources of variance described above call for a systematic analysis in which the effects of these different sources of variability can be disentangled. For example, if a writing test involves the evaluation of two tasks scored by two independent raters, one could estimate internal consistency from classical test theory, such as Cronbach's alpha, for the four scores assigned to each test taker (i.e., 2 tasks x 2 raters x number of test takers). However, this would tell us little about the sources of variance in the measurement. In other words, we would not know which facets of the measurement contribute substantially to the observed score variance. Is it the examinee’s writing ability, or is it characteristics of the

50 Taejoon Park
writing task, the rater, the rubric, or an interaction among any of these facets? We need to gain clearer insight into the amount of variance introduced by the different facets of measurement to improve our understanding of the complex construct of writing ability and also to support decisions made based on the scores from writing performance assessments.

In this regard, the purpose of this study was two fold: (a) to examine the factorial structure of a writing assessment administered to a group of English learners, and (b) to examine the extent to which the test takers’ performance on the writing assessment is affected by the underlying trait (i.e., the aspects of writing ability being tabbed into by the analytic scoring rubrics) and the test methods (i.e., writing tasks and raters) used to elicit and score test performance.

II. LITERATURE REVIEW

In the context of writing performance assessment, the process of investigating the validity of score-based interpretations can be described in terms of building a validation argument and collecting evidence in support of that argument. In this validation framework, the validation argument we make for a particular score interpretation consists of claims and counterclaims about specific factors that are likely to affect performance on a test (Bachman, 2004; Chapelle, 1998; Kane, 1992, 2002; Kane, Crooks, & Cohen, 1999). According to Kane (1992), the central claim that we make when we use a language test is that performance on test tasks is affected primarily by the language ability we want to measure. The primary articulation of this claim is the definition of the construct to be measured. However, researchers in the field of language testing have long recognized that performance on language tests is also affected by factors other than the language ability of interest (e.g., Brown, Hilgers, & Marsella, 1991; Carlson, Bridgeman, Camp, & Wanders, 1985; Lumley, 2002; Powers & Fowles, 1999). Kane (2002) argues that the potential effects of other factors that may affect performance
on a language test need to be articulated as *counterclaims* in a validation argument. In Kane’s framework of validation, therefore, the examination of validity of score-based interpretations depends on our ability to distinguish the effects of the abilities we want to measure from the effects of other factors on test scores. That is, if we want to investigate how valid our score-based interpretations are, we must begin with a set of definitions of the abilities we want to measure, and of “the other factors” that we expect to affect test scores. Bachman (1990) grouped these factors into the following broad categories: (1) test method facets, (2) attributes of the test takers that are not considered part of the language abilities we want to measure, and (3) random factors that are largely unpredictable and temporary. Test method facets are of particular importance in language testing because the correspondence between the characteristics of test methods used to elicit test performance and the features of language use contexts will have a direct effect on the authenticity of the test and test tasks (Bachman, 1990). That is, in general, the closer the correspondence between the characteristics of the test method and the essential features of language use contexts, the more authentic the test task will be for test takers. Bachman and Palmer (1996) presents a framework for characterizing the facets of test method that affect performance on language tests. The five major categories of test method facets are: (1) the testing environment, (2) the testing rubric, (3) the nature of the input the test taker receives, (4) the nature of the expected response to that, and (5) the relationship between input and response.

Attributes of test takers that are not related to language ability include individual characteristics such as cognitive style and knowledge of particular content areas, and group characteristics such as sex, race, and ethnic background. Like test method facets, these attributes can have systematic effects on test performance and thus are considered as “potential sources of error in our measurement of language abilities” (Bachman, 1990, p.164).

In addition to these systematic sources of error, an examinee’s test
score is affected to some degree by unsystematic or random factors. These include unpredictable and largely temporary conditions, such as the test takers’ mental alertness or emotional state, and uncontrolled differences in test method facets, such as changes in the test environment from one day to the next (Brennan, 1992, 2001; Kane, 1992, 2002; Kane, Crooks, & Cohen, 1999).

The factors that affect test scores as described above are represented graphically in Figure 1 below.

Figure 1. Factors that Affect Language Test Scores (Bachman, 1990, p.165)

In sum, a primary purpose of using language tests such as a writing test is to make score-based inferences about test takers’ language ability. As Bachman (1990, 2002) and Kane (1992, 2002) pointed out, to the extent that an examinee’s test score is affected by test method facets, personal attributes other than the abilities we want to measure, and random factors, the validity of inferences made about his or her level of language ability on the basis of the test score will be undermined. Therefore, the interpretations and inferences we make based on test scores must be justified in a validation study by investigating the extent to which we can support our claim that students’ performance on test tasks is primarily
affected by the language ability we intend to measure.

III. METHODOLOGY

1. Participants & Instruments

The participants in this study were 280 high school students who took a writing test consisting of two writing tasks. The background information collected from the students showed that 58% of the test takers were female while 42% were male.

The two writing tasks used in this study represented two different intentions or purposes of writing (i.e., “writing to keep in touch” and “writing to convince/persuade). The rationale behind using the two different types of writing tasks was to achieve appropriate construct relevance and construct representativeness.

Students’ responses to the two writing tasks were scored by eight high school English teachers using analytic scoring scales, consisting of four subscales: task fulfillment, content control, organizational control, and language control. Each subscale was divided into six-point bands (ranging from 0 to 5) with descriptors for each band. The four subscales were carefully selected based on the aspects of writing that were found to be critical in previous research (e.g., Bridgeman & Carlson, 1983; Canesco & Byrd, 1989; Horowitz, 1987). The design of the rating scales was also informed by existing writing rating scales such as the TOEFL Internet-based test (iBT) rating scales. The main reason for using the analytic rating scales was to obtain a profile of scores on different aspects of writing so as to better understand the construct of writing ability being tapped into in writing performance assessment.

2. Procedures

After collecting the written responses from the test takers, eight raters who were all high school English teachers scored them. Although most of the
teachers had experience with scoring students’ writing, they were all given a program of rater training immediately before the actual scoring session. The training included an orientation to the writing test, a discussion of the analytic scoring rubrics, rating practice, and a discussion of pre-scored writing samples that represented the whole range of the scoring rubrics. During the actual scoring session, each response was rated by two teachers independently using the analytic scoring rubrics, resulting in the final rating pattern used to estimate students’ writing ability being composed of eight ratings (2 raters x 4 scoring domains).

In this study, no adjudication was made during the scoring session to resolve discrepancies in ratings from the two independent raters. In other words, even though one rater’s score was more than two points above or below the other rater’s score on a certain domain, adjustments were not made.

3. Computer Equipment

SPSS version 10.0 for PC (SPSS Inc., 2000) was used for computing descriptive statistics, inter-rater reliability, and internal consistency reliability. Confirmatory factor analysis (CFA) was conducted using the computer program EQS version 5.7b for PC (Bentler & Wu, 1998).

4. Statistical Analyses

First, descriptive statistics were computed to check if the scores in each of the four domains are normally distributed. Then inter-rater reliability was computed in order to estimate the degree of agreement between the two independent raters. Internal consistency reliability was also computed to examine how the four domains of the scoring rubrics performed as a group.

In addition, confirmatory factor analysis, which allows us to define models that posit a priori trait and method factors and test the ability of such models to fit the data based on substantive theory, was performed. In the present study, a series of substantively plausible
CFA models were tested to seek the best representation of the underlying trait and method factor structure of the writing assessment consisting of two writing tasks. To this end, a baseline CFA model was first established based on substantive theory of writing performance assessment. Then, the relative fit of the baseline model and other substantively plausible CFA models were compared by conducting chi-square difference tests. In addition to the chi-square difference tests, the competing CFA models hypothesized in this study were compared using three widely used goodness of fit indexes: comparative fit index (CFI), standardized root mean square residual (SRMR), and root mean square error of approximation (RMSEA).

In the present study, based upon the multiple fit indexes discussed above as well as substantive theory, a final model that would best represent the underlying trait and method structure of the writing assessment was selected from a number of substantively plausible competing CFA models. The parameter estimates for the final model were then examined to investigate the extent to which the students’ scores on the writing test were affected by the underlying traits (i.e., the aspects of writing ability being tapped into by the four analytic rating scales) and the test methods (i.e., the tasks and raters) used to elicit and score test performance.

IV. RESULTS

1. Descriptive Statistics

The descriptive statistics for the 16 measured variables (4 scoring criteria x 2 tasks x 2 raters) are presented in Table 1. The means ranged from 2.35 to 3.42 and the standard deviations ranged from 1.10 to 1.56. All values for skewness and kurtosis were within the accepted limits (i.e., greater than −3 and less than 3), indicating that the students’ scores on the writing test appeared to be univariately normally distributed. Table 1 presents two notable patterns in the
mean ratings for the two tasks across the four scoring criteria (i.e.,
task fulfillment, content control, organizational control, and language
control). First, the scores assigned by two independent raters were
very similar across all the rating scales, indicating a high degree of
agreement between the two ratings. Second, the students tended to
do better on the first task (i.e., writing to keep in touch) than on the
second task (i.e., writing to convince/persuade) across all the scoring
criteria.

Table 1. Descriptive Statistics of Measured Variables

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Task</th>
<th>Rater</th>
<th>Mean</th>
<th>S.D.</th>
<th>Skewness</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task fulfillment</td>
<td>1</td>
<td>1</td>
<td>3.42</td>
<td>1.50</td>
<td>-.67</td>
<td>-.47</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>2</td>
<td>2.78</td>
<td>1.56</td>
<td>-.48</td>
<td>-.27</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2.87</td>
<td>1.54</td>
<td>-.32</td>
<td>-.85</td>
</tr>
<tr>
<td>Content control</td>
<td>1</td>
<td>1</td>
<td>2.96</td>
<td>1.30</td>
<td>-.72</td>
<td>-.17</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>2</td>
<td>3.02</td>
<td>1.32</td>
<td>-.63</td>
<td>-.27</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2.60</td>
<td>1.45</td>
<td>-.25</td>
<td>-.83</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2.63</td>
<td>1.44</td>
<td>-.21</td>
<td>-.79</td>
</tr>
<tr>
<td>Organizational control</td>
<td>1</td>
<td>1</td>
<td>3.10</td>
<td>1.32</td>
<td>-.60</td>
<td>-.15</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>2</td>
<td>3.03</td>
<td>1.32</td>
<td>-.43</td>
<td>-.48</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2.41</td>
<td>1.47</td>
<td>-.06</td>
<td>-.90</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2.44</td>
<td>1.46</td>
<td>-.03</td>
<td>-.82</td>
</tr>
<tr>
<td>Language control</td>
<td>1</td>
<td>1</td>
<td>2.92</td>
<td>1.10</td>
<td>-.55</td>
<td>.17</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>2</td>
<td>2.80</td>
<td>1.20</td>
<td>-.52</td>
<td>-.31</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2.35</td>
<td>1.27</td>
<td>-.25</td>
<td>-.74</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2.42</td>
<td>1.26</td>
<td>-.26</td>
<td>-.79</td>
</tr>
</tbody>
</table>

In addition to the univariate normality checked above, other
distributional assumptions were also examined in order to ascertain
the data are appropriate for confirmatory factor analysis, which is the
main analysis tool used in this study. The examination of multivariate
normality, the presence of univariate and multivariate outliers, and the
linearity of relations among the 16 measured variables all indicated
that the data met the assumptions required for conducting CFA.

Table 2 presents the inter–rater reliability correlations between
two independent raters for each of the four subscales. Because the
two independent ratings were considered as ordinal variables,

Examining the Structure of a Writing Assessment
Using Confirmatory Factor Analysis 57
inter-rater reliability was obtained by computing Spearman rank-order correlation coefficients. These values were adjusted by using the Spearman-Brown Prophesy Formula, following the recommendation of Henning (1987). The values of inter-rater reliability ranged from .74 to .88, suggesting that there existed some disagreement between the two independent raters, although there was a fair amount of consistency in assigning scores to the students’ responses to the two writing tasks.

| Table 2. Inter-rater Reliability |
|-------------------------------|-------------------|
| Rating scale                  | Task              | Inter-rater reliability |
| Task fulfillment              | 1                 | .80                    |
|                               | 2                 | .88                    |
| Content control               | 1                 | .78                    |
|                               | 2                 | .87                    |
| Organizational control        | 1                 | .74                    |
|                               | 2                 | .87                    |
| Language control              | 1                 | .78                    |
|                               | 2                 | .85                    |

Coefficient alpha estimates of reliability for the two writing tasks are presented in Tables 3 and 4. As shown in these tables, all the subscales displayed acceptable internal consistency (α > .70), suggesting that the same construct (i.e., writing ability) was being measured on each subscale.

<table>
<thead>
<tr>
<th>Table 3. Internal Consistency Reliability for Task 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rating scale</td>
</tr>
<tr>
<td>Task fulfillment</td>
</tr>
<tr>
<td>Content control</td>
</tr>
<tr>
<td>Organizational control</td>
</tr>
<tr>
<td>Language control</td>
</tr>
</tbody>
</table>

58 Taejoon Park
2. Confirmatory Factor Analysis (CFA)

The primary purpose of this study was to investigate the extent to which the students’ performance on the writing assessment is affected by the underlying traits and the test methods used. To this end, a number of substantively plausible CFA models were tested within the framework of a multitrait–multimethod (MTMM) design by which multiple traits are measured by multiple methods. In the MTMM design employed in this study, the aspects of writing ability being tapped into by the four analytic rating scales were treated as multiple traits, whereas the tasks and raters used to elicit and score test performance were treated as multiple methods.

An advantage of this CFA approach to the analysis of MTMM data is the apparently unambiguous interpretation of method effects, with large method factor loadings indicating the existence of method effects. The results of the CFA modeling discussed in this section are based on the covariation matrix of 16 observed variables (4 analytic rating scales x 2 raters x 2 tasks). Since the observed data met the assumption of multivariate normality, the maximum likelihood parameter estimation method was used in the present study.

1) Baseline Model

The baseline model, by definition, is the foundation against which a series of increasingly stringent hypotheses are tested. Nested models are hierarchically related to one another in the sense that their parameter sets are subsets of one another. In this study, in addition to the chi-square difference test, the competing CFA models
hypothesized in this study were evaluated on the basis of three widely used fit indexes: comparative fit index (CFI), standardized root mean square residual (SRMR), and root mean square error of approximation (RMSEA). For these three fit indexes, the cutoff criteria recommended by Hu and Bentler (1999) were adopted in the current study. That is, .95 or more for CFI, .05 or less for SRMR, and .06 or less for RMSEA was considered to be indicative of good fit.

The MTMM model shown in Figure 2 represents the baseline model against which alternatively nested models were compared in this study.

Figure 2. Baseline Model (correlated traits & correlated methods)

Note.
1. Errors of measurement associated with each observed variable are omitted in this figure due to space limitation
2. Variable names:
   - TFul1R1 = Task fulfillment score of task 1 assigned by rater 1
   - TFul1R2 = Task fulfillment score of task 1 assigned by rater 2
   - TFul2R1 = Task fulfillment score of task 2 assigned by rater 1
   - TFul2R2 = Task fulfillment score of task 2 assigned by rater 2
   - Cont1R1 = Content control score of task 1 assigned by rater 1

60 Taejoon Park
Cont1R2 = Content control score of task 1 assigned by rater 2
Cont2R1 = Content control score of task 2 assigned by rater 1
Cont2R2 = Content control score of task 2 assigned by rater 2
Org1R1 = Organizational control score of task 1 assigned by rater 1
Org1R2 = Organizational control score of task 1 assigned by rater 2
Org2R1 = Organizational control score of task 2 assigned by rater 1
Org2R2 = Organizational control score of task 2 assigned by rater 2
Lang1R1 = Language control score of task 1 assigned by rater 1
Lang1R2 = Language control score of task 1 assigned by rater 2
Lang2R1 = Language control score of task 2 assigned by rater 1
Lang2R2 = Language control score of task 2 assigned by rater 2

Table 5 presents selected model fit statistics for the baseline CFA model. As shown in this table, the baseline model displayed good fit to the observed data, meeting all the criteria for selecting well-fitting models (e.g., CFI ≥ .95; SRMR ≤ .05; RMSEA ≤ .06) recommended by Hu and Bentler (1999).

<table>
<thead>
<tr>
<th>Model</th>
<th>chi-square</th>
<th>df</th>
<th>chi-square to df ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>chi-square = 111.584</td>
<td>df = 64</td>
<td>chi-square to df ratio = 1.744</td>
</tr>
</tbody>
</table>

**Fit indexes**
- comparative fit index (CFI) = .990
- standardized root mean-square residual (SRMR) = .035
- root mean-square error of approximation (RMSEA) = .052

2) Alternative Models

After ascertaining that the baseline model displayed adequate fit to the observed data, the model was compared against a set of substantively plausible alternative models in order to select a model that would best represent the underlying trait and method factor structure of the writing assessment. For example, in order to test the multidimensionality of the construct (i.e., writing ability) being tapped into by the four analytic rating scales, the baseline model was compared with an alternative model positing a single general factor that represents a unidimensional construct of writing ability. As
discussed earlier, each of the alternative CFA models hypothesized in this study can be considered to be a “nested” or more “restricted” version of the baseline model. For instance, the trait factor covariances in the baseline model were constrained to equal 1.0 to obtain an alternative model positing a unitary trait factor. The five alternative nested models evaluated in this study are: (1) second-order traits & correlated methods, (2) a unitary trait & correlated methods, (3) correlated traits & correlated tasks only (no rater factors), (4) no traits: correlated methods only, and (5) no traits: correlated tasks only (no rater factors).

3) Comparison of CFA Models

Each of the alternative models tested in this study yielded a proper solution, meaning that the estimated parameters were all within the range of permissible values (i.e., no negative or zero variance estimates, no correlations greater than 1.0). Table 6 summarizes the fit statistics related to the alternative models, and Table 7 presents the results of the chi-square difference tests conducted between these nested models.

<table>
<thead>
<tr>
<th>Model</th>
<th>chi-square</th>
<th>df</th>
<th>chi-square</th>
<th>df</th>
<th>CFI</th>
<th>SRMR</th>
<th>RMSEA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>111.584</td>
<td>64</td>
<td>1.744</td>
<td>.990</td>
<td>.035</td>
<td>.053</td>
<td></td>
</tr>
<tr>
<td>Model 2</td>
<td>121.319</td>
<td>66</td>
<td>1.838</td>
<td>.988</td>
<td>.035</td>
<td>.055</td>
<td></td>
</tr>
<tr>
<td>Model 3</td>
<td>153.204</td>
<td>70</td>
<td>2.189</td>
<td>.983</td>
<td>.018</td>
<td>.066</td>
<td></td>
</tr>
<tr>
<td>Model 4</td>
<td>296.371</td>
<td>81</td>
<td>3.659</td>
<td>.955</td>
<td>.040</td>
<td>.098</td>
<td></td>
</tr>
<tr>
<td>Model 5</td>
<td>228.802</td>
<td>86</td>
<td>2.660</td>
<td>.970</td>
<td>.031</td>
<td>.077</td>
<td></td>
</tr>
<tr>
<td>Model 6</td>
<td>558.662</td>
<td>103</td>
<td>5.424</td>
<td>.905</td>
<td>.044</td>
<td>.126</td>
<td></td>
</tr>
</tbody>
</table>

Table 6. Summary of Fit Statistics for the Hypothesized Models

Note.

1. Model specifications
   - Baseline: correlated traits: correlated methods
   - Model 2: second-order traits: correlated methods
   - Model 3: a unitary trait: correlated methods
   - Model 4: correlated traits: correlated tasks only (no rater factors)
   - Model 5: no traits: correlated methods
   - Model 6: no traits: correlated tasks only (no rater factors)
2. Criteria for selecting well-fitting models (Hu & Bentler, 1999)
   - Chi-square/df (Chi-square to degrees of freedom ratio): 2 or less
   - CFI (comparative fit index): .95 or above
   - SRMR (standardized root mean-square residual): .05 or less
   - RMSEA (root mean-square error of approximation): .06 or less

| Table 7. Summary of the chi-square tests between hypothesized CFA models |
|-----------------------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| Baseline                   | 111.584         |                 |                 |                 |                 |                 |
| Model 2                    | 9.735*          | 121.319         |                 |                 |                 |                 |
| Model 3                    | 41.62**         | 31.885**        | 153.204         |                 |                 |                 |
| Model 4                    | 184.787**       | 175.052**       | 143.167**       | 296.371         |                 |                 |
| Model 5                    | 117.218**       | 107.483**       | 75.598**        | 67.569**        | 228.802         |                 |
| Model 6                    | 447.078**       | 437.343**       | 405.458**       | 262.291**       | 329.860**       | 558.662         |

Note.
1. * indicates the chi-square difference between the two models is significant at the .01 level.
2. ** indicates the chi-square difference between the two models is significant at the .001 level.
3. Values on the diagonal (in bold) are chi-square values; in parentheses are degrees of freedom.
4. Values in the lower diagonal are chi-square differences; in parentheses are differences in degrees of freedom.

Examination of the fit statistics reported in Table 6 reveals that the baseline model and Model 2 were the only models that met all of the criteria specified for good model fit.

The results of the chi-square difference tests summarized in Table 7 show that the baseline model and Model 2 were significantly better than the other CFA models. The only difference between the two models is that the baseline model specifies four intercorrelated first-order trait factors (i.e., task fulfillment, content control, organizational control, and language control), whereas Model 2 specifies a higher order factor (i.e., writing ability) explaining the intercorrelations among the four first-order factors. Compared to the baseline model and Model 2, the
other models in general displayed relatively poorer fit to the data although they met some of the criteria specified for good model fit. For example, Model 3 positing a single general trait factor (a unidimensional construct) demonstrated poorer fit to the data (e.g., chi-square = 153.204; chi-square/df = 2.189; RMSEA = .066) compared to the first two models positing a multidimensional construct. This provides evidence supporting the notion that the construct being tapped into by the four analytic scales of the writing assessment is multidimensional rather than unidimensional.

When Model 4 in which rater factors were not specified was compared with the baseline model and Model 2 in which they were, the former yielded poorer fit to the data (e.g., chi-square = 296.371; chi-square/df = 3.659; RMSEA = .098) than the latter models, suggesting that although the method effect associated with raters may not be large, it still contributes somewhat to explaining the underlying factor structure of the writing assessment. Finally, Model 5 and 6 in which only method factors were specified displayed much poorer fit to the data than the baseline model and Model 2 in which trait factors as well as method factors were specified. This suggests that the method factors (i.e., tasks and raters) alone cannot sufficiently represent the underlying factor structure of the writing assessment.

As noted earlier, both the baseline model (i.e., the first-order model) and Model 2 (i.e., the hierarchical model) yielded good fit to the data, meeting all the criteria of good fit recommended by Hu and Bentler (1999). The hierarchical model posits that a higher-order factor (writing ability) would explain the intercorrelations among the first-order factors (task fulfillment, content control, organizational control, and language control).

When a chi-square difference test was conducted to compare these two models, the results showed that they differed significantly (chi-square difference = 9.735, difference in degrees of freedom = 2, \( p < .01 \)). This means that the baseline model (i.e., the first-order model) is statistically better than Model 2 (i.e., the hierarchical
model) because the former has a significantly lower chi-square value than the latter. However, the difference between the two models in terms of goodness of fit indexes was extremely small (e.g., CFI = .990 vs. .988; SRMR = .035 vs. .035; RMSEA = .053 vs. .055), suggesting that these two models fit the observed data almost equally well. Therefore, in order to select a final model that would best represent the underlying factor structure of the writing assessment from these two models (i.e., the first-order model vs. the hierarchical model), the first-order trait factor correlations of the baseline model were examined because high intercorrelations among the first-order factors necessitate the concept of a higher-order factor (i.e., writing ability) whereas low intercorrelations among them inhibit the use of a higher-order factor in summarizing the intercorrelations among the first-order factors (Vlachopoulos, Karageorghis, & Terry, 2000). Table 8 presents the first-order trait factor correlations for the baseline model.

<table>
<thead>
<tr>
<th>Factors</th>
<th>Task fulfillment</th>
<th>Content control</th>
<th>Organizational control</th>
<th>Language control</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task fulfillment</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Content control</td>
<td>0.887*</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Organizational control</td>
<td>0.901*</td>
<td>0.927*</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>Language control</td>
<td>0.758*</td>
<td>0.858*</td>
<td>0.906*</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Note. * indicates the correlation is significant at the .005 level.

As shown in Table 8, the correlations among the first-order trait factors derived from the baseline model were high except for the correlation between task fulfillment and language control, suggesting that the concept of a higher-order factor might be necessary to explain the high intercorrelations among the first-order factors. Thus, to further ascertain the plausibility of Model 2 comprising four first-order trait factors all subsumed by a higher-order general factor (i.e., writing ability) as a better alternative to the baseline model, the second-order factor loadings (i.e., factor loadings of each
first-order factor on the higher order factor) and the disturbance terms (i.e., residual variances associated with the first-order factors) of Model 2 were examined. Figure 3 presents the standardized second-order factor loadings and the disturbance terms of Model 2 (the hierarchical trait and correlated method model).

**Figure 3. Standardized Second-order Factor Loadings and Disturbances for Model 2**

Note.
1. D1–D4: disturbances or residual variances associated with the four first-order factors.
2. * indicates that the parameter was freely estimated.
3. Statistically significant parameter estimates (p < .05) are underscored.

In order for a second-order model such as Model 2 to be considered to be an adequate representation of the observed data, the second-order factor loadings should be large and statistically significant, whereas the disturbance or residual variance associated with each first-order factor should be small (Vlachopoulos, Karageorghis, & Terry, 2000). The residual variance associated with each first-order factor indicates the proportion of the variance of the first-order factor unexplained by its higher-order factor after all measurement error has been removed. In this respect, the
disturbance terms have to do with the predictive validity of the second-order model. Thus, if the disturbance terms are large, the plausibility of the second-order structure might be questioned.

As shown in Figure 3, the factor loadings of each first-order factor on the higher-order factor were large, ranging from .886 to .986, whereas the residual variance estimates for each first-order factor were extremely small, ranging from .007 to .095. As noted earlier, these residual variances are the proportion of true score variance in each first-order factor that was “unexplained” by the higher order factor. For example, the residual variance estimate of .015 (D2) associated with the “content control” factor shown in Figure 3 indicates that only 1.5% of the variance in the first-order factor was not explained by the higher-order factor (i.e., writing ability).

In sum, it seems clear that an adequate representation of the underlying factor structure of the writing assessment requires the use of a higher-order factor explaining the intercorrelations among the four first-order trait factors given (1) the high intercorrelations among the first-order trait factors, (2) the large and statistically significant factor loadings of the first-order factors on the higher-order factor, and (3) the extremely small residual variance (i.e., the disturbance term) associated with each first-order factor. Therefore, among the six substantively plausible CFA models evaluated in this study, Model 2 appears to be the best representation of the underlying trait and method factor structure for the writing assessment used in this study.

After Model 2 was selected as the final model from the six competing models based on its substantive and statistical adequacy, factor loadings of the model were examined to investigate the extent to which the students' performance on the writing assessment is affected by the underlying traits and the test methods. Table 9 presents the standardized trait and method factor loadings for the final model (i.e., second-order traits: correlated methods).
### Table 9. Standardized Trait and Method Factor Loadings for the Final Model

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>TFul1R1</td>
<td>.288&lt;sup&gt;1&lt;/sup&gt;</td>
<td>.590</td>
<td>.642</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TFul1R2</td>
<td>.157</td>
<td>.548</td>
<td>.664</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TFul2R1</td>
<td>.621</td>
<td>.258</td>
<td>.834</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TFul2R2</td>
<td>.583</td>
<td>.642</td>
<td>.682</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cont1R1</td>
<td>.324&lt;sup&gt;1&lt;/sup&gt;</td>
<td>.548</td>
<td>.664</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cont1R2</td>
<td>.187</td>
<td>.250</td>
<td>.840</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cont2R1</td>
<td>.615</td>
<td>.138</td>
<td>.700</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cont2R2</td>
<td>.575</td>
<td>.434</td>
<td>.594</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Org1R1</td>
<td>.386&lt;sup&gt;1&lt;/sup&gt;</td>
<td>.488</td>
<td>.608</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Org1R2</td>
<td>.259</td>
<td>.235</td>
<td>.805</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Org2R1</td>
<td>.493</td>
<td>.126</td>
<td>.786</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Org2R2</td>
<td>.500</td>
<td>.454</td>
<td>.614</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lang1R1</td>
<td>.594&lt;sup&gt;1&lt;/sup&gt;</td>
<td>.421</td>
<td>.515</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lang1R2</td>
<td>.455</td>
<td>.214</td>
<td>.719</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lang2R1</td>
<td>.420</td>
<td>.126</td>
<td>.819</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lang2R2</td>
<td>.395</td>
<td>.429</td>
<td>.718</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Second-order</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Writing</td>
<td>.910&lt;sup&gt;1&lt;/sup&gt;</td>
<td>.959</td>
<td>.986</td>
<td>.886</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note.
1. The superscript 1 indicates the factor loading has been fixed to 1.0 for latent factor scaling.
2. Statistically significant factor loadings (p < .05) are underscored.
3. Trait factor names: TFul. = Task fulfillment; Cont. = Content control; Org. = Organizational control; Lang. = Language control
4. Variable names:
   - TFul1R1 = Task fulfillment score of task 1 assigned by rater 1
   - TFul1R2 = Task fulfillment score of task 1 assigned by rater 2
   - TFul2R1 = Task fulfillment score of task 2 assigned by rater 1
   - TFul2R2 = Task fulfillment score of task 2 assigned by rater 2
   - Cont1R1 = Content control score of task 1 assigned by rater 1
   - Cont1R2 = Content control score of task 1 assigned by rater 2
   - Cont2R1 = Content control score of task 2 assigned by rater 1
   - Cont2R2 = Content control score of task 2 assigned by rater 2
   - Org1R1 = Organizational control score of task 1 assigned by rater 1
   - Org1R2 = Organizational control score of task 1 assigned by rater 2
   - Org2R1 = Organizational control score of task 2 assigned by rater 1
   - Org2R2 = Organizational control score of task 2 assigned by rater 2
   - Lang1R1 = Language control score of task 1 assigned by rater 1

68 Taejoon Park
As shown in Table 9, all trait factor loadings are statistically significant except those fixed to 1.0 for the purpose of latent factor scaling. However, a comparison of factor loadings across traits and methods shows that the magnitude of the task factor loadings exceeded that of the trait factor loadings for all but one of the observed variables (i.e., Lang1R1), suggesting the attenuation of traits by strong method effects, especially those related to tasks. This means that students’ performance on the writing assessment is significantly affected by test methods, especially those associated with writing tasks used to elicit test performance.

VI. DISCUSSION

Among the six competing CFA models evaluated in this study, Model 2 consisting of hierarchical traits (i.e., four first-order trait factors all subsumed by a higher-order general factor) and correlated methods (i.e., two correlated rater factors and two correlated task factors) was selected as the final model that provides the best representation of the underlying trait and method factor structure for the current measurement design of the writing assessment based on (1) the good fit to the observed data, as indicated by multiple fit indexes, (2) the high intercorrelations among the four first-order trait factors, which necessitate the concept of a higher-order factor explaining the relationships among the first-order factors, (3) the large and statistically significant loadings of the first-order factors on the higher-order general factor, which indicates that the test takers’ performance on the four first-order factors was affected by their underlying writing ability factor, and (4) the extremely small disturbance terms, which indicate that large proportion of the variance of each first-order factor was explained by the underlying writing
ability factor.

The comparison of the factor loadings of the final model across traits and methods revealed that for all but one of the sixteen observed variables, the magnitude of the task factor loadings exceeded that of the trait factor loadings.

According to Widaman (1985), the underlying assumption taken by the traditional multitrait-multimethod (MTMM) paradigm is that the observed variance attributed to the traits that the test is intended to measure is considered evidence for the test’s validity, whereas the observed variance attributed to the test methods is viewed as evidence against the test’s validity. In other words, the MTMM paradigm typically treats all method effects as error, regardless of their relevance to the construct of interest, predicting that ideally strong covariance should occur only as the result of similar traits underlying test performance. This approach, as Chapelle (1998) notes, is consistent with a trait perspective whereby all method effects refer to error introduced into performance data resulting in inconsistent performance (i.e., variation) across test methods. In the present study, however, the two sets of data elicited by means of two different writing tasks varied because they were samples from two different “contexts.” Therefore, the method effect involving the test tasks could be considered evidence of the expected influence of context on test performance. It is for this reason that the observed variance attributed to the task facet could be viewed as construct-relevant and thus should not be treated as error. This approach is consistent with an interactionalist perspective whereby test takers’ performance is the result of traits, contextual features, and their interaction (Chapelle, 1998).

The large task sampling variability observed in this study can be attributed to the nature of the writing tasks. That is, the two writing tasks used in this study were very different in terms of task type and in the ways in which they were contextualized. Thus, the distinct nature of the two writing tasks may have led to the substantial variation in test takers’ performance across tasks.
The substantial score variability due to tasks found in this study suggests that there is an inevitable tradeoff between reliability and validity in writing performance assessments. That is, narrowing and structuring of writing tasks might be expected to increase score reliability in the same way that writing homogeneous objective test items does. However, doing so inevitably narrows the domain to which results generalize. While more narrowly defined tasks might solve the problem of score variability due to tasks, they may do so at an exacting price. That narrowing, as pointed out by Dunbar, Koretz, and Hoover (1991), poses an unattractive choice in terms of validity. That is, if inferences are kept broad enough to be meaningful, their validity can be undermined, whereas if inferences are narrowed to maintain validity in the face of the restricted definition of tasks, they can become less meaningful.

Another important theoretical implication of this study has to do with how we should go about interpreting method effects, especially those involving tasks, in the context of writing performance assessment. In this study, the two writing tasks used to elicit test performance were very different in the ways that they were contextualized (e.g., informal vs. formal). As a result, the test takers performed differently on the two writing tasks. As discussed in Chapelle (1998), there are two distinct perspectives on how to interpret method effects in validity inquiry. From a trait perspective, which defines constructs in terms of the knowledge and fundamental processes of the test taker, method effects refer to error introduced into performance data resulting in inconsistent performance across test methods. On the other hand, from an interactionalist perspective, which sees performance as the result of traits, contextual features, and their interaction, method effects would not always be considered as error. Rather, method effects could be evidence of the expected influence of “context” on performance. That is to say, from the interactionalist perspective, the test takers’ performance on the writing assessment varied substantially from one task to another because the two sets of data elicited by using the two writing tasks...
were samples from two different “contexts.”

In sum, as pointed out by Chapelle (1998), it seems that language testers need to reconceptualize the “trait as good variance” and “method as systematic error variance” as more general notions, “construct relevant variance” and “construct irrelevant variance,” respectively (p.56). In other words, especially in the context of writing performance assessment, language testers need to reconceptualize what should be considered relevant sources of variance and what should be considered irrelevant sources of variance instead of just assuming that only trait variance is construct relevant and all method variance is construct irrelevant. In the present study, the observed variance attributed to the writing tasks is evidence of the expected influence of “context” on performance (i.e., construct relevant variance) and thus should not be viewed as measurement error. Given the instability of test takers’ performance across tasks observed in the current study and many others in the performance assessment literature (e.g., Brennan & Johnson, 1995; Lane, Liu, Ankenmann, & Stone, 1996; Shavelson, Baxtor, & Gao, 1993), the way forward seems to recognize that, while some contexts activate stable ability features, others produce more variable performance from test takers. Thus, as Bachman (2002) emphasizes, language testers need to take into account both a construct-based and a task-based approach to test design and score interpretation.

REFERENCES


Examining the Structure of a Writing Assessment
Using Confirmatory Factor Analysis


### APPENDIX

**Covariance Matrix for the Observed Variables**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>(1)</td>
<td>2.26</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(2)</td>
<td>1.55</td>
<td>2.16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(3)</td>
<td>1.06</td>
<td>0.96</td>
<td>2.43</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(4)</td>
<td>1.23</td>
<td>1.13</td>
<td>1.94</td>
<td>2.37</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(5)</td>
<td>1.61</td>
<td>1.27</td>
<td>0.95</td>
<td>1.14</td>
<td>1.69</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(6)</td>
<td>1.25</td>
<td>1.55</td>
<td>0.83</td>
<td>0.97</td>
<td>1.20</td>
<td>1.74</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(7)</td>
<td>1.00</td>
<td>0.89</td>
<td>1.88</td>
<td>1.74</td>
<td>0.98</td>
<td>0.82</td>
<td>2.10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(8)</td>
<td>0.99</td>
<td>0.93</td>
<td>1.66</td>
<td>1.37</td>
<td>0.99</td>
<td>0.84</td>
<td>1.68</td>
<td>2.10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(9)</td>
<td>1.54</td>
<td>1.20</td>
<td>0.99</td>
<td>1.10</td>
<td>1.34</td>
<td>1.05</td>
<td>1.03</td>
<td>0.97</td>
<td>1.76</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(10)</td>
<td>1.29</td>
<td>1.47</td>
<td>0.92</td>
<td>1.02</td>
<td>1.20</td>
<td>1.34</td>
<td>0.86</td>
<td>0.91</td>
<td>1.14</td>
<td>1.76</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(11)</td>
<td>1.01</td>
<td>1.00</td>
<td>1.86</td>
<td>1.77</td>
<td>1.00</td>
<td>0.85</td>
<td>1.08</td>
<td>1.62</td>
<td>1.03</td>
<td>0.95</td>
<td>2.17</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(12)</td>
<td>1.10</td>
<td>1.02</td>
<td>1.70</td>
<td>1.87</td>
<td>1.02</td>
<td>0.86</td>
<td>1.58</td>
<td>1.75</td>
<td>0.97</td>
<td>0.97</td>
<td>1.66</td>
<td>2.14</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(13)</td>
<td>1.15</td>
<td>0.88</td>
<td>0.82</td>
<td>0.87</td>
<td>1.05</td>
<td>0.81</td>
<td>0.86</td>
<td>0.81</td>
<td>1.10</td>
<td>0.86</td>
<td>0.83</td>
<td>0.80</td>
<td>1.22</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(14)</td>
<td>1.15</td>
<td>1.21</td>
<td>0.88</td>
<td>0.97</td>
<td>1.05</td>
<td>1.15</td>
<td>0.87</td>
<td>0.96</td>
<td>1.04</td>
<td>1.19</td>
<td>0.88</td>
<td>0.89</td>
<td>0.91</td>
<td>1.46</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(15)</td>
<td>0.87</td>
<td>0.89</td>
<td>1.50</td>
<td>1.38</td>
<td>0.85</td>
<td>0.74</td>
<td>1.48</td>
<td>1.29</td>
<td>0.96</td>
<td>0.85</td>
<td>1.58</td>
<td>1.27</td>
<td>0.73</td>
<td>0.82</td>
<td>1.61</td>
<td></td>
</tr>
<tr>
<td>(16)</td>
<td>0.97</td>
<td>0.94</td>
<td>1.41</td>
<td>1.54</td>
<td>0.93</td>
<td>0.86</td>
<td>1.40</td>
<td>1.48</td>
<td>0.92</td>
<td>0.9</td>
<td>1.42</td>
<td>1.50</td>
<td>0.72</td>
<td>0.85</td>
<td>1.24</td>
<td>1.61</td>
</tr>
</tbody>
</table>

Variable names:

1. TFull1R1 = *Task fulfillment* score of task 1 assigned by rater 1
2. TFull1R2 = *Task fulfillment* score of task 1 assigned by rater 2
3. TFull2R1 = *Task fulfillment* score of task 2 assigned by rater 1
4. TFull2R2 = *Task fulfillment* score of task 2 assigned by rater 2
5. Cont1R1 = *Content control* score of task 1 assigned by rater 1
6. Cont1R2 = *Content control* score of task 1 assigned by rater 2
7. Cont2R1 = *Content control* score of task 2 assigned by rater 1
8. Cont2R2 = *Content control* score of task 2 assigned by rater 2
9. Org1R1 = *Organizational control* score of task 1 assigned by rater 1
10. Org1R2 = *Organizational control* score of task 1 assigned by rater 2
11. Org2R1 = *Organizational control* score of task 2 assigned by rater 1
12. Org2R2 = *Organizational control* score of task 2 assigned by rater 2
13. Lang1R1 = *Language control* score of task 1 assigned by rater 1
14. Lang1R2 = *Language control* score of task 1 assigned by rater 2
15. Lang2R1 = *Language control* score of task 2 assigned by rater 1
16. Lang2R2 = *Language control* score of task 2 assigned by rater 2