Pan-Pacific Association of Applied Linguistics 14(2), 1-23

Examination of Rater Training Effect and Rater Eligibility in L2 Performance Assessment[1]

Yusuke Kondo
Ritsumeikan University

Kondo, Y. (2010). Examination of rater training effect and rater eligibility in L2 performance assessment. ***Journal of Pan-Pacific Association of Applied Linguistics, 14*****(2), 1-23.**

The purposes of this study were to investigate the effects of rater training in an L2 performance assessment and to examine the eligibility of L2 users of English as raters in L2 performance assessment. Rater training was conducted in order for raters to clearly understand the criteria, the evaluation items, and the evaluation procedure.In this evaluation, the rater training was conducted based on Common European Framework of Reference (CEFR).In the training, the raters watched the videos (North and Hughes, 2003), and discussed the learners' characteristics at each level. The analyses of the evaluations were done before and after the rater training based on Generalizability (G-Theory) and Multifaceted Rasch Analysis (MFRA). In the analyses based on G-Theory, the variance related to the items was reduced to about one sixth after the training, though no difference was found in the rater characteristics before and after the training in the analysis based on MFRA. Comparing the results of Kim (2009) with those of the present study, the raters are equally self-consistent with the raters of native speakers of English in Kim (ibid). Furthermore, it is legitimate to adopt L2 users as the raters, because, in countries where English is a foreign or second language, the non-native users teach and learn English. In this situation, teachers of L2 users are the most appropriate in L2 performance assessment if they are self-consistent in their ratings.

Key Words: CEFR, L2 performance assessment, rater, G-Theory, MFRA

1 Introduction

In second language (L2) performance assessment, there is much debate about the eligibility of the raters. According to Canagarajah (1999), eighty per cent of English language teachers in the world are non-native speakers of English. Japanese secondary education follows the similar pattern: there are only

[1]An earlier version of this paper first appeared as Kondo, Y. (2010). A tentative method of reforming your assessment of English abilities into international standards such as Common European Framework of Reference (CEFR) (1): The eligibility of raters and rater training effect in L2 performance assessment. *Proceedings of the 15th International Conference of Pan-Pacific Association of Applied Linguistics*, 436-443.

about 4700 English teachers who are the native speakers of English. That means that one native English teacher has 1200 students in the secondary education (Takanashi, 2009). The situation indicates that, generally speaking, learners of English have the slightest chance to be evaluated by the native speakers of English. In addition, from the view point of World Englishes, English users are considered to be more eligible as educators than the native speakers of English (McKay, 2002). Furthermore, the results of Kim (2009) indicate that non-native speakers of English were able to function as reliable raters in L2 performance evaluation.

Two approaches to L2 performance evaluation, Generalizability Theory (G-Theory) and Multifaceted Rasch Analysis (MFRA) have been used as complementary methods to investigate the reliability of the evaluation and the consistency of raters' evaluation.The previous studies indicated the usefulness of the information on the evaluation produced by these two methods in the reliability examination in L2 performance evaluation (Lumley and McNamara, 1995).

In the evaluation reported here, raters evaluated recorded self-introduction speech made by Asian learners of English before and after rater training.The purpose of the study is to investigate the effect of rater training.The study focuses on the change of reliability of the evaluation applying the information provided by G-Theory and the changes of raters' internal consistency, and also investigates the change of raters'consistency and severity, applying the information on raters' behaviors produced by MFRA through the rater training.

2 Literature Review

2.1 Raters in second language performance assessment

From the view point of World Englishes, English users are considered to be more eligible as educators than the native speakers of English (McKay, 2002).Now English is an international language that serves communities of businessmen and researchers all over the world.English "provides for effective communication, but at the same time it establishes the status and stability of the institutional conventions which defines these international activities". New Englishes are locally developed in such community.The native speakers of English, British or American are irrelevant to such Englishes (Widdowson, 2003, p. 40).Widdowson implies that learners of English have various purposes of learning English; to be a member of the native speakers' community is one of their purposes of English learning.To acquire the competence of the native speakers of English is one of the final goals of their English learning.The final goal of vast majority of learners of English is to be a member of international communities where English is used as a communication tool (McKay, ibid).Against this background,

although the quality of communication and standards of intelligibility are not assured if we fail to preserve standard (if Englishes used in the world are not mutually intelligible, the purpose of learning English disappears), English users are more eligible as educators than the native speakers of English, because English users are more knowledgeable in English learning in their community, which is no longer relevant to the native speakers of English.

Norcini and Shea (1997) mentioned, in the context of standard settings, that the most important factor in developing a credible standard is qualified standard setters.The same can be said on L2 performance assessment.Raters must be knowledgeable in their evaluation and their examinees, and particularly must be certificated.Furthermore, in L2 performance assessment, they must understand the context of learning the target language.The eligibility of raters is one of the issues to be considered in L2 performance assessment, because the property of raters, such as severity and consistency, might be influenced by their experience and language background.For these reasons, experienced Japanese language teachers were chosen as raters in the present study, because they are conversant with Asian learners of English and with the context of English language education in a situation where English is learnt as a foreign language.In addition, the rating by non-native language teachers of English is fairly realistic for Asian learners of English.According to Canagarajah (1999), eighty per cent of English language teachers in the world are non-native speakers of English.Japanese secondary education follows the similar pattern: there are only about 4700 English teachers who are the native speakers of English.That means that one native English teacher has 1200 students in the secondary education (Takanashi, 2009).In the Takanashi's data, only the students in public school were included.If the number of the students in private schools is added, that of students per one native English teacher will explode.The situation indicates that, generally speaking, learners of English have the slightest chance to be evaluated by the native speakers of English.

However, the eligibility of L2 users as a rater in L2 performance evaluation is questionable.Kim (2009) gave an answer to this question.She investigated the differences of rating behaviors between Korean teachers and Canadian teachers in evaluation of an oral proficiency test administered to ten Korean students at a university.The evaluations were analyzed based on MFRA.The index of self-consistency in the evaluation adopted in this study were fit statistics, proportions of large standard residuals between observed and expected scores, and a single rater-rest of the raters correlation.The results revealed that in the severity and the self-consistency, there was little difference between non-native speakers and native speakers of English.The two groups of teachers showed the same pattern in the severity of the evaluation, and all teachers fell into the acceptable range of the self-consistency.Kim (ibid), according to the results, concluded that non-native speakers of English were able to function as reliable raters in L2

performance evaluation, with the caveat that the results of the study might not be applied to other L2 performance evaluation, because only Canadian and Korean teachers were included as the raters.

Assessments of human performance require a number of raters, because no one evaluation can be definitive.A number of raters will be needed to obtain valid evaluation of human performance.Raters do not always agree, however.Therefore, rater training is usually conducted in order to achieve certain agreement among raters.As recent studies on L2 performance evaluation revealed (Lunz, Wright, and Linacre, 1990; Weigle, 1998), rater training is not capable of letting raters to achieve the same level of severity, but to make the raters self-consistent.As shown in Weigle (1998), rater variability cannot be eliminated, but extreme differences can be reduced.However, because the difference of the severity among raters can be modeled in MFRA to some extent (McNamara, 1996), the reduction of the variability in raters' severity is not a main purpose of rater training, but the focal point of rater training is to let raters to be internally consistent in their evaluation.

2.2 Reliability measurement[2]

Reliability, which is generally examined by statistical analysis, is defined as the degree of coincidence of test scores when two or more tests are conducted to measure the same characteristic of examinees (Ikeda 1994).In this study, the reliability of performance evaluation by raters was examined by using G-Theory (Brennan, 1992).In this section, we review the reliability of measurement in CTT, and then, outline the concept and the procedure of G-Theory.

In CTT, it is assumed that a test score consists of true score and error as shown in the equation (1):

$$(1)\quad X = T + E$$

where X is an observed score; T, a true score; and E, an error of measurement.The correlation between the true score T and the error E is expected to be zero. As a result, the variance of the observed score is expressed as in the equation (2):

$$(2)\quad \sigma_X^2 = \sigma_T^2 + \sigma_E^2$$

where σ_X^2 is the variance of the observed score X; σ_T^2, that of the true score T;and σ_E^2, the error E.Under the hypothesis of the equation (1), the index of reliability is expressed as in the equation (3):

[2] This section is based on Ikeda (1994) and Brennan (1992)

$$(3)\quad \rho_X = \frac{\sigma_T^2}{\sigma_X^2}$$

This index $\rho_X$is called coefficient of reliability. Since the variance of the true score is indeterminate, the various methods are adopted to estimate the reliability of measurement such as split-halves method, parallel test method, and so on. However, in CTT, the components of the error are not specified in a single analysis.

G-Theory, on the other hand, specifies multiple sources of measurement error in performance test. It is based on CTT and adopts the method of Analysis of Variance (ANOVA). Furthermore, the cost of performance evaluation can be estimated, such as number of raters and evaluation items.

G-Theory is a measurement model by which we can detect two or more sources of measurement error in test scores.In this section, the procedure of a two-crossed design in G-Theory is described. This design is of typical in performance assessment where we have two facets of measurements: raters and evaluation items.The analysis based on G-Theory consists of two steps: a generalizability study (G study) and a decision study (D study).The purpose of the G study is to estimate the relative effects of the respective sources of variance.In the D study, using the information of the variance components estimated in G study, we can assume the reliability of the test scores under several operational conditions.In this case, using the information estimated in the G study, we can assume the reliability of the test scores if we change the number of the items and the raters in the evaluation.

Suppose that examinees (e) do self-introduction task, and raters (r) evaluate the examinees performance using evaluation items (t). In the method of ANOVA, any observed score for a single evaluation item evaluated by a single rater can be expressed as:

$$(4)\quad X_{etr} = \mu + V_e + V_t + V_r + V_{et} + V_{er} + V_{tr} + V_{etr}$$

where μ is the grand mean in the population, and V stands for variances.Because of the orthogonality of each variance component, the population variance of X_{etr}can be deconstructed as:

$$(5)\quad \sigma^2(X_{etr}) = \sigma^2(e) + \sigma^2(t) + \sigma^2(r) + \sigma^2(et) + \sigma^2(er) + \sigma^2(tr) + \sigma^2(etr)$$

This is also represented in Figure 1 in terms of Venn diagram. Table 1 summarizes the expected mean squares and estimated variances of each variation factor.

Figure 1.Venn diagram for the variances of person, task, and rater based on Brennan (1992)

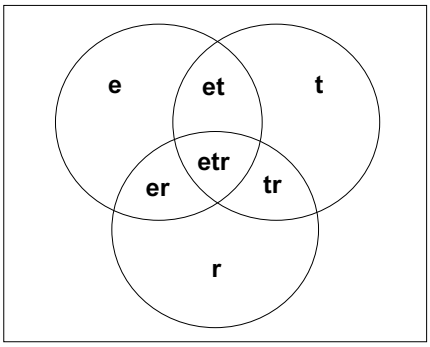


Table 1.Expected Mean Squares and Estimated Variances Based on Ikeda (1994)

Variation factor	Expected mean square	Estimated variance
e	$\sigma^2_{etr} + n\sigma^2_{er} + r\sigma^2_{et} + nr\sigma^2_{e}$	$\hat{\sigma}^2_{e} = [MS_{e} - MS_{et} - MS_{er} + Ms_{etr}] / nr$
t	$\sigma^2_{etr} + N\sigma^2_{tr} + r\sigma^2_{et} + Nr\sigma^2_{t}$	$\hat{\sigma}^2_{t} = [MS_{t} - MS_{et} - MS_{tr} + Ms_{etr}] / Nr$
r	$\sigma^2_{etr} + N\sigma^2_{tr} + n\sigma^2_{et} + Nn\sigma^2_{t}$	$\hat{\sigma}^2_{r} = [MS_{r} - MS_{er} - MS_{tr} + Ms_{etr}] / Nn$
et	$\sigma^2_{etr} + r\sigma^2_{et}$	$\hat{\sigma}^2_{et} = [MS_{et} - MS_{etr}] / r$
er	$\sigma^2_{etr} + n\sigma^2_{er}$	$\hat{\sigma}^2_{er} = [MS_{er} - MS_{etr}] / n$
tr	$\sigma^2_{etr} + N\sigma^2_{tr}$	$\hat{\sigma}^2_{tr} = [MS_{tr} - MS_{etr}] / N$
etr	σ^2_{etr}	$\hat{\sigma}^2_{etr} = MS_{etr}$

*Notes.*e stands for examinee; t, item; and r, rater.

Utilizing the information of the variance components specified in G study, the cost of performance evaluation can be estimated in D study.Adopting the model expressed in the equation (4), the score of an examinee (e) is expressed as:

(6) $\mu_e = \mu + V_e = \tau_e$

The difference between the grand mean and an observed score expressed in (7) is called absolute error:

(7) $\Delta_e = X_{etr} - \mu_e$

Utilizing these variables, index of dependability (Φ) is defined as:

$$(8)\quad \Phi = \frac{\sigma^2(\tau)}{\sigma^2(\tau)+\sigma^2(\Delta)}$$

$\sigma^2(\tau)$and$\sigma^2(\Delta)$can be found by the variance components specified G study as:

$$(9)\quad \hat{\sigma}^2\,(\tau) = \hat{\sigma}_e^2$$

$$(10)\quad \hat{\sigma}^2\,(\Delta) = \frac{\hat{\sigma}_t^2+\hat{\sigma}_{et}^2}{n'} + \frac{\hat{\sigma}_r^2+\hat{\sigma}_{er}^2}{r'} + \frac{\hat{\sigma}_{tr}^2+\hat{\sigma}_{etr}^2}{n'r'}$$

Utilizing the variance components specified in G study, Φ can be found by assigning value to n'and r' in the equation (10). This index is the G-Theory analogue of a reliability coefficient in CTT. This procedure helps us to estimate the cost of performance evaluation.

2.3 Item analysis

In the examination of test items, we usually use two indices of item characteristics: correction rate and discrimination power. The correction rate of binary item p_j is defined as:

$$(11)\quad p_j = \frac{1}{N}\sum_{i=1}^{N} u_{ij}$$

where N is the number of examinees, and u_{ij} is examinees' responses: 0-1.The correction rate falls between 0 and 1, and an easy item obtains larger value.This rate is used as the index of item difficulty.The discrimination power is defined as the correlation coefficient between item responses and the sum of the test scores of examinees'.The items with high discrimination power are considered to reflect the sum of the test scores if the test is composed to measure a single trait.Although these two indices give useful information to test designers, there is fundamental limitation to the item analysis in CTT.This model depends on the abilities of the test takers and the test itself, even if the scores are normalized.The indices of item difficulty and item discrimination power totally depend on sampling population in CTT.Hence, we cannot predict the test results of a given learner in CTT.Transgressing the limitation, however, we can analyze the items based on IRT.

IRT hypothesizes latent trait independent of examinee group.This trait is considered to be the same as factor one in factor analysis where items are dealt with as variables.In IRT, adopting cumulative normal distribution function, we can draw item characteristics curve (ICC) where y-axis indicates probability of correct response, and x-axis, the latent trait as described the equation (12):

$$(12)\ \Phi\big(f(\theta)\big) = \int_{-\infty}^{f(\theta)} \phi(z)dz$$

To describe item difficulty, function of θ in (12) is defined as $f(\theta) = a(\theta - b_j)$, and ICC of a given item, $item_j$ is defined as $p_j(\theta) = \Phi\left(a\big(\theta - b_j\big)\right)$.This is called one parameter normal ogive model.In this equation, a is the constant value in one parameter model, and b_j is the item difficulty index.Because only b_j determines the property of the ICC, it is called one parameter model.

Since the equation (12) includes integral equation, the approximate formula (13) is used for convenience in which D is a scaling factor, 1.7.When D equals 1.7, it is noted that the discrepancy of estimated θ is below .01.This one-parameter logistic model is called Rasch model.

$$(13)\ \int_{-\infty}^{f(\theta)} \phi(z)dz \cong \frac{1}{1+\exp\left(-Df(\theta)\right)}$$

In the actual situation, only available is examinees' responses, such as u'_i = [10110011].In the estimation in IRT, fixing u_i,θ is estimated by optimizing the equation below.

$$(14)\ L(u_i|\theta_i) = \prod_{j=1}^{n} p_j(\theta_i)^{u_{ij}}\, q_j(\theta_i)^{1-u_{ij}}$$

In this study, the evaluation scores are analyzed based on MFRA, which is an extension of Rasch model.Itis adopted because item properties, trait level, and rater's severity can be separately estimated. The model is depicted in the equation below:

$$(15)\ \log(P_{nmijk}/P_{nmijk}\text{-}1)=B_n - A_m - D_i - C_j - F_k$$

where
B_n = ability of examinee n
A_m = difficulty of task m
D_i = difficulty of skill item i
C_j = severity of judge j
F_k = difficulty of category k relative to category k - 1
P_{nmijk} = probability of rating of k under these circumstances
P_{nmijk}-1 = probability of rating of k – 1

In MFRA in the present study, the Rating Scale Model was adopted, because the model assumes that the relative difficulties of the steps (intersections) within items (Embredson and Reise, 2000).The model is expressed as follows (Embredson and Reise, ibid):

$$(16)\ P_x(\theta) = \frac{\exp[\psi_x + x(\theta - \lambda_i)]}{\sum_{x=0}^{M} \exp[\psi_x + x(\theta - \lambda_i)]}$$

where$\psi_x = -\sum_{j=0}^{x} \delta_j$ and $\psi_0 = \psi_m = 0.\delta_j$is a category intersection parameter which describes each of the J = K – 1 category thresholds, and λ_i is a scale location parameter which expresses the relative difficulty of the particular item.

Raters and items can be excluded,based on the scores of infit calculated by MFRA. The score of infit "provides the size of the residuals, the differences between predicted and observed scores (McNamara, 1996).The infit is the weighted mean-squared residual which is the index of unexpected responses near the point in which decisions are made.In the case of raters, the infit of the raters indicates whether or notevaluations by the raters are inconsistent with the estimated ability of the examinees. The fit statistics produced by MFRA indicate the degree of individual raters' consistency in their ratings.An acceptable range of fit statistics can be fixed, but it depends on the context of the evaluation and the use of the results (Myford and Wolfe, 2004a; 2004b).The acceptable range of infit is "the mean ± twice the standard deviation of the mean score statistics" in the case where the population exceeded thirty (McNamara, ibid).In this study, this criterion was adopted.

Kondo-Brown (2002) analyzed the assessment of Japanese L2 writing, based on MFRA.Three examinees out of 234 were identified as misfits (they obtained extremely high/low fit scores).She examined the examinees with high infit scores, and found out that two of them were children of Japanese immigrants: one was who had lived in Japan for several years, and the other was who demonstrated fluent and accurate expressions, but could write neither kana nor kanji and wrote the essay in alphabet.Kondo-Brown (ibid) eliminated these examinees in the subsequent analyses, because they were not candidates who the test developers had assumed as the examinees of the test.In MFRA, in this way, it is possible to detect an examinee based on the fit statistics.

2.4 Standards in L2 performance assessment

Attempts have been made to describe the development of L2 learners' proficiency, which is essential in composing a test, developing a language learning curriculum, and self-evaluating language ability.However, as North and Schneider (1998) indicate, there is no language proficiency model that is empirically and theoretically valid, and the examination of validity of proficiency scales or descriptors involves extensive research.Therefore, as of this moment, we cannot obtain proficiency scales or descriptors based on an established language proficiency model.

An early study of the description of the development of L2

learners'proficiency, Foreign Service Institute (FSI) scales were developed in 1950s.FSI comes down to American Council on the Teaching of Foreign Language (ACTFL) Proficiency Guidelines (American Council for the Teaching of Foreign Languages, 1999).In ACTFL, learners are evaluated with ten levels in four language skills: listening, speaking, reading, and writing.In the evaluation methods provided by ACTFL, Oral Proficiency Interview (OPI) which take fifteen minutes to twenty five minutes, interviewers control the levels of questions to examinees and tasks for examinees to accomplish.Standing on the theoretical foundation of OPI, Standard Speaking Test (SST) was developed by ALC Press to meet the needs of Japanese learners of English (ALC Press, 2006).However, these two tests have been criticized for the low validity and reliability (e.g. Lee and Musumeci, 1988; Salaberry, 2000).Lee and Musumeci (1998) pointed out that the tasks in OPI and SST were not hierarchically arranged: the skills and the ability required in the tasks of higher levels do not postulate those required in the task of lower levels.Furthermore, Salaberry (2000) noticed that improvement had not been occurred in the ACTFL Tester Training Manual published in 1999 from the previous manual published in 1986.

Another framework of foreign or second language learning related to ACTFL is Canadian Language Benchmarks (CLB: Centre for Canadian Language Benchmarks, 2000).The purposes of CLB are to provide learners with indices to be used in the self-evaluation of L2 ability, and provide a commonly understood framework for language programs in Canada.In CLB, in terms of four language skills: listening, speaking, reading, and writing, learners are divided into twelve levels.In each level, in addition to can-do statements, typical examples of tasks and texts, performance indicators, and strategies to be taught are provided.However, CLB does not include descriptions of discrete knowledge and skills (e.g. pronunciation, grammar, and vocabulary).

The Europeancounterpart of ACTFL is CEFR (Council of Europe, 2001).CEFR is a widely used guideline on learning, teaching, and assessing L2 and describes six levels of learners with descriptors.In reception, production, and interaction, the descriptors of language proficiency in relation to learners' activities are listed with respect to the six levels. In addition to the descriptors in global scales, such as spoken interaction, and written production, CEFR presents the descriptors in local scales such as phonological control and grammatical accuracy.CEFR presents detailed descriptors which capture various aspects of learners' activities.The descriptors of CEFR are written, based on theories of language competence and scaled based on a theory of measurement.In CEFR, learners are initially divided into three levels; basic user, independent user, and proficient user, and then each level is divided into two levels, which makes the six levels; Breakthrough, Waystage, Threshold, Vantage, Effective Operational Proficiency, and Mastery. Each level is usually called A1, A2, B1, B2, C1,

and C2 respectively. The number of the levels is largely based on the works by Trim and Wilkins (e.g. Trim, 1978).The scaling of the descriptors has been examined by a large number of researches (Council of Europe, 2001).

In North and Schneider (1998), two projects were reported: the one is for English, and the other for French and German.The aim of the projects was to develop a scale of language proficiency in the forms of descriptors.This is a fundamental research on validation of descriptors and levels in CEFR.The projects consisted of three stages to scale the descriptors.In the first stages, descriptors were created based on models of communicative competence and language use, and then, the created descriptors were categorized into some groups, such as reception, interaction, and production.In the second stage, which they called qualitative validation, the quality and the classification of the descriptors were examined by language teachers.They held thirty two workshops attended by more than 292 teachers through these two projects for the qualitative validation of the descriptors.The purpose of this procedure was to ensure that teachers' thoughts were well represented in the pool of the descriptors.In this workshop the teachers discussed learners'performances and sorted the descriptors into some provisional ranks.Based on the discussion and the levels of descriptors sorted by the teachers, questionnaires were composed, and the teachers evaluated learners'performances by using the questionnaires.In the third stage, the statistical analyses of the questionnaires were done based on Multifaceted Rasch Analysis (MFRA).Some descriptors were excluded based on the fit statistics produced in MFRA and Differential Item Functioning.The quality, the classification, and the levels of the descriptors were validated by comparing the results of the two projects.Although these two projects were conducted in different context of language learning: the first project was for English, and the second was for French and German, the correlation of the difficulty of the descriptors in the two projects were almost identical (r = .99), and descriptors on similar issues were adjacently aligned.North and Schneider (ibid) concluded that these results, the coherence and the consistency of the scaling of the descriptors were attributed to the facts that the descriptors were organized and selected according to the models of the communicative competence and language use, that the quality of the descriptors were examined by language teachers, and that the analyses were done based on Item Response Theory (IRT).However, they reminded us that the interpretation of the descriptors were subject to the context of language learning, and mentioned that the provision of the validated scale of language proficiency was only the first step to the establishment of an assessment framework.

2.5 Two techniques in the examination of reliability in performance assessment

In this study L2 speech evaluations are analyzed based on G-Theory and MFRA.These two techniques work in a mutually complementary manner in

the analysis of performance assessment.While G-Theory detects the source of error in each facet: rater, item, and examinee, and on the other hand, MFRA provides information on specific raters, items, and examinees that reduce the reliability of the performance assessment.These two approaches to the analysis of performance assessment were often adopted by studies on L2 performance (Bachman, Lynch and Mason, 1995; Lumlyand McNamara, 1995; Weigle, 1998; Kozaki, 2004; Bonk and Ockey, 2003; Kondo-Brown, 2002).

Bachman, Lynch and Mason (ibid) and Lumley and McNamara (1995) adopted these two techniques to analyze the performance assessment of L2 speaking ability.Bachman et al. (ibid) used these two techniques to analyze the data of a foreign language (Spanish) performance assessment for the placement of students at University of California, and investigated the reliability of the assessment.They mentioned that test users must have adopted some models to detect multiple sources of measurement errors, and G-theory and MFRA were not anti-theoretical model of measurement, but they give us complementary information in the analysis of performance assessment.Lumley and McNamara (ibid), using these two approaches, analyzed a test of communicative skills in English as a Second Language for intending immigrants to Australia.They also concluded that G-theory and MFRA complemented one another: while G-theory provided general information to decide test design, and MFRA, on the other hand, provided specific information on individual examinees, raters, and items.These two early studies indicated the potentials of G-Theory and MFRA in the analysis of L2 performance assessment.

MFRA is adopted in several rating situations to investigate rater characteristics in L2 performance assessments.In Lumley and McNamara (1995), MFRA was adopted to investigate the stability of rater characteristics over a certain period.They set three rating occasions of the evaluations of a speaking test for health professionals.In the first two occasions, rater training was conducted to establish their reliability, and in the last occasion, no rater training was included.They made a comparison of rater characteristics among three occasions.The results showed the change of the rater characteristics through the three rating occasions, and Lumley and McNamara (ibid) concluded that the effect of rater training could not endure for long.This analysis was made possible by MFRA, estimating the severity of the raters independently of the data set.

Weigle (1998), furthermore, investigated the rater training effects in L2 essay writing.Sixty compositions in UCLA's English as a Second Language Placement Examination were evaluated by eight experienced and eight inexperienced raters with three evaluation items, rhetorical control, content, and language of 10-point scale.The rater training was conducted by the composition supervisor.In this rater training, the raters read "norming packets" with sample compositions rated in the previous examination,

compared their own rating with the rating in the previous examination, and lastly discussed the ratings with the supervisor.To investigate the effects of the rater training on the severity and the inconsistency of the experienced and inexperienced raters, the two sets of the evaluations were examined based on MFRA.Based on the comparison between the evaluations before and after the rater training, the following points were implied as the effects of the rater training.The raters tend to be in the similar levels of severity after the rater training, but this tendency was only for the inexperienced raters who showed extreme severity before the rater training.As for the experienced raters, almost no effect was found on rater variability in severity.The remarkable effect of the rater training, however, was the reduction of the inconsistency of raters' evaluation.This means that the individual raters evaluated the compositions more consistently after the rater training.Weigle (ibid) concluded that the rater training affects intra-rater reliability more strongly than inter-rater reliability.

MFRA is also applied in standard setting on performance assessment for certification in Japanese medical translation into English.In Kozaki (2004), the performances by trainees supervised by a translation expert were rated by translation experts and medical doctors.The raters evaluated the performance data along the analytic scales, such as schema conventions, information structure, grammar and vocabulary and graded pass-fail on the examinees.The ground rule of passing the examination was that at least three judges agreed to pass the examinee.Kozaki (ibid) firstly analyzed the evaluation ratings, based on the descriptive analysis and set the cut-off point of pass-fail in the analytic scale.The cut-off point in the analytic scales was the minimum of the average scores of the passers.In this analysis, some analytic scales were found to be against her assumption.In one analytic scale, the average score of the fail group was higher than that of the pass group.An examinee in the fail group obtained a higher score above the cut-off point than the others in the pass group.Based on the information provided by MFRA, Kozaki (ibid) concluded that these results attributed to the inconsistency and the severity of the raters and the difficulty of the scales.This study is an example indicating the advantage of Rasch Analysis over CTT.

As the previous studies indicated, these two techniques are useful; while G-Theory detects relative effects of variability attributable to facets, MFRA provides information on specific elements of evaluation, raters, items and examinees.G-Theory allows investigators to handle sources of error in performance assessment, and it is possible that it predicts the dependability (reliability) according to conditions of manipulating the number of items and raters.In MFRA, examinees' ability is estimated independently from the severity of the particular raters and the difficulty of particular evaluation items.Examinees' ability is estimated in relation to the severity of raters and the difficulty of items.Moreover, the inconsistency of raters and items with

the model can be excluded.In the present study, L2 performance evaluations were analyzed based on these two models.The analyses with G-Theory and MFRA were performed by the computer programs, GENOVA (Crick and Brennan, 1984) and FACETS (Linacre, 2006) respectively.

2.6 Method

2.6.1 Participants

Seventy three Asian learners of English participated as an examinee in this study.Their first languages are Thai, Japanese, Korean, Tagalog, Mandarin, and Taiwanese.They are graduate or undergraduate students.Their L2 background is summarized in Table 2.

Five Japanese raters with the master degree of Applied Linguistics participated in this study.Their average year of learning English was 18.3 with S.D. 6.5 and that of teaching English, 10.9 with S.D. 8.7.Their experiences of teaching English were not only in primary, secondary, and high school and university in Japan, but also some of them have taught English as an L2 to non-Japanese learners.Language teachers of non-native speakers of English were chosen, because of their knowledge on the context of learning English.

Table 2.Key Information of the Participants in Self-introduction Task

	M	*SD*	Range
Age	20.77	3.14	13
Study of English (year)	10.38	3.94	22

Notes.N = 73.

2.6.2 Recording procedure

All the recording was made in soundproof rooms in the universities which the participants belonged to.The participants were called in the room and given the instruction of recording individually.Their self-introductions without preparation were digital-tape recorded by using Roland R-09 and a condenser microphone, SONY ECM-MS957. In the recording, the participants gave their self-introduction to an interviewer, and the interviewer only gave approving nods. After the recording, the participants were given a small gift for their participation.It took about ten minutes for each participant to complete the recording.

2.6.3 Rating procedure

Evaluation items were selected from those in Yashiro, Araki, Higuchi,

Yamamoto, and Komissarov (2001), and each item was thoroughly reviewed in order to make the items suitable in the evaluation of unprepared L2 speech.The items are listed below.

1. Loudness
2. Sound pitch
3. Quality of vowels
4. Quality of consonants
5. Epenthesis
6. Elision
7. Word stress
8. Sentence stress
9. Speech rate
10. Prosody
11. Fluency
12. Place of fillers
13. Frequency of fillers
14. Place of pause
15. Frequency of pause
16. Length of silent pause
17. Paralinguistic cues
18. Confidence
19. Try to sound cheerful
20. Try to sound friendly
21. Grammatical accuracy
22. Coherency
23. Absence of tension
24. Foreign accentedness

Figure 2.A sample of the evaluation website

The raters evaluated the participants' speech on the website individually.On the website, the raters listened to and evaluated the recorded participants' speeches in view of overall proficiency and twenty four subcategories of overall proficiency where a 6-point Likert scale was

adopted.A sample of the evaluation website is shown in Figure 2.All the raters evaluated every speech in this evaluation in the same order.

2.6.4 Rater training procedure

Rater training was conducted according to the manual provided by Council of Europe (Council of Europe, 2003).The procedure of linking a test to CEFR consists of five steps: Familiarization, Specification, Standardization training and benchmarking, Standard setting, and Validation (Council of Europe, ibid: 10-11) in the present study.Firstly, raters received the overview of the speech data.They are unprepared self-introduction speech and recorded in the universities where the speakers belonged to, and the speakers were Asian learners of English who were graduate or undergraduate students.Then, the raters discussed the speech characteristics of the learners' and selected the evaluation items from Yashiro, et al (2001), listening to a couple of speech data.This stage is "Specification" of the evaluation in the manual.After the discussion on the speech characteristics of the learners, the raters were given the descriptors and the levels in CEFR and watched the video (North and Hughes, 2003) which depicted the learners divided into six levels.This stage is "Familiarization" to the descriptors and the levels in CEFR.Lastly, the raters discussed the descriptors and the levels in CEFR, watching the video, and discussed the characteristics of learner language in each level.This is the stage of "Standardization Training and Benchmarking" and a part of "Standard Setting" in the manual.Rater training was conducted three times during two weeks.This activity led the raters to establish the images of the learners of six levels.

2.7 Examination of rater training effects based on generalizability study

In this section the effect of rater training are reported in terms of reliability improvement.In the present study, the raters and the items were a random facet, because they could be exchanged with other raters, and evaluation items were also exchangeable, which could be taken from any other items related to the L2 speech assessment.All the examinees were evaluated by all the raters.Hence, the design of the Generalizability study (G study) was examinee × items × raters.The design of the G study is a random effect model with two facets: twenty four items and five raters, which assumes that the raters and the items interacted interchangeably.The focus of this study is dependability (reliability) of test scores with full facets.The estimated variances of each facet (e.g. examinee, item, and rater) were examined, and the indices of dependability were compared before and after our rater training.In this experiment, fifteen learners randomly selected from the participantsdescribed above,and five language teachers described above as raters.

Tables 3 and 4 show the results of the G study before and after the rater training.Comparing the estimated variances before and after the training, the

examinees' ability accounts for 43 per cent and 63 per cent, and the rater related variables, for 12 per cent and 8 per cent.A remarkable difference before and after our rater training is the difference in the estimated variances of the items.The estimated variance of items after the training is about one-sixth of that of items before the training.This suggests that the items (rating criteria) before the training differ much more in average difficulty than these after the training.In the rater training our raters watched the video where the learners of six levels were depicted.It must have helped the raters to clarify how they should scale.

Table 3.G Study before the Rater Training

	SS	*df*	*MS*	EV
e (examinee)	1518.55	14	108.47	0.43
t (item)	1886.89	23	82.04	0.51
r (rater)	425.41	9	47.27	0.12
Et	1056.77	322	3.28	0.28
Er	296.39	126	2.35	0.08
Tr	651.91	207	3.15	0.18
etr (residual)	1419.09	2898	0.49	0.49
Sum	7255.01	3599	247.05	2.08

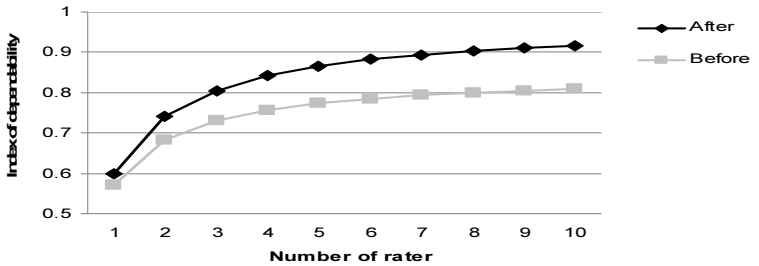
Note: *SS*: sum of squared deviation, *df*: degree of freedom, *MS*: Mean square, EV: Estimated variance.

Table 4. G Study after the Rater Training

	SS	*df*	*MS*	EV
e (examinee)	2221.45	14	158.68	0.63
t (item)	342.15	23	14.88	0.08
r (rater)	440.75	9	48.97	0.11
Et	397.76	322	1.24	0.08
Er	814.28	126	6.46	0.25
Tr	351.33	207	1.70	0.09
etr (residual)	1177.84	2898	0.41	0.41
Sum	5745.66	3599	232.32	1.66

Note: *SS*: sum of squared deviation, *df*: degree of freedom, *MS*: Mean square, EV: Estimated variance.

Figure 3.Change of index of dependability



Utilizing the information of the estimated variance specified in the G study, the indices of dependability Φ were calculated before and after the training. Generalizability coefficient is an index only for examinee-related factors, and the index of dependability, on the other hand, is an index for all variation factors including examinee-related factors. The former is larger than the latter.In the results of the D-studies, according to conditions; manipulating the number of items and raters, we can predict dependability in several conditions.

Table 5.The Index of Dependability before the Rater Training

		Rater									
		1	2	3	4	5	6	7	8	9	10
Item	1	.21	.27	.29	.31	.31	.32	.33	.33	.33	.34
	2	.33	.40	.44	.46	.47	.48	.48	.49	.49	.50
	3	.40	.49	.53	.55	.56	.57	.58	.58	.59	.59
	4	.45	.54	.58	.61	.62	.63	.64	.65	.65	.65
	5	.48	.58	.63	.65	.67	.68	.68	.69	.70	.70
	6	.51	.61	.66	.68	.70	.71	.72	.72	.73	.73
	7	.53	.64	.68	.71	.72	.73	.74	.75	.75	.76
	8	.55	.65	.70	.73	.74	.75	.76	.77	.77	.78
	9	.56	.67	.72	.74	.76	.77	.78	.79	.79	.80
	10	.57	.68	.73	.76	.77	.78	.79	.80	.80	.81

By comparing the results of the D-studies before and after the rater training, the cost reduced by the rater training is revealed.Evaluation conditions were simulated where one to ten rater(s) evaluated examinees using one to ten evaluation item(s).All the simulations are found in Tables 5 and 6.The evaluation condition of one to ten rater(s) using the ten items is described in Figure 3.With the acceptance that this index is the analogue of a reliability coefficient, the minimum value is .85 for a reliable evaluation.The change of Φdescribed in Figure 3 is the simulation where one to ten rater(s) evaluate(s) examinees using ten items.If the rater training is conducted, above 0.85 of Φcan be obtained by only four raters.

Table 6.The Index of Dependability after the Rater Training

		Rater									
		1	2	3	4	5	6	7	8	9	10
Item	1	.39	.52	.59	.63	.66	.68	.69	.70	.71	.72
	2	.48	.62	.69	.73	.76	.78	.79	.80	.81	.82
	3	.52	.67	.74	.77	.80	.82	.83	.84	.85	.86
	4	.55	.69	.76	.80	.82	.84	.85	.86	.87	.88
	5	.56	.71	.77	.81	.84	.85	.87	.88	.88	.89
	6	.57	.72	.78	.82	.85	.86	.88	.89	.89	.90
	7	.58	.73	.79	.83	.85	.87	.88	.89	.90	.91
	8	.59	.73	.80	.83	.86	.88	.89	.90	.90	.91
	9	.59	.74	.80	.84	.86	.88	.89	.90	.91	.91
	10	.60	.74	.81	.84	.87	.88	.89	.90	.91	.92

2.8 Examination of rater training effects based on MFRA

The evaluation scores before and after the rater training were independently analyzed based on MFRA.In the process of the analysis of the evaluation scores before the rater training, three items, “Paralinguistic cues”, “Absence of tension”, and “Foreign accentedness”, were found to be extremely inconsistent evaluations items, whose infits surpassed 3.00.Hence, these three items were excluded in this analysis.Table 7 shows the infits and the severity measures of the raters and the infits and the difficulty measures of the evaluation items before and after the rater training respectively.

Table 7.Infits and Severity of Raters before and after Rater Training

	Before training		After training	
	Infit	Severity	Infit	Severity
Rater 1	1.14	-0.74	1.24	-1.33
Rater 2	1.11	-0.16	1.22	-0.49
Rater 3	0.95	-0.40	0.91	-0.48
Rater 4	0.93	-0.12	0.86	-0.09
Rater 5	0.73	-0.47	0.82	-0.07
M	0.97	-0.33	1.01	-0.42
SD	0.16	-0.32	0.20	-0.67

The logit values of the severity before and after the rater training need to be adjusted to make them comparable with each other (Lumley and McNamara, 1995).Adding -0.09 to the each value of the severity before the training, the two sets of the severity were compared by t-test.There were no difference in the severity measure of raters before and after the training ($t(4) = 0.56$, $p = .60$ (two-tailed)).As for the index of the self-consistency in the raters, the infit, no inconsistent raters were found both before and after the

training in the condition that the upper and lower limit of the fit statistics are set to 1.4 and 0.6, respectively (Wright and Linacre, 1994).

3 Discussion and conclusion

The purpose of the study was to investigate the effects of rater training in an L2 performance evaluation.Rater training was conducted in order for raters to clearly understand the criteria, the evaluation items, and the evaluation procedure.In the training, the raters watched the videos (North and Hughes, 2003), and discussed the learners' characteristics at each level.The analyses of the evaluations were done before and after the rater training based on G-Theory and MFRA.In the analyses based on G-Theory, the variance related to the items was reduced to about one sixth after the training, though no difference was found in the rater characteristics before and after the training in the analysis based on MFRA.

These results might be mainly attributed to the background of the raters in this study.The raters in these evaluations are familiar with the context of learning English in Asia.They also know the learners themselves.It is the reason why the raters were self-consistent before the training.In the analysis by Weigle (1998), inexperienced raters tended to be self-inconsistent, while experienced raters were self-consistent before the training.However, as the results of G study in the present study shows, the variance related to the evaluation items were reduced after the training.This is because the raters might have understood the contents of the evaluation items better through the training.In performance evaluation, the difficulty and the consistency of the evaluation items are greatly influenced by raters' understanding of the contents of items.In the present study, no difference were found in raters' characteristics in the results of MRFA, but the variance related to the evaluation items were found to be reduced in the results of G study.This can be said to be one of the effects of the rater training.

The other finding of this study is about the eligibility of the raters whose first language is not English in L2 performance evaluation.Comparing the results of Kim (2009) with those of the present study, our raters were equally self-consistent with the raters of native speakers of English in Kim (2009).Furthermore, it is legitimate to adopt L2 users as the raters, because, in countries where English is a foreign or second language, the non-native users teach and learn English.In this situation, teachers of L2 users are the most appropriate in L2 performance evaluation if they are self-consistent in their ratings.

The raters in this study were Japanese language teachers of English, though the learners' speech data were collected widely from Asia.If raters share their first language with learners, it may influence on their evaluation.The answer to this question could not be found in the results of the present study.

References

ALC Press. (2006). SST: *Standard Speaking Test*. Retrieved May 30, 2010, from http://www.alc.co.jp/edusys/sst/english.html

American Council for the Teaching of Foreign Languages.(1999). *ACTFL Proficiency Guidelines*.Retrieved March, 20, 2010, from http://www.sil.org/lingualinks/languagelearning/OtherResources/ACTFLProficiencyGuidelines/contents.htm

Bachman, L. F., Lynch, B. K., & Mason, M. (1995).Investigating variability in tasks and rater judgments in a performance test of foreign language speaking.*Language Testing*, *12*(2), 238-257.

Bonk, W. J., &Ockey, G. J. (2003). A many-facet Rasch analysis of the second language group oral discussion task.*Language Testing*, *20* (1), 89-110.

Brennan, R. L. (1992). *Generalizability theory. ITEMS: The instructional topics in educational measurement series*. Module 14. Madison: NCME.

Canagarajah, A. S. (1999). Interrogating the "native speaker fallacy": Non-linguistic roots, non-pedagogical results,In G. Braine (Ed.),*Non-native educators in English language teaching*. Mahwah, NJ: Lawrence Erlbaum Associates.

Centre for Canadian Language Benchmarks. (2000). *Canadian language benchmarks 2000*. Retrieved March 20, 2010, from http://www.language.ca/pdfs/clb_adults.pdf

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge: CUP.

Crick, J. E., & Brennan, R. L. (1984). GENOVA: A general purpose analysis of variance system. Version 2.2 [Computer software]. Iowa: American College Testing Program.

Embretson, S. E., &Reise, S. P. (2000).*Item response theory for psychologists*.Mahwah: Lawrence Erlbaum Associates, Inc.

Ikeda, H. (1994). Gendai test riron. [Contemporary test theory]. Tokyo: AsakuraShoten.

Kim, Y. (2009). An investigation into native and non-native teachers' judgments of oral English performance: A mixed methods approach. *Language Testing*, *26*(2), 187-217.

Kondo, Y. (2010). A tentative method of reforming your assessment of English abilities into international standards such as Common European Framework of Reference (CEFR) (1): The eligibility of raters and rater training effect in L2 performance assessment. *Proceedings of the 15th International Conference of Pan-Pacific Association of Applied Linguistics*, 436-443.

Kondo, Y., &Brown, K. (2002).A FACETS analysis of rater bias in measuring Japanese second language writing performance.*Language Testing*, *19*(1), 3-31.

Kozaki, Y. (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.

Lee, J. F., &Musumeci, D. (1988).On hierarchies of reading skills and text types.*Modern Language Journal*, *72*, 173-187.

Linacre, J. M.(2006) Facets Rasch measurement [Computer software]. Chicago: Winsteps.com.

Lumley, T., & McNamara, T. F. (1995).Rater characteristics and rater bias: implications for training. *Language Testing*, *12*(1).54-71.

Lunz, M. E., Wright, B. D., & Linacre, J. M. (1990).Measuring the impact of judge severity on examination scores.*Applied Measurement in Education*, *3*(4), 331-345.

McKay, S. L. (2002). *Teaching English as an international language*. Oxford: OUP.

McNamara, T. F. (1996). *Measuring second language performance*. Essex: Pearson Education Limited.

Myford, C. M. & Wolfe, E. W. (2004a). Detecting and measuring rater effects using many-face Rasch measurement: Part I. In Jr.Smith, E. V., &R. M. Smith(Eds.),*Introduction to Rasch measurement*. Maple Grove, MN: JAM Press. 460-517.

_____.(2004b). Detecting and measuring rater effects using many-face Rasch measurement: Part II. In Jr.Smith, E. V., &R. M. Smith (Eds).*Introduction to Rasch measurement*. Maple Grove, MN: JAM Press. 518-574.

Norcini, J. J., & Shea, J. A. (1997).The credibility and comparability of standards.*Applied Measurement in Education*, *10*(1), 39-59.

North, B.,& Hughes, G. (2003). CEF illustrative performance samples for relating language examinations to the CEF of languages: Learning, teaching, assessment (CEF) English (Swiss adult learners). Eurocentres.

North, B., & Schneider, G. (1998). Scaling descriptors for language proficiency scales. *Language Testing*, *14*(2), 217-263.

Salaberry, R. (2000). Revising the revised format of the ACTFL oral proficiency interview.*Language Testing*, *17*, 289 - 310.

Takanashi, Y. (2009). *Data de yomueigokyouiku no joushiki*.[Understanding common knowledge of English language education by data]. Tokyo: Kenkyusha.

Trim, J. L. M. (1978). *Some possible lines of development of an overall structure for a European unit credit scheme for foreign language learning by adults*. Council of Europe.

Weigle, S. C. (1998). Using FACETS to model rater training effects.*Language Testing*, *15*(2),263-287.

Widdowson, H. G. (2003). *Defining issues in English language teaching*. Oxford: OUP.

Wright, B. D., & Linacre, L. M. (1994).Reasonable mean-square fit values.*Raschmeasurement: Transaction of the Rasch MeasurementSIG*, *8*, 370.

Yashiro, K., Araki, A., Higuchi, Y., Yamamoto, S., &Komissarov, K. (2001).*Ibunka communication workbook*.[A workbook for cross-cultural communication]. Tokyo: Sanshusha.

Yusuke Kondo
Language Education Center, Ritsumeikan University
377-302 Motohonnojicho Nakagyo-ku Kyoto-shi,
Kyoto, Japan 604-8244
Tel& Fax: +81 (0)75-708-2523
Email: ykondo@fc.ritsumei

Received: August 31, 2010
Revised: November 30, 2010
Accepted: December 5, 2010