Investigating Complex Interaction Effects Among Facet Elements in an ESL Writing Test Consisting of Integrated and Independent Tasks

Yong-Won Lee
(Seoul National University)
Robert Kantor
(Educational Testing Service)

Lee, Yong-Won and Kantor, Robert. (2015). Investigating Complex Interaction Effects Among Facet Elements in an ESL Writing Test Consisting of Integrated and Independent Tasks. *Language Research*, 51.3, 653-678.

The main purposes of the current study are to: (a) examine the interactional effects among test-takers, tasks, and raters, as well as the main effects of these facets, in an ESL writing test consisting of both integrated and independent writing tasks and (b) thereby identify additional sources of score variability and error in the rating of test-taker responses. A total of 162 test-takers with 29 different L1 backgrounds participated in the study, each of whom took the same six writing tasks, which included 3 Listening-Writing (LW), 2 Reading-Writing (RW), and 1 Independent Writing (IW) tasks. Each of the essays was rated by each of the 6 trained raters to obtain a completely-crossed data matrix for the test-takers, tasks, and raters. A computer program, FACETS (Linacre, 1998), was used to calibrate the test-takers, tasks, and raters and conduct interaction analysis on 970 essays. Results of the analyses revealed that raters seemed to be having slight difficulty in maintaining a consistent level of severity across all of the 6 tasks and that a close inspection of the rating patterns of selected test-takers demonstrated the usefulness of interaction analysis in pinpointing particular combinations of facet elements with unusual interaction patterns. The implications of these findings for writing assessment are discussed along with the avenues for further investigation.

Keywords: ESL writing assessment, integrated tasks, independent tasks, main effects, interaction effects, Many-faceted Rasch Measurement, Facets

1. Introduction

In recent years, writing tests comprising integrated tasks as well as independent tasks are widely used in English as second/foreign language (ESL/EFL) testing and assessment (Knoch & Sitajalabhornet, 2013; Plakans, 2015; Sawaki, Quinlan, & Lee, 2013). *Integrated* tasks require test-takers to integrate multiple language skills (or modalities) in a substantial way to complete a constructed writing task at hand (e.g., to understand academic texts and/or lectures and compose written responses that demonstrate understanding of such stimulus material), while *independent* tasks require the test-takers to use their personal experiences or general knowledge rather than stimulus material to respond to a writing prompt (Lee & Kantor, 2007).

Integrated tasks are advocated on two major grounds (Lewkowics, 1977; Lee & Kantor, 2007; Plakans, 2015): (a) test-takers are less likely to suffer from the lack of information on which to base their argument (Read, 1990; Weir, 1993) and (b) test validity would be improved by emulating authentic communication tasks used in academic settings (Wesch, 1987). When the integrated tasks are used together with independent tasks in the same test, it becomes possible for us to represent the writing construct more fully in assessing the test-taker's writing proficiency. However, one should also note that there have been some concerns raised about writing tests comprising both integrated and independent writing tasks, which include the low generalizability of writing scores across task types and the built-in interdependency between modalities or different sections of the test (Lee & Kantor, 2007). These concerns may also be associated with the potential impact of input stimuli on test-takers' elicited, written responses. Since these task types are dissimilar in input stimuli characteristics (e.g., an audio recording, a reading passage, a short stand-alone statement), one can possibly claim that each of these task types assesses a somewhat distinct writing construct, which can provide a justification for reporting separate scores for each of these distinct constructs.

Similar arguments could be put forward about the test-takers' test-taking process and the raters' scoring process (Lee, 2006; Lee & Kantor, 2007). First, the test-takers might use different cognitive skills and processes in

responding to these different types of writing tasks. Moreover, the raters can also apply a somewhat different set of rating criteria for each of the task types. When evaluating test-taker responses for independent tasks, for instance, raters can mostly focus on language. When scoring test-taker responses from integrated tasks, however, raters have to deal with content accuracy as well in order to assess whether the test-takers have grasped what was described or presented in the text or audio-recording. Despite these concerns, it is also possible that there are sets of core rating criteria that are used in common by different raters across task types and thus what are measured by these distinct-looking task types can turn out to be extremely highly correlated empirically.

A number of previous studies using *generalizability* (G) *theory* analyses have examined the effects of these facet elements on the measurement errors and score reliability and have shown that the major source of measurement error is generally related to tasks rather than raters in writing assessments (Brennan, Gao, & Colton, 1995; Lee & Kantor, 2007). However, it has also been pointed out that G-theory analysis is limited in providing specific information about individual facet elements (*e.g.*, each individual test-taker, rater, task) and each combination of facet elements (*e.g.*, each individual rater-by-test-taker pairs, rater-by-task pairs) that can be fed into the test development and refinement process (Bachman, Lynch, & Mason, 1995; Lynch & McNamara, 1998). The focus of investigation in G-theory is on "groups" rather than "individuals." In other words, G-theory is interested in investigating the overall effects of facets (main effect) and combination of facets (interaction effects) rather than the more finer-grained effect of individual facet elements contributing to the overall effect.

An alternative, promising psychometric tool for analyzing rater-mediated assessment is the *many-faceted Rasch measurement* (MFRM; also called FACETS) procedure (Linacre, 1989, 1998) which was developed within the framework of item response theory (IRT), particularly *Rasch measurement*. This MFRM procedure makes it possible for us to not only put test-takers, raters, and tasks in the same frame of reference (on the same logit scale) but also identify unusual combinations of facet elements for further examination (Kondo-Brown, 2002; Kozaki, 2004; Linacre,

1989, 1998; Lynch & McNamara, 1998; McNamara, 1996; Myford & Wolfe, 2000a, 2000b; Tyndall & Kenyon, 1995; Upshur & Turner, 1999; Weigle, 1998; Wiggleworth, 1993). However, one important limitation of MFRM application in writing assessment is that the focus of MFRM analysis has been on estimating and comparing parameters of facets elements that represent the main effects of the facets (*e.g.*, test-taker proficiency, task difficulty, and rater severity) but that little research has been done to examine interaction effects among facet elements. In this regard, the MFRM bias/interaction in particular has been known to be effective in identifying unusual interaction patterns among various facet elements in our rating data and finding areas of quality improvement and refinement for writing tests being developed or existing tests (Kondo-Brown, 2002; Linacre, 1998).

With these as a backdrop, the current study aims to examine not only the main effects of test-takers, tasks, and raters on the writing scores but also the interaction effects among these facet elements in an ESL (English as a second language) writing test consisting of both integrated and independent writing tasks and thereby identify the potential sources of error in the rating of test-taker responses for the writing test.

2. Literature Review

2.1. Potential Sources of Error in Rater-Mediated Writing Assessment

In the context of rater-mediated, performance-based language assessment, tasks and raters have been studied as two major sources of score variability and measurement error (Bachman, Lynch, & Mason, 1995; Brennan, Gao, & Colton, 1995; Eckes, 2005; Engelhard, 1994; Lynch & McNamara, 1998; Lee & Kantor, 2007; Myford & Wolfe, 2000a). Particularly, when new task types drastically different from the traditional types of tasks are used in the test, it is critically important not only to carefully examine the impact of the new task types on the test-taking and rating processes and the overall psychometric qualities of assessment but also to find ways to minimize the construct-irrelevant impact on the scores.

First of all, when a performance-based writing test consists of multiple tasks, particularly of different types, one critical quality control issue is to examine whether there is a high degree of consistency and stability in test-takers' performance (and scores) across different tasks or task types. Especially, since only a small set of tasks can usually be used in large-scale writing tests due to time constraints, it would be even more important to ensure the generalizability of writing scores across different tasks and task types (Lee & Kantor, 2007). If the test-takers' performance fluctuates across different tasks or task types in great magnitude, however, this suggests that there is a significant level of test-taker-by-task interaction effects and that these different types of tasks might not be truly additive in terms of a construct that all of the these tasks are designed to measure as a whole. If this is found out to be true, this warrants a more in-depth analysis of task content and a range of knowledge, skills, and processes engaged by these tasks.

Second, raters and rater judgment-related error need to be need to be closely monitored and investigated in rater-mediated writing assessments (Eckes, 2005; Engelhard, 1994; Myford & Wolfe, 2000a), which are in one way or another related to the appropriateness of the rating rubrics and scales employed for a particular test. Engelhard (1994) identifies four different types of rater error in the content of writing assessment, which include rater severity/leniency difference, the Halo effect, central tendency, and restriction of range. Among these, the rater severity/leniency differences can clearly be linked to the rater main effect in the G-theory framework. It has been found that rater training/retraining can help the raters more internally consistent across tasks and test-takers but do not remove the severity differences among raters completely (Weigle, 1998). In addition, numerous studies have also found significant effects of not only rater-by-test-taker interaction (Kondo-Brown, 2002; Lee & Kantor, 2007; Lynch & McNamara, 1998) but also rater-by-task type interaction (Lee, 2006; Lynch & McNamara, 1998; Wiggleworth, 1993) and rater-by-rating criteria interaction.

One additional issue that is worth discussing here in relation to measurement error in the current study is the increasing need to carefully examine a full range of complex interaction effects among facet elements, partic-

ularly involving raters, in the contexts of rater-mediated writing assessment. When it comes to the issue of task- and rater-related error, research has primarily focused on the main effects of tasks and raters and the two-way interaction effects between test-takers and each of these two facets (e.g., test-taker-by-task, test-taker-by-rater interaction). It should also be noted that there can also be an interaction effect between raters and tasks. More importantly, when we use integrated and independent tasks in combination in the same writing test, it would be interesting to see whether raters tend to exhibit different patterns of interaction with test-takers, depending upon the tasks or task types.

2.2. MFRM Interactional Analyses in Writing Assessment

MFRM (many-faceted Rasch measurement) is an extended version of the Rasch measurement model that enables more than two measurement facets to be investigated simultaneously in the same frame of reference (Linacre, 1989, 1998). In the traditional dichotomous and polytomous Rasch models for standardized tests, "test-takers" and "items/tasks" are usually the two major facets of measurement.[1] MFRM can be expanded to include a "rater" facet for performance-based assessment situations, so that all three of the test-taker, tasks, and rater facets are put into the same frame of reference for analysis and interpretation. In addition, after logit measures for the main facet elements (*e.g.*, test-taker proficiency, task difficulty, and rater severity) are estimated, measures of interaction (or bias) expressed in terms of logit values can also be estimated for all possible pairs (or combinations) of the facet elements in subsequent interaction analyses.

The base form of the partial credit model used in this study can be expressed as:

---

1) It should be noted that measurement facets are conceptualized somewhat differently in MFRM and generalizability (G) theory. In G-theory, there is a clear distinction made between the object of measurement (usually test-takers) and measurement facets, whereas both the object of measurement and measurement facets are all counted as "facets" in MFRM. Thus, rater-mediated assessment involving test-takers, tasks, and raters typically is regarded as a two-facet scenario in G-theory, whereas it is regarded as a three-facet one in MFRM. For the sake of convenience, the MFRM convention is followed in this paper.

$$\log\left(\frac{P_{nijk}}{P_{nijk-1}}\right) = \beta_n - \delta_i - \lambda_j - \tau_{ik},$$

where

P_{nijk} = the probability of test-taker *n* when rated by rater *j* on task *i* receiving a score in category *k*,

P_{nijk-1} = the probability of test-taker *n* when rated by rater *j* on task *i* receiving a score in category *k*,

β_n = the proficiency of test-taker *n*,

δ_i = the difficulty of task *i,*

λ_j = the severity of rater *j*,

τ_{ik} = the difficulty of category *k* relative to category *k-1* on item *i.*

Once the main analysis is done by using the base form of the model presented above, the MFRM bias/interaction analysis can then be conducted on the residuals of the main analysis, with the facet parameters from the main analyses fixed (or anchored) (Linacre, 1998). More specifically, the residuals between raw and expected scores are computed for each combination of facet elements, and the residual scores for each element are converted into logit (log-odds) measures and standardized z-scores. The logit measures permit the effect of interaction to be expressed in the same frame of reference as the element measure, while the z-scores give us a measure of statistical significance for the interaction effect. The expected mean and standard deviation of the z-scores are 0 and 1, respectively, and absolute z-scores greater than 2 are usually seen as an indication of significant interaction effect (Linacre, 1998). The bias/interaction analysis identifies unusual interaction patterns among various facet elements in our data matrix, especially patterns that suggest a consistent deviation from what is expected by a specified MFRM model, given the total response matrix. Patterns of interaction are identified in relation to a pair or a combination of facets (*e.g.*, rater-by-test-taker, rater-by-task, task-by-test-taker, and rater-by-task-by-test-taker).

In a sense, the expanded, partial credit model of MFRM (with these two- and three-way interaction terms) used discussed above is comparable to a two-facet, fully-crossed design with tasks (t) and raters (r) as random

facets (p×t×r) in the G-theory framework. Nevertheless, it should be pointed out that G-theory analysis focuses on estimating and comparing variances associated with the main and interaction effect at group levels, whereas MFRM analysis is centered on estimating parameters for each individual facet element (main) or combination of facet elements (interaction).

2.3. Research Questions

1. Are raters internally consistent? Are raters consistent in rating essays across test-takers and tasks?
2. Are raters interchangeable? Are raters homogeneous in terms of overall rater severity/leniency?
3. Do some individual raters score particular test-takers more harshly or more leniently than other test-takers?
4. Do individual raters score particular tasks or task types more harshly or more leniently?
5. Do individual raters score particular test-takers on particular tasks (i.e., particular test taker-by-task combinations) more harshly or more leniently?

3. Method

3.1. Participants

The data set analyzed in the present study is a subset of the larger data set obtained from a prototyping study in which the main focus of the study was to evaluate a range of possible task types that can be used in the final configuration of the test within the framework of G-theory (Enright *et al.*, 2008; Lee & Kantor, 2007). In the current study, MFRM bias/interaction analysis is the focus of investigation. Participants were 162 *English as a Second (or Foreign) Language* (ESL/EFL) students recruited from 3 domestic and 5 overseas (*i.e.*, Mexico, Australia, Canada, Hong Kong, Taiwan) testing sites. Participants completed a battery of English

assessments containing a prototype version of writing tasks that are examined in this study (see Enright, Bridgeman, Eignor, Lee, & Powers, 2008, for more details). Of the 162 test-takers used in this study, there were 83 males (51.2%), 71 females (43.8%), and 8 test-takers of unidentified gender. The average age of the test-takers was 21.9. Paper-based ITP (Institutional Testing Program) TOEFL scores of the test-takers ranged from 403 to 673, with a mean of 559 and a standard deviation of 61. The participants were from 29 language backgrounds, with the 5 largest native language groups being Chinese (22.2%), Spanish (18.5%), Cantonese (16.1%), Korean (9.3%), and Japanese (6.2%). Among the 162 test-takers, 134 (82.7%) test-takers word processed their essays on a computer, and 28 (17.3%) test-takers hand-wrote their essays.

3.2. Instrument and Procedure

Three major types of writing tasks have been considered for the writing section of the English language assessment, which included integrated (*i.e.*, listening-writing, reading-writing) and independent writing tasks (*e.g.*, a stand-alone prompt). Two separate scoring rubrics were used for integrated and independent writing tasks, because raters had to evaluate content accuracy in test-taker responses for integrated writing tasks. For the current study, a total of 6 writing tasks of 3 task types were prepared and administered as part of a larger prototyping study for an ESL proficiency test, as described in Enright *et al.* (2008).

To be more specific, these writing tasks included 3 Listening-Writing (LW), 2 Reading-Writing (RW), and 1 Independent Writing (IW) tasks administered to 162 ESL/EFL test-takers. The stimuli for the three LW tasks (Tasks 1 to 3) were lectures on "plate tectonics," "biological symmetry," and "communications," respectively. The stimulus materials for the two RW tasks (Tasks 4 and 5) were reading passages on "zooplankton" and "dance", respectively. The IW task (Task 6) was a typical TOEFL CBT writing prompt requiring writers to express their personal experiences and opinion.

A total of six trained raters participated in the rating of the test-taker responses to the tasks. These raters were the ones who had participated

in the operational rating of the TOEFL CBT essays for many years before participating in the current study. In this study, all of the 162 test-takers took the same 6 writing tasks, and each test-taker's essay for each of the 6 tasks was rated by 6 raters on a scale of 1-5 according to a fully-crossed design (p×t×r). To minimize the potential "halo" effect, raters were asked to rate all essays for a specific task for all test-takers and move on to the next task.

3.3. Data analysis

Data analyses were done in two phases. In Phase 1, the computer program FACETS (Linacre, 1998) was used to calibrate test-takers, raters, and tasks to put them in the same frame of reference in Phase 1. The partial credit model implemented by the FACETS program was used for calibration in this study. In the partial credit model, the structure of scale is allowed to vary across tasks. The maximum number of iterations was set at an unlimited number, but the estimation process converged before 100 iterations. Both rater and task facets were centered so that means of logit measures for the two facets were constrained to be zero, but the test-taker was non-centered. Spreads of logit measures for these 3 facet elements were examined and compared by looking at a map produced by the program. In particular, rater severity/leniency measures were examined carefully along with rater fit measures. In Phase 2, FACETS was again used to conduct interaction analyses for: (1) rater-by-test-taker pairs, (2) rater-by-task pairs, (3) task-by-test-taker pairs, and (4) rater-by-task-by-test-taker combinations. Flagged pairs (or combinations) of facet elements were identified and examined.

4. Results

4.1. Calibration of Test-takers, Raters, and Tasks

Good model-data fit was achieved overall.[2)] Two hundred and forty six (about 4.2%) of the total 5,820 responses (or ratings) had absolute

standardized residual values equal or greater than 2, whereas only 27 responses were flagged as unexpected responses that had absolute standardized residuals equal or greater than 3 (Linacre,1998). Figure 1 show the plot of test-taker ability, rater severity, and task difficulty logit measures along with threshold (or step) difficulty parameters for each of the six tasks, which were estimated from the Phase 1 analysis.

Test-takers. Ability logit values of the 162 test-takers ranged from -4.72 to 6.19 (a spread of about 11 logits), with a mean of -0.22 and a standard deviation of 1.74. The spread of test-taker proficiency logit measures was 11 times larger than that of rater severity measures and 5.5 times larger than that of item difficulty logit measures, as seen in Figure 1. The test-taker separation index (G) and separation reliability were 6.63 and 0.98, respectively. This indicates that test-takers can be separated into distinct strata of proficiency. When the (4G+1)/3 formula was used to compute the number of statistically significant strata of proficiency among test-takers (see Myford & Wolfe, 2000a, for more detail about the formula), the sample of test-takers that took the current test form could be separated into 9 distinct levels of writing proficiency. The unstandardized "infit mean square" (information-weighted, inlier sensitive, mean-square fit) statistics ranged from 0.3 to 2.9, whereas the standardized infit z-scores ranged from -4 to 5. Nevertheless, the number of test-takers with unusual infit z scores (i.e., outside +/-2) was rather small. Of the 162 test-takers, only 10 test-takers (6%) turned out to have an absolute infit z-score greater than 2. Five of them had negative z-score values (overfitting), whereas the rest of them had positive z-score values (misfitting).

Raters. The rater severity logit values of the 6 raters ranged from -0.46 to 0.42 (a spread of about 1 logit), with a mean of 0 and a standard deviation of 0.28. The most lenient rater turned out to be Rater 4 (-0.46), whereas the most stringent rater was Rater 5 (0.42). Rater separation index and reliability were 6.22 and 0.97, respectively. This indicates that raters cannot be thought of as equally severe, and thus raters were not strictly interchangeable. However, given all the severity measures clustered

2) According to Linacre (1998:48), when the data fit the model, about 5% of standardized residuals (z-scores) are outside of +/-2, and about 1% are outside +/-3.

within the 1-logit distance on the severity/leniency scale, the severity differences among raters are rather moderate. The "infit" mean square (MnSq) statistics ranged from 0.8 to 1.1, with a mean of 1 and a standard deviation of 0.1, while the "infit" standardized z-scores ranged from -4 to 1.The rater infit statistics from MFRM provides us with a measure of how internally consistent each rater is in assigning scores across tasks and test-takers.

```
-----------------------------------------------------------------------
|Measr| +test-taker  | -Rater  | -Task|S.1  |S.2  |S.3  |S.4  |S.5  |S.6  |
-----------------------------------------------------------------------
+   7 +    More    +         +      +(5)  +(5)  +(5)  +(5)  +(5)  +(5)  +
|     | Proficient | Severe  | hard |     |     |     |     |     |     |
|     |            |         |      |     |     |     |     |     |     |
|     | .          |         |      |     |     |     |     |     |     |
+   6 +            +         +      +     +     +     +     +     +     +
|     |            |         |      |     |     |     |     |     |     |
|     |            |         |      |     |     |     |     |     |     |
|     |            |         |      |     |     |     |     |     |     |
+   5 + .          +         +      +     +     +     +     +     +     +
|     |            |         |      |     |     |     |     |     |     |
|     |            |         |      |     |     |     |     |     |     |
|     |            |         |      |     |     |     |     |     |     |
+   4 +            +         +      +     +     +     +     +     +     +
|     | .          |         |      |     |     |     |     |     |     |
|     |            |         |      |     |     |     |     |     | --- |
|     | .          |         |      |     |     |     |     |     |     |
+   3 + *          +         +      +     +     + --- +     +     +     +
|     |            |         |      |     |     |     | --- | --- |     |
|     | *          |         |      | --- | --- |     |     |     |     |
|     | *.         |         |      |     |     |     |     |     | 4   |
+   2 + *          +         +      +     +     +     +     +     +     +
|     | *****      |         |      |     |     | 4   | 4   | 4   |     |
|     | ***.       |         |      | 4   | 4   |     |     |     |     |
|     | *          |         |      |     |     |     |     |     | --- |
+   1 + **         +         + 5    +     +     + --- +     +     +     +
|     | *******    |         |      | --- | --- |     | --- | --- |     |
|     | ***        | 5       | 4    |     |     |     |     |     |     |
|     | *****.     | 6       | 1 2  |     |     |     |     |     | 3   |
*   0 * ******.    * 1 2     *      * 3   * 3   * 3   * 3   * 3   *     *
|     | ******.    | 3       |      |     |     |     |     |     |     |
|     | **         | 4       |      |     |     |     |     |     |     |
|     | ***.       |         | 6    | --- | --- |     | --- | --- | --- |
+  -1 + *****      +         + 3    +     +     + --- +     +     +     +
|     | ****       |         |      |     |     |     |     |     |     |
|     | ****       |         |      | 2   | 2   |     |     |     |     |
|     | ***.       |         |      |     |     | 2   | 2   | 2   |     |
+  -2 + ***.       +         +      +     +     +     +     +     +     +
|     | *.         |         |      |     |     |     |     |     | 2   |
|     | **         |         |      | --- | --- |     |     |     |     |
|     | .          |         |      |     |     |     | --- | --- |     |
+  -3 + *.         +         +      +     +     + --- +     +     +     +
|     |            |         |      |     |     |     |     |     |     |
|     | *          |         |      |     |     |     |     |     |     |
|     | .          |         |      |     |     |     |     |     |     |
+  -4 + .          +         +      +     +     +     +     +     + --- +
|     |            |         |      |     |     |     |     |     |     |
|     |            |         |      |     |     |     |     |     |     |
|     | *.         |         |      |     |     |     |     |     |     |
+  -5 +    Less    + Lenient + Easy + (1)  + (1)  + (1)  + (1)  + (1)  + (1)  +
|     | Proficient |         |      |     |     |     |     |     |     |
-----------------------------------------------------------------------
|Measr| * = 2      | -Rater  | -Task|S.1  |S.2  |S.3  |S.4  |S.5  |S.6  |
-----------------------------------------------------------------------
```

Figure 1. Calibrations of Test-takers, Raters, and Tasks.

Among 6 raters, only Rater 3 turned out to have a somewhat unusual infit z-score value (*i.e.*, -4.0). On close inspection, however, it was found that although this particular rater had an absolute infit z-score value greater than 2, an infit mean square value (0.8) was within the expected (or tolerable) range of 0.7 to 1.3 (Myford & Wolfe, 2000a). Generally speaking, the negative sign of the z-score value (or the mean square value less than 1.0) for a rater means that there was somewhat less variation in the rater's ratings than expected by the MFRM model (or lack of independence among ratings). Such a rater is usually known to have a "flat-line" profile consisting of identical or very similar ratings across tasks or employ a "play-it-safe" strategy in rating test-taker responses. Such rating behaviors can also be linked to the "central tendency" and "restriction of range" errors mentioned by Engelhard (1994). For this reason, the negative infit z-score value (or an infit mean square value less than 1) is usually regarded as less problematic than the positive z-score values (or an infit mean square value greater than 1) (Myford & Wolfe, 2000a). Given that the infit mean square value for all 6 raters were greater than 0.7, none of the raters seem to have exhibited an extremely inconsistent rating behavior overall.

Tasks. The task difficulty logit values of the 6 tasks ranged from -0.99 to 0.88 (a spread of about 2 logits), with a mean of 0.00 and a standard deviation of 0.67. Task separation index and separation reliability were 14.94 and 1.00, respectively. This suggests that there are statistically significant differences in difficulty among 6 tasks. The most difficult task was Task 5 (a RW task based on a reading passage about "dance"), whereas the easiest task was Task 3 (an LW task based on a lecture about "communications"). It should be noted that the task difficulty in a polytomous item is an average of estimated *threshold difficulties* (or difficulty of transition from a lower category to higher one) parameters. Because there were 5 possible score points (*i.e.*, 1, 2, 3, 4, 5), a total of 4 threshold difficulty parameters had to be estimated for each task. In relation to the structure of intervals between threshold difficulties for tasks, the last six columns to the right in Figure 1 show that Task 6 (an Independent Writing task) turned out to cover the widest range of the test-taker ability scale, while Tasks 1 and 2 (LW tasks on "tectonic plates" and

"zooplankton") were found out to span the narrowest range of the abilit scale. This means that the independent task used in this study have the higher discriminating power for the test-takers than integrated writing tasks.

The "infit" mean square statistics ranged from 0.8 to 1.1, whereas the "infit" standardized z-scores ranged from -4 to 2. The task infit statistics from MFRM gives us a measures of how consistent each of the tasks is in rank-ordering test-takers. Among the 6 tasks, only Task 1 (LW task based on "tectonics") turned out to have an absolute z-score greater than 2. This task was actually an overfit rather than a misfit, because the sign of the infit z-score value for this task was negative (*i.e.*, -4), with its mean square value less than 1. However, the infit mean square value of 0.8 for this task was located within the tight quality control range of 0.7 to 1.3, which was employed by Myford and Wolfe (2000a). Given that the infit mean square value for all 6 tasks were greater than 0.7, none of the tasks seem to be functioning in a redundant manner. By the same token, since none of the infit mean square indices was greater than 1.3, there was no clear evidence of "psychometric multidimensionality" (Henning, 1992; McNamara, 1996; Myford & Wolfe, 2000a).

4.2. MFRM Interaction Analysis

Rater-by-Test-taker Interactions. This "rater-by-test-taker" interaction analysis provides a tool for investigating whether a particular rater is behaving in a similar way for all test-takers or whether a certain rater is giving overgenerous or too harsh scores for a certain test-taker (Lynch & McNamara, 1998). There were a total of 972 interaction pairs (*i.e.*, 6 raters x 162 test-takers) for raters and test-takers. The interaction logit measure for the "rater-by-test-taker" pairs ranged from -1.64 to 1.38, with a mean of 0.01 and a standard deviation of 0.48, while their standardized z-score values ranged from -2.91 to 2.62, with a mean of 0.00 and a standard deviation of 0.78.

Of the 972 pairs, only 18 pairs (2%) turned out to have absolute z-scores greater than 2 (12 negative and 6 positive values). A negative standardized

z-score (or logit measure) value for a particular rater-by-test-taker pair means that the observed score was higher than the expected score and thus the rater is more lenient for the particular test-taker than she usually is on average for other test-takers; but a positive z-score value indicates the other way around. It was found out that Rater 2 had the largest number of rater-by-test-taker pairs with significant interaction effect (11 pairs), with their z-score values ranging from -2.47 to 2.62. About half of the flagged pairs for Rater 2 had negative z-score values and the other half, positive values. On the other hand, Raters 1, 4, and 5 had much smaller numbers of pairs with significant interaction effects (4, 2, and 1, respectively), but the z-score values for all of the flagged pairs for the 3 raters were negative. This suggests that the rater-by-test-taker pairs for these 3 raters were flagged due to their leniency for these test-takers. Nonetheless, there were no pairs with significant interaction effect flagged for Raters 3 and 6. It means that both raters seemed to have exercised a similar level of severity consistently for all the test-takers.

Task-by-Test-taker Interactions. This "task-by-test-taker" interaction analysis enables us to answer the questions of whether a particular task behaves in a similar way for all test-takers or whether a certain task is significantly more difficult or easier for a certain test-taker. There were a total of 970 interaction pairs for tasks and test-takers[3]. The interaction logit measure for the "task-by-test-taker" pairs ranged from -3.61 to 5.53, with a mean of 0.06 and a standard deviation of 1.09, while their standardized z-score values ranged from -4.97 to 4.77, with a mean of -0.03 and a standard deviation of 1.53. Spreads of both the interaction logit measures and standardized z-scores for the task-by-test-taker interaction were much larger than those for the rater-by-test-taker interaction.

Compared to the "rater-by-test-taker" pairs, a much larger percentage of the "task-by-test-taker" pairs were also flagged for significant interaction effect, and these flagged pairs were distributed rather evenly among the

3) Theoretically, there should be 972 "task-by-test-taker" pairs (6 tasks × 162 test-takers). However, one test-taker's essay for Task 2 and another test-taker's essay for Task 3 were missing at the time of rating. That is why the valid number of the task-by-test-taker pairs analyzed was 970. For the same reason, the total number of observations (or the rater-by-task-by-test-taker combinations) was 5,820, not 5,832.

6 tasks. Of the 970 pairs, 178 of the pairs (18%) turned out to have absolute z-scores greater than 2 (93 negative and 85 positive values). It was found out that the number of the flagged task-by-test-taker pairs were 30, 31, 23, 24, 35, and 35, respectively, for Tasks 1, 2, 3, 4, 5, and 6. In each task, about a half of the flagged pairs had negative standardized z-scores, and other half, positive. Such a trend may be consistent with the results of the G-theory analysis on the same data (Lee & Kantor, 2007). It was found in the G-theory analysis that the test-taker-by-task variance was much larger than the test-taker-by-rater variance.

Rater-by-Task Interactions. This "rater-by-task" interaction analysis helps us to investigate whether a particular rater behaves in a similar way for all tasks or whether she treats a particular task differently from other tasks. There were a total of 36 pairs of rater-by-task combinations. The standardized z-score values for the "rater-by-task" pairs ranged from -5.33 to 5.11, with a mean of 0.00 and a standard deviation of 2.32. Of the 36 pairs, about 33% of rater-by-task pairs (7 negative and 5 positive) turned out to have absolute z-scores greater than 2. It means that raters are not as consistent in severity as desired, when they move from one task (or task type) to another.

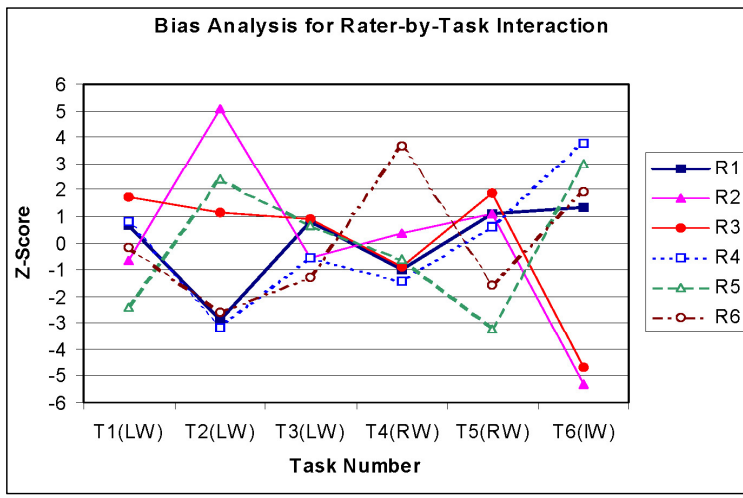

Figure 2. Plot of Standardized Z-Scores for Rater-by-Task Pairs (6 × 6).

Figure 2 shows that each of the raters seemed to have different rater-by-task interaction patterns across the 6 tasks. Rater 1, for instance, exhibited relatively a high degree of consistency across all of the 6 tasks, with the smallest variation in severity across the tasks, whereas Rater 2 has shown the largest variation across the 6 tasks overall. In particular, Rater 2 was extremely severe on Task 2 (with a z-score of 5.10), but extremely lenient on Task 6 (with a z-score of -5.33). The general rater-by-task interaction pattern for the rest of the raters was that the raters were consistent on most of the tasks, but showed some significant interaction effect on one task or two. Rater 3, for example, has exercised an acceptable range of severity on the first 5 tasks, but was extremely lenient on Task 6.

When patterns were examined by task in Figure 2, all of the 6 raters seem to be behaving in a most consistent manner on Task 3 (an LW task based on a lecture about "communication"). It should be reminded that this task was the easiest task among the 6 tasks in terms of MFRM difficulty logit (-0.99) and task mean score (3.4). Task 1 (an LW task based on a lecture about "tectonics") was another task exhibiting a lower but similar level of consistency across the 6 raters. As previously mentioned, this particular task was of medium difficulty, but the most discriminating one among the 6 tasks. In stark contrast, however, raters were polarized in terms of rater severity on Tasks 6 (a IW task) and 2 (a LW task based on a lecture about "symmetry"). On Task 6, for instance, Raters 2 and 3 were clustered together on the extremely lenient side of a severity/leniency z-score scale, whereas the remaining 4 raters were somewhat spread toward the more severe side of the scale. Similarly, Raters 2, 3, and 5 were on the severe side of the severity scale on Task 2, but Raters 1, 4, and 6 were narrowly clustered on the lenient side of the scale (around the z-score of -2.5).

Rater-by-Task-by-Test-taker Interactions. This "rater-by-task-by-test-taker" interaction analysis makes it possible to examine whether a particular "rater-by-task" pair behaves in a similar way for all test-takers or whether a particular test-taker receives overgenerous or too harsh treatment from a certain rater-by-task pair. In total, there were 5,820 combinations of rater, tasks, and test-takers. The z-score values for the "rater-by-task-by-test-taker"

combinations ranged from -4.00 to 3.68, with a mean of -0.03 and a standard deviation of 0.82. Of the 5,820 combinations, 93 combinations (1.6%) turned out to have absolute z-scores greater than 2 (43 negative and 50 positive values). The number of flagged combinations varied across raters. Raters 1, 2, and 4 had a much larger number of flagged combinations (*i.e.*, 24, 27, 18) than Raters 3, 5, and 6 (*i.e.*, 10, 8, 6). When the number of flagged combinations was examined task by task, Task 6 (an IW task) had the largest number of flagged combinations (24); Tasks 2, 3, 4, and 5 had somewhat smaller numbers (15, 15, 17, 14); Task 1 had the smallest number (8).

Table 1. Rating Patterns for Flagged Rater-by-Task-by-Test-taker Combinations

(a) Rater 1 (Severity = 0.05); Task 4 (Difficulty = 0.45);
Test-taker A (Proficiency = -3.33)

Rater ID	Task 1	Task 2	Task 3	Task 4	Task 5	Task 6
1	1	1	2	**4**	2	2
2	1	1	1	3	2	3
3	1	1	1	1	1	1
4	1	2	1	1	1	1
5	1	1	1	1	1	2
6	1	1	1	1	1	1

(b) Rater 2 (Severity = -0.06); Task 4 (Difficulty = 0.45);
Test-taker B (Proficiency = 5.08);

Rater ID	Task 1	Task 2	Task 3	Task 4	Task 5	Task 6
1	5	5	5	5	5	5
2	5	5	5	**3**	5	5
3	5	5	5	5	4	5
4	5	5	5	5	5	5
5	5	5	5	5	5	5
6	5	5	5	5	5	5

(c) Rater 4 (Severity = -0.46); Task 1 (Difficulty = 0.12); Test-taker C (Proficiency = 2.61)

Rater ID	Task 1	Task 2	Task 3	Task 4	Task 5	Task 6
1	4	4	5	5	5	5
2	4	3	5	4	4	4
3	3	5	5	5	4	5
4	**2**	5	5	5	5	5
5	3	4	5	5	5	5
6	3	4	5	4	4	5

A general pattern across 6 tasks was that most of the combinations were within the acceptable z-score range of -2 to 2 for the 6 raters. However, there were some outstanding outliers on one or more tasks for most of the raters, except Rater 3. Three of these outliers with extreme z-scores were selected for further investigation of rating patterns. Tables 1-a, 1-b, and 1-c show the rating patterns for the 3 selected test-takers with a "rater-by-task-by-test-taker combination" of extreme z-score values (-4, 3.68, 3.45). In the first example (Table 1-a), Test-taker A has received a very generous score of 4 (expected score=1.2) on Task 4 from Rater 1, given the total score pattern of the test-taker across tasks and raters. In the second example (Table 1-b), however, Test-taker B has received a much lower score than expected again on Task 4 from Rater 2. The expected score was 4.9 for this combination, but the observed score was 3. In the third case (Table 1-c), Test-taker C has received a lower score than expected on Task 1 from Rater 4. The observed score was 2, even though the expected score was 4.6. This particular rater gave ratings of 5's to the same test-taker for all of the remaining 5 tasks. The rest of 5 raters assigned scores of either 3's or 4's for the same task. Given that Rater 4 was the most lenient rater (-0.46) among the 6 raters overall, such a comparatively low score for Task 1 seems clearly very unusual for this rater.

5. Discussion

The main purpose of the present study was to investigate patterns of interaction among test-takers, raters, and tasks on a writing assessment through MFRM bias/interaction analyses and identify potential sources of rater error in scoring test-taker responses for integrated and independent tasks. It was found that (a) the current 6-task writing assessment distinguishes among test-takers on the writing construct measured by the tasks as a whole; (b) that the major source of construct-irrelevant variation in writing scores is related to tasks rather than raters; and (c) that although overall rater severity/leniency differences were not large, some of the raters seemed to be having difficulty in maintaining a consistent level of severity across all of the 6 tasks.

First, we found that the spread of test-takers proficiency measures was much wider than those of rater severity measures (11 times greater) and task difficulty measures (5.5 times greater). One important thing to mention about raters was that the raters did not seem to be interchangeable in a strict sense. Nevertheless, the severity differences among the raters were rather moderate. On the other hand, the spread of task difficulty measures was twice as wide as that of rater severity measures. Unlike in rater severity measures, differences in difficulty among tasks in the test might be desirable in the sense that the test can be discriminating for test-takers of wider range of proficiency levels. Moreover, none of the tasks seems to be functioning in a redundant manner, and there was no clear evidence of "psychometric multidimensionality" of the tasks (Henning, 1992; McNamara, 1996; Myford & Wolfe, 2000a).

Second, the results of the "rater-by-test-taker" and "task-by-test-taker" interaction analyses have shown that tasks might be a larger source of measurement error (psychometric noise) than raters in this study. Only 2% of the rater-by-test-taker pairs were flagged due to standardized z-scores greater than 2 for interaction effects, and more than half of these flagged pairs were related to a particular rater (Rater 2). In contrast, as much as 18% of the task-by-test-taker pairs were flagged due to unusually large z-scores. Moreover, these flagged pairs were also distributed rather evenly

among the 6 tasks rather than isolated for a particular task. Such a trend is consistent with the results of the previous G-theory analysis on the same data (Lee & Kantor, 2005). Lee and Kantor (2007) have reported based on the previous G-theory analysis on the same data that a much larger portion of the total score variance was explained by the "test-taker-by-task" interaction variance (17.9%) than the "test-taker-by-rater" interaction variance (1.2%). In most of the previous studies on large-scale performance-based assessments, test-taker-by-rater interaction variances were found to be relatively very small, compared to the test-taker-by-task interaction variances.

Third, the plot of "rater-by-task" interaction measures have also made it possible to examine patterns of rater-by-task interactions across different raters and tasks. It seemed that some of the raters were having difficulty in maintaining a consistent level of severity across all of the 6 tasks. Especially, one rater (Rater 2) was extremely severe on one Listening-Writing task (Task 2) but lenient on an Independent Writing task (Task 6). It should also be remembered that this particular rater had the largest number of rater-test-taker pairs with significant interaction effects among the 6 raters. Even though Rater 1 was most consistent in exhibiting a similar level of severity across most tasks, even this rater was somewhat more lenient than expected on Task 2. The remaining 4 raters seemed to be having a similar problem on one or more tasks.

Such rater inconsistency identified across tasks demonstrates the power of bias/interaction analysis in identifying particular combinations of raters and tasks with unusual interaction patterns, but at the same time raises a series of issues related to rating design and rater training for operational testing situations. It would be necessary eventually to investigate the potential causes of such inconsistencies in depth, which include examining raters' familiarity with new scoring rubrics and the structure of tasks and rubrics. In fact, different kinds of rubrics were used for integrated (*i.e.*, LW, RW) and independent writing tasks in this study. Even for the two integrated task types of Listening-Writing and Reading Writing, two different modes of stimulus material were used to provide information to write on: audio-taped lectures for LW tasks and reading passages for RW tasks.

Thus, the scoring rubrics for LW and RW tasks may be the same in form, but different in substance, because raters may need to understand the content of stimulus material in two different modes to score test-taker responses.

Finally, a close inspection of the total rating patterns for 3 selected test-takers with extreme 3-way interactions has revealed that interaction analysis can really pinpoint particular test-takers receiving overgenerous or too harsh treatment from a particular "rater-by-task" pair. Interestingly, both of the first two examples of extreme 3-way interaction occurred on a reading-writing task (Task 4; based on a passage about "zooplankton"), but the direction of bias was opposite in the two cases. One of the potential factors for the first two examples could be related to difficulty in scoring essays made up of sentences or phrases copied verbatim from a reading passage. Raters were instructed to assign the lowest possible score (*i.e.*, 1) to an essay containing copied material from the stimulus material, even though the resulting essays appears to be long enough and well-organized. However, it would be difficult in some circumstances to determine a clear borderline between "copying" and "paraphrasing." It can also be unavoidable for test-takers to borrow some key terms and vocabulary directly from the text in their essay. Possibly, some careful thought in rater training and rubric refinement for operational tests should be given to the issue, then, of how much is an acceptable or unacceptable range of overlap between the reading passage and the test-taker essay. Test-takers may need to be made more aware of the penalty associated with plagiarizing and encouraged to paraphrase the borrowed phrases and sentences in a proper way or create their responses in their own language.

6. Conclusions

MFRM interaction analyses have proven to be a powerful analysis tool in this study. This analysis has not only cross-validated the previous findings from the G-theory analysis but also provided additional diagnostics

on exactly which combination (or pairs) of facet elements may require further examination. A close inspection of rating patterns for selected test-takers has demonstrated the usefulness of interaction analysis in pinpointing particular combinations of facet elements with unusual interaction patterns. Overall, it turned out that a combination of both integrated and independent writing tasks used in this study can work well to discriminate the test-takers on the intended writing construct. Nevertheless, the study suggests that some additional efforts need to be made to prevent the test-takers from using the copy-and-paste strategy to respond to Reading-Writing tasks and refine the rating rubric and scoring guidelines to help the raters better distinguish copied and paraphrased portions in test-takers' essays. More careful rater training on these issues may also help raters more internally consistent in rating test-takers across different types of tasks. Continued future investigation on the potential causes of inconsistency in rater severity across integrated and independent tasks will provide valuable information about the designing and refinement of the new writing tasks and rater training procedures.

Acknoweldgements

This paper is partly based on research work done while the first author was working at Educational Testing Service (ETS). We want to thank the TOEFL Program and ETS for allowing us to publish the work and Don Powers and three anonymous reviewers for their review and helpful feedback for the initial manuscript of the paper. Needless to say, the responsibility for any errors that remain is solely ours, and the ideas and opinions expressed in the paper are those of the authors, not necessarily of ETS, TOEFL Program or Seoul National University (SNU).

References

Bachman, Lyle F., Brian K. Lynch, and Maureen Mason. (1995). Investigating variability in tasks and rater judgments in a performance test of foreign lan-

guage speaking. *Language Testing* 12.2, 238-257.

Brennan, Robert L., Xiaohong Gao, and Dean A. Colton. (1995). Generalizability analyses of Work Key listening and writing tests. *Educational and Psychological Measurement* 55, 157-176.

Cumming, Alister, Robert Kantor, Don Powers, and Terry Santos. (2000). TOEFL 2000 writing framework: A working paper (TOEFL monograph MS -19). Princeton, NJ: ETS.

Eckes, Thomas. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly* 2.3, 197-221.

Engelhard, George Jr. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement* 31, 93-112.

Enright, Mary K, Brent Bridgeman, Dan Eignor, Yong-Won Lee, and Don Powers. (2008). Designing new measures of listening, reading, speaking, and writing. In Chapelle, Carol, Mary K. Enright, and Joan Jamieson, (Eds.), *Building a validity argument for the Test of English as a Foreign Language*. Mahwah, NJ: Erlbaum.

Henning, Grant. (1992). Dimensionality and construct validity of language tests. *Language Testing* 9.1, 1-11.

Knoch, Ute and Woranon Sitajalabhorn. (2013). A closer look at integrated writing tasks: Towards a more focussed definition for assessment purposes. *Assessing Writing* 18, 300-308.

Kozaki, Yoko. (2004). Using GENOVA and FACETS to set multiple standards on performance assessment for certification in medical translation from Japanese into English. *Language Testing* 21.1, 1-27.

Lewkowicz, Jo A. (1997). The integrated testing of a second language. In C. Clapham & D. Corson (Eds.), *Encyclopedia of language and education volume 7 language testing and assessment* (pp. 121-130). Dordrecht, Netherlands: Kluwer Academic Publishers.

Lee, Yong-Won. (2006). Dependability of scores for a new ESL speaking assessment consisting of integrated and independent tasks. *Language Testing* 23.2, 131-166.

Lee, Yong-Won and Robert Kantor. (2007). Evaluating prototype tasks and alternative rating schemes for a new ESL writing test through G-theory. *International Journal of Testing* 7.4, 353-385.

Linacre, John M. (1989). *Many-facet Rasch measurement*. Chicago: MESA Press.

Linacre, John M. (1998). *A user's guide to FACETS: A Rasch measurement computer*

program. Chicago: MESA Press.

Lunz, Mary. E. and John A. Stahl. (1990) Judge consistency and severity across grading periods. *Evaluation and the Health Profession* 13, 425-44.

Lynch, Brian K., and McNamara, Tim F. (1998). Using G-theory and many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing* 15.2, 158-80.

Kondo-Brown, Kimi. (2002). A FACETS analysis of rater bias in measuring Japanese writing performance. *Language Testing* 19.1, 3-31.

McNamara, Tim. (1996). *Measuring second language performance.* New York: Addison Wesley Longman.

Myford, Carol and Edward Wolfe. (2000a). *Monitoring sources of variability within the Test of Spoken English assessment system* (TOEFL Research Report 65). Princeton, NJ: ETS.

Myford, Carol and Edward Wolfe. (2000b). *Strengthening the ties that bind: Improving the linking network in sparsely connected rating designs* (TOEFL Research Report 65). Princeton, NJ: ETS.

Plakans, Lia. (2015). Integrated second language writing assessment: Why? what? how? *Language and Linguistic Compass* 9.4, 159-167.

Read, John. (1990). Providing relevant content in an EAP writing test. *English for Specific Purposes* 9, 109-121.

Sawaki, Yasuyo, Thomas Quinlan, and Yong-Won Lee. (2013). Understanding learner strengthsand weaknesses: Measuring performance on an integrated writing task. *Language Assessment Quarterly* 10.1, 73-95.

Tyndall, Belle and Dorry M. Kenyon. (1995). Validation of a new holistic rating scale using Rasch multifaceted analysis. In Alister Cumming and Richard Berwick (Eds*), Validation in language testing* (pp. 39-57). Clevedon, England: Multilingual Matters.

Upshur, John A. and Carolyn E. Turner. (1999). Systematic effects in the rating of second language speaking ability: Test method and learner discourse. *Language Testing* 16.1, 82-111.

Weigle, Sara C. (1998). Using FACETS to model rater training effects. *Language Testing* 15.2, 263-87.

Weir, Cyril J. (1993). *Understanding and developing language tests.* Hemel, Hamstead: Prentice Hall.

Wesche, Marjorie B. (1987). Second language performance testing: The Ontario test of ESL as an example. *Language Testing* 4, 28-47.

Wiggleworth, Gillian. (1993). Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing* 10.3, 305-335.

Yong-Won Lee
Department of English Language and Literature
Seoul National University (SNU)
1 Gwanak-ro, Gwanak-gu, Seoul, Korea
Tel: (02) 880-2536/H.P.: 010-2796-1777
Email: ylee01@snu.ac.krr

Robert Kantor
Research Division,
Educational Testing Service (ETS)
660 Rosedale Road
Princeton, New Jersey 08540, USA
Email: rkantor@ets.org

Received: October 31, 2015
Revised version received: December 16, 2015
Accepted: December 20, 2015