JOURNAL OF APPLIED MEASUREMENT, *10*(4), 408-423

A Comparative Analysis of the Ratings in Performance Assessment Using Generalizability Theory and The Many-Facet Rasch Model

Sungsook C. Kim
Korea Institute for Curriculum and Evaluaton (KICE)

Mark Wilson
University of California, Berkeley

The purpose of this study is to compare two different methods for modeling rater effects in performance assessment: Generalizability (G) Theory and the Many-facet Rasch Model (MFRM). The view that G theory and the MFRM are alternative solutions to the same measurement problem, in particular, rater effects, is seen to be only partially true. G theory provides a general summary including an estimation of the relative influence of each facet on a measure and the reliability of a decision based on the data. MFRM concentrates on the individual examinee or rater and provides as fair a measure as it is possible to derive from the data as well as summary information such as reliability indices and ways to express the relative influence of the facets. These conclusions are illustrated using data for ratings of student writing assessments.

Requests for reprints should be sent to Sungsook C. Kim, Korea Institute for Curriculum and Evaluation, 142 Gahoe-ro, Jongno-gu, Seoul 110-230, Korea, e-mail: sungs@kice.re.kr.

Introduction

Today, many educators are considering augmenting their multiple-choice items with performance-based items or tasks. Performance-based assessment is an authentic evaluation which attempts to grasp what and how well students can perform in life-like situations. Examples of performance assessment are observation, portfolio, reports, etc. However, when teachers assess students' achievement by these methods, one problem that must be faced is controlling the dependability of scores in different rating or grading systems. In order to improve the reliability of performance-based assessment, we need a technical framework to control it. The process of rating each performance contains a number of potential sources of error associated with raters, with tasks, with occasion, with rating criteria, with different scales, and with their interactions.

More specifically, the scoring of performance assessments contains many possible errors such as rater disagreement, lack of objectivity, the potential effects of an unclear rating guide, and non-systematic changes over time in raters and environment. More specifically, for instance, we can view a performance assessment as a sample of student achievement drawn from a complex universe defined by a combination of all possible tasks, occasions, raters and rating standards. We can view the task facet to be representative of the content in a subject-matter domain. The occasion facet potentially includes all possible occasions on which a student would perform at that particular time. We can view the rater facet as including all possible individuals who could be trained to score the tasks reliably. Also, the type of rating system includes different types of ratings or standards used in the rating process.

To better understand the structure of performance assessments, it is useful to break the measurement down into its facets so that the influence of each facet on the score can be observed. Recently, Stahl (1994), Linacre (1989, 1995), Lunz and Schumacker (1997), and Marcoulides (1997, 1999) provided comparisons of generalizability theory (G theory) and the many-faceted Rasch IRT model (MFRM) and illustrated how both methods can be used to provide information concerning future test construction. Also, MacMillan (2000) conducted a study of classical test theory (CTT), G theory and MFRM approaches to detecting and correcting for rater variability. Although G theory and MFRM were found to provide similar information with respect to measurement designs that involve judged ratings, G theory was criticized for its focus on group behavior and the interaction of groups. They noted that a MFRM analysis produced calibrated person measures that were adjusted for the facets included in the measurement design, something that G theory did not provide. Recently, Wilson and Hoskens (2001) have demonstrated that MFRM systematically overestimates reliability where multiple raters rate the same piece of work, and have demonstrated a method to deal with the problem.

The purpose of this study is to compare the results of analyzing data from a performance assessment (written composition) using different approaches. The methods to be studied are G theory and the Many-Facet Rasch Model (MFRM) which have been the most frequently applied psychometric methods lately in evaluating the reliability of performance assessment. In particular, the study discusses the similarities and differences of results related to each main and interaction effect. The study also presents a further interpretation based on understanding the merits of the two methods for multifaceted observations.

In general, the G theory approach is fundamentally an observed score analysis procedure and provides a general summary including the relative influence of each facet on a measure and the reliability of the decision estimated from data collection. The MFRM approach is derived from a series of mathematical axioms that can possess (so long as the data conform to the model to a reasonable extent) the desirable properties of invariance of parameters and objective linear measures and concentrates on a measure as fair as it is possible to derive from the data. It too provides summary indices, such as reliability

coefficients and indicators of facet impact. In order to develop a perspective from which to understand the two methods, the basic concepts of G theory and MFRM and their approaches to performance assessment are reviewed in the next section.

Review of the Methods

Generalizability Theory

Cronbach, Gleser, Nanda, and Rajaratnam (1972) introduced G Theory as a measurement theory about the dependability of behavioral measures. Dependability refers to the accuracy of generalizing from a person's observed score on a measure or a test to the score that the person would have received averaged over all possible conditions. In G theory, a behavioral measurement is considered a sample from a universe of admissible observations described by one or more sources of variation, called facets. To apply G Theory to a real situation, two studies may be conducted, a G study and a D study. The G study defines the universe of admissible observations. To do this, (1) it simultaneously analyzes the multiple sources of error in a measurement and (2) it compares the relative influence of each facet on a measure. Based on the results of the G study, the D study defines the universe of generalization. It also (3) provides a generalizability coefficient reflecting the 'reliability' of generalizing from a sample score and (4) allows one to estimate the number of levels of each facet that are needed to attain a certain level of generalizability (reliability).

G Theory uses an analysis of variance approach based on the raw scores to provide estimates of scoring variation due to raters, occasions, evaluation standards, or other sources of error. By estimating the magnitude of the variance components, the sources of the greatest measurement error can be pinpointed. It is important to recognize that the purpose of a G study is to obtain estimates of variance components associated with the universe of admissible observations. More importantly, these estimates can be used to design efficient measurement procedures that provide information for making substantive decisions about objects of measurement (in various D studies). The D study considers the specification of a universe of generalization, which is the universe to which the decision maker wants to generalize in the D study. The universe score is much like a true score. In classical test theory, one assumes that error variance is all of one kind and that a person has one true score. In contrast, G Theory recognizes alternative universes of generalization, therefore there are many universe scores. The assumption of G Theory is that the observed behavior is a random sample of all behavior. In G Theory, the generalizability of a measure depends on how the data will be used to in the decision study. Decisions based on the standing of individuals relative to one another are called relative decisions. In contrast, when people index the absolute level of an individual's performance without reference to how well or poorly his or her peers performed—these are called absolute decisions. The variance components contributing to measurement error are somewhat different for relative and absolute decisions. For relative decisions all variance components that influence the relative standing of individuals contribute to measurement. For absolute decisions, all error variance components except the object of measurement contribute to measurement error. Consider a study with two-facets and folly-crossed behavior observations in the classroom. For example, in the Beginning Teacher Evaluation Program (BTEP), teachers are coded by two raters on each of two occasions. The teachers are the object of measurement; teachers, raters and occasions are considered to be random effects, that is, randomly sampled from large universe. This is the persons (p) by raters (r) by occasions (o) design and is denoted as $p \times r \times o$. Any person's observed score in this design can be decomposed into seven variance components:

$$\sigma_x^2 = \sigma_p^2 + \sigma_r^2 + \sigma_o^2 + \sigma_{pr}^2 + \sigma_{po}^2 + \sigma_{ro}^2 + \sigma_{pro,e}^2. \quad (1)$$

The variance component for person, σ_p^2, also called universe-score variance, shows how much teachers' differences in their performances affect the ratings.

The variance component for raters, σ_r^2, shows the extent to which rater differences affect the ratings.

The variance component for occasions, σ_o^2, shows the extent to which the occasion differences affect the ratings.

The person-rater interaction, σ_{pr}^2, shows the extent to which the relative standing of teachers in their performances affect the ratings.

The variance component for the person-occasion interaction, σ_{po}^2, shows the extent to which the inconsistencies in relative standing of teachers from one occasion to the next affect the ratings.

The variance component for the rater-occasion interaction, σ_{ro}^2, shows the extent to which the inconsistency in raters' average rating of teachers from one occasion to the next affect the ratings.

Finally, the residual variance component, $\sigma_{pro,e}^2$, reflects the three way interactions between persons, raters, and occasions, confounded with unmeasured sources of variation, affect the ratings.

Although it focuses on variance components, G theory also has a reliability coefficient, analogous to the reliability coefficient in classical test theory, called the *generalizability coefficient*. It is defined as the universe score variance divided by the expected observed-score variance—i.e., it is the proportion of expected observed-score variance that is also the universe score variance. The teacher behaviors are the object of measurement in this example, and the error variance includes the variation related to the interaction between teachers and raters, the interaction between teachers and occasions, and the residual. Focusing on a relative decision, all error variance components that influence the relative standing of individuals contribute to measurement error, which is called the estimated relative error variance ($\sigma(\delta)^2$)). Then the estimated generalizability coefficient ($\rho(\delta)^2$)) can be expressed as follows:

$$\sigma(\delta)^2 = \frac{\sigma_{pr}^2}{n_r} + \frac{\sigma_{po}^2}{n_o} + \frac{\sigma_{pro,e}^2}{n_r n_o} \quad (2)$$

$$\rho(\delta)^2 = \frac{\sigma_{(\tau)}^2}{\sigma_{(\tau)}^2 + \sigma_{(\delta)}^2} = \frac{\sigma_p^2}{\sigma_p^2 + \left(\frac{\sigma_{pr}^2}{n_r} + \frac{\sigma_{po}^2}{n_o} + \frac{\sigma_{pro,e}^2}{n_r n_o}\right)}. \quad (3)$$

Since the generalizability coefficient is calculated based on one rater and one occasion, the D study provides a summary table which shows how to improve the coefficient by increasing the number of raters and occasions. The choice of number of raters and occasions depends on the magnitude of the measurement error, on the level of generalizability desired, and on practical considerations.

Several studies have applied generalizability theory (Brennan, 1983; Shavelson and Webb, 1991) to analyze multiple sources of variation in performance assessment. In addition, the effects of changing the number of observations in a single facet or two facet design on attaining a satisfactory level of the generalizability coefficient has been investigated extensively in the literature (Lane, Liu, Ankenmann and Stone, 1996; Baxter, et al., 1992; Kim, 1996, 2000; Lehmann, 1990). Marcoulides (1997) introduced an extension to G theory that can be used to provide specific information on the reliability of individual ability estimates and diagnostic information at the individual and group levels. This extension to G theory can be used to analyze performance assessments to provide three important pieces of information, (1) a diagnostic scatter diagram-that can be examined for unusual patterns for each examinee and each judge in the measurement design, (2) an index for examinees—that can be needed to examine person ability, and (3) an index for judges—that can be used to examine judge severity. More recently, Marcoulides (1999) emphasized that this extension to G theory can be considered a special type of IRT model capable of estimating latent traits such as examinee ability estimates, rater severity, and item difficulties.

The Many-Facet Rasch Model (MFRM)

The MFRM which is based on the linear logistic test model (Fischer, 1983) extends the basic Rasch model (Rasch, 1960, 1980) so that facets for effects such as judge severity are added to the model (Linacre, 1989). Facets are measurement conditions that are hypothesized to effect a persons' score, such as examinees being rated by different judges or being tested at different times or over different topics or tasks. This allows one to model the impact of error variance within each facet on the person's ability estimate. The probability of a satisfactory performance is a function of the difference between the student's ability and the task or step difficulty, after adjustment for the severity of the raters. The rater severity or leniency is the tendency on the part of raters to consistently provide ratings that are lower or higher than the means of the other raters. It can be viewed as a continuum on which raters range from lenient to severe. If the student's ability is higher than the difficulty of the tasks after adjustment for the judge severity, then the probability of a satisfactory performance is greater than .5. Using an extension of the partial credit model to this multifaceted context, Linacre (1989, 1995) has shown that this approach can be used to identify raters who are harsher or more lenient than others, who exhibit different patterns in the way they use rating schemes, and who make judgments that are inconsistent with judgments made by other raters. From a measurement design perspective, raw scores are obtained from individuals under certain defined conditions or facets, which in the MFRM are converted to logit measures. Given that a particular facet element may affect student's scores, it is important to determine whether the elements within each facet are significantly and meaningfully different.

Consider, for example, the assessment of writing where a stimulus is presented to a student, the student prepares a piece of writing, and then a rater makes a judgment about the quality of the writing performance. Here, the object of measurement is the student, but the agent (facet) is a combination of the rater who makes the judgment and the stimulus that serves as a prompt for the student's writing, and the interaction of all three. The response that is analyzed by the item response model is influenced by the characteristics of the student, the characteristics of the stimulus, and the characteristics of the rater. The exploration of rater effects is an important application of many-faceted models. Engelhard (1994) and Du and Wright (1997) have carried out foundational work on writing assessment using MFRM. The MFRM can be used to examine variation in the harshness and leniency of raters, to examine the propensity of raters to favor different response categories, and to examine the fit of individual raters with the model. It can also be used to examine the impact of various facets and to predict what the effects will be of changing the conditions with in the facets.

This model for this writing assessment has three facets—student competence, item difficulty and rater severity. The data are modeled to be the stochastic outcome of the logit-linear probability model based on an extension of the simple Rasch model. The MFRM model is a unidimensional model with a single student competence parameter, and a collection of other facets, such as items and raters. The probability of student n with competence B_n obtaining a rating of x (x = 0,1,...,k) on item D_i from rater C_j with category step difficulty F_k is given as

$$P(X_{nijk}=x)=\frac{\exp\left[x(B_n-D_i-C_j)-\sum_{s=1}^{x}F_k\right]}{1+\sum_{k=1}^{m}\exp\left[(B_n-D_i-C_j)-\sum_{s=1}^{k}F_k\right]}, \quad (4)$$

where

P_{nijk} = the probability of student n being given a rating of k on item i by rater j,

$P_{nij(k-1)}$ = the probability of student n being given a rating of k–1 on item i by rater j,

B_n = the ability of student n,

D_i = the difficulty of item i,

C_j = the severity of rater j,

F_k = the difficulty of being rated in category k rather than category k–1.

A simpler way to write this is:

$$\log(P_{nijk}/P_{nij(k-1)})=B_n-D_i-C_j-F_k. \quad (5)$$

The fit of rating scale data to the MFRM model can be examined in various ways. Many of the criteria proposed for examining the quality of ratings are based on the standardized residuals summarized as unweighted mean squares (OUTFIT) and weighted mean squares (INFIT) over the three facets in the model. A useful statistic for examining how well the latent variable has been defined is the reliability of separation index. This index provides information about how well the elements within a facet are separated in order to reliably define the facet. This index is analogous to a traditional index of reliability in the sense that it reflects the ratio of true score variance to observed score variance (in the logit metric rather than in raw scores). Also, a chi square statistic is provided for judging the statistical significance of the differences between the elements within a facet.

Comparing some issues

As described in the previous two sections, both G theory and MFRM can be applied to analyze data with multiple facets or sources of errors. At a foundational level, the two approaches may seem incompatible. For example, Brennan (2001) comments that G theory is primarily a sampling model whereas IRT (e.g., MFRM) is principally a scaling model. The comparison issues are: (a) major research questions addressed, (b) statistical model, (c) design issues, (d) methods of data collection, (e) standard results, (f) limitations of the two approaches.

Major research questions addressed. G Theory asks: How reliably can observed scores can be generalized to make inferences about person's behavior in a defined universe? But, MFRM asks: What is a person's ability estimate after adjusting for the influence of item difficulty and rater harshness? Thus, G theory primarily focuses on analyzing multiple sources of error simultaneously and comparing the relative impact of each source of error on a measure, and so, to provide a generalizability coefficient, and estimate of the number of conditions that should be selected for each facet. Alternatively, the MFRM primarily focuses on searching for the simplest model best-fitting model to allow an unbiased person estimate.

Statistical model. G Theory is a random effects ANOVA model. The MFRM is a generalized linear mixed model (De Boeck and Wilson, 2005).

Design. In G Theory, the facets may be either random or fixed, whereas in the Rasch model, usually only the person is seen as a random effect.

Data collection. In terms of collecting data, G Theory requires a formulated experimental design in order for the complete set of variance components to be estimable. MFRM does not require such a strict design, though the constraints on data collection designs are that the measures be estimable unambiguously in one frame of reference or that the relationship between disjoint observations can be specified.

Standard results. Major findings of a G theory analysis are to estimate and optimize the reliability (generalizability) of decisions based on the data collection and to design and modify an efficient measurement process and test construction. The results of a MFRM analysis include an estimate of each person's location, as well as a reliability estimate, and a map of the distribution of persons, items, and raters.

Limitations. As mentioned above, an assumption of G theory is that the observed behavior is a random sample of all behaviors, and that the underlying distributions are normal (to calculate standard errors). MFRM analysis does not require the normality assumption that underlies G theory analysis. However, Rasch analysis is based to need to link responses and requires that local independence holds.

Methodology

To better understand the structure of a writing performance assessment, two approaches, G theory and MFRM were used to analyze data. In the first part of the analysis, using G theory, we

estimate the variance component of items, raters, rating criteria and their interactions to compare the relative influence of each facet, and determine the optimal number of grading conditions of each facet that maximizes the generalizability coefficient. In the next part of the analysis, by applying the MFRM approach, we investigated the inconsistency in examinee scores caused by raters, items, and criteria, and provide a better model by reporting quality control fit statistics.

Data

The data used in this study were based upon the scoring of written compositions by 2nd year high school students for the annual teacher training program at KICE (Korea Institute of Curriculum and Evaluation) conducted in August, 2000. The sample consisted of 229 students—each student wrote two compositions. Each of these compositions was rated by two raters (teachers) on each of three scoring criteria. A pair of raters in the training session were exchanged randomly according to each item. For each of the three criterion (analytic proof, structure, grammar and vocabulary), a five-point rating scale was used (1 to 5). Each item was assigned 15 points, 5 points per each criteria, therefore, the possible total score for the written composition is 30 points. Thus, for each of the 229 students, there are four sets of data (two topics by two raters) using the three criteria. The data for this situation can be arranged as shown in Figure 1.

G Theory Analysis

The research questions for G theory related to the following general questions: Is the scoring of student's written composition generalizable across raters, items, and criteria? This general question is usually decomposed into the following research questions:

1. What are the differences in the relative magnitudes of error variance due to raters, items, criteria and the interactions between these factors?
2. Is the generalizability coefficient improved by increasing the number of conditions in each facet? If so, what is the optimal number of conditions of each facet to maximize the generalizability coefficient?

For this G study design, the object of measurement is student (p:person) and sources of error are rater (r), item (i), and criteria (c). As described above, each composition was scored by two raters assigned to each item against three rating criteria, therefore, the student (person) effect is crossed with raters, items, and criteria. In order to check the rater effects, a pair of raters were crossed with each other rating the same item of each student. However, since all raters

	Essay Item 1						Essay Item 2					
	Rater A			Rater B			Rater C			Rater D		
	C1	C2	C3	C1	C2	C3	C1	C2	C3	C1	C2	C3
student 1	3	3	4							3	2	1
student 2	3	3	5							1	3	3
student 3	4	2	4							3	3	1
student 4	4	3	2							2	2	3
student 115				2	3	4	2	2	1			
student 116				4	3	5	1	3	3			
student 117				3	4	3	1	1	1			
student 229				4	3	3	1	2	3			

Figure 1. Arrangement of the data set for Essay Items

did not score all items, raters and rating criteria are considered to be nested within each item. The conditions of each facet can be defined as a sample of a complete set of conditions (i.e., fixed effect) or as the infinite set of conditions (i.e., random effect). For this design, students were considered to be a random effect because the students were chosen from possible examinees. Raters were also treated as a random effect since raters were randomly chosen from a pool of eligible high school language/literature teachers. In addition, items and rating criteria were also considered to be random effects because items and criteria were developed from an infinite possible set of items or rating criteria.

Therefore, the generalizability study is a random, partially nested design, p x ((c x r) : i). It is conducted in the following two steps: (1) estimating the variance components of raters, items, and scoring criteria to compare the relative influence of each facet (the G study), and (2) determining the optimal numbers of grading conditions of each facet that maximize the generalizability coefficient (the D study).

The design addressing the questions includes a three-facet generalizability study with raters (r) crossed with rating criteria (c) nested within items(i), and examinees (p) crossed with the other three factors; p x ((c x r) : i) design. The variance of the observed score can be decomposed into nine variance components as follows:

$$\sigma_x^2 = \sigma_p^2 + \sigma_i^2 + \sigma_{pi}^2 + \sigma_{ci}^2 + \sigma_{ri}^2 + \sigma_{pri}^2 + \sigma_{pci}^2 + \sigma_{rci}^2 + \sigma_{prci,e}^2. \quad (6)$$

In other words, the G theory assumes that the variance of the ratings can be partitioned into independent sources of variation due to difference between students, items, criteria, raters, their interactions, and the residual. The focus of the G study is on these variance components because their magnitude provides information about the sources of error influencing a particular measurement.

The student was the object of measurement in the scoring system, therefore, the variance component for students represents the universe score variance. The relative error variance includes the variation related to the interaction between items and students, the interaction between students and raters, and the interaction between students and rating criteria and residual. A generalizability study can generate several coefficients, each corresponding to a different universe of conditions. The resulting estimated relative error variance and estimated generalizability coefficient can be expressed as follows:

$$\hat{\sigma}(\delta)^2 = \frac{\sigma_{pi}^2}{n_i} + \frac{\sigma_{pri}^2}{n_r n_i} + \frac{\sigma_{pci}^2}{n_c n_i} + \frac{\sigma_{prci}^2}{n_r n_c n_i} \quad (7)$$

$$\hat{\rho}(\delta)^2 = \frac{\sigma_{(\tau)}^2}{\sigma_{(\tau)}^2 + \sigma_{(\delta)}^2} = \frac{\sigma_p^2}{\sigma_p^2 + (\frac{\sigma_{pi}^2}{n_i} + \frac{\sigma_{pri}^2}{n_r n_i} + \frac{\sigma_{pci}^2}{n_c n_i} + \frac{\sigma_{prci}^2}{n_r n_c n_i})}. \quad (8)$$

The estimated error variance components are then compared for relative magnitudes. Each variance component contributes to several types of error variance for mean scores in a D study $p \times ((C \times R) : I)$ design and their contribution to such error variances can be reduced by increasing the D study sample size for each facet. The data were analyzed with the GENOVA program developed by Crick and Brennan (1983).

MFRM Analysis

In contrast to the G theory analysis, the standard MFRM analysis primarily investigates the individual performances. The general research question of the MFRM might be: What are the major effects of each facet? This question implies further possible questions such as: Are the raters behaving consistently? How does a student's performance differ from the first item to the second item? etc.

The MFRM specifies that the data patterns result from independent contributions of each student, item, criterion, as well as their interactions. Inevitable random measurement error is also incorporated explicitly into the model by the probability terms. Since the intention of this study is to investigate the interaction effect of a

measure and compare the results with those of applying G theory, the interaction terms are also parameterized.

The data analyzed here are three-faceted data, with item, rater, and criterion facets. The model allows for the raters to differ in harshness, the criteria to differ in difficulty, and the rating structure to vary across the criteria. The model includes a main effect for all three facets. The model also includes the set of three two-way interactions, and a step structure. The model, therefore, contains seven terms: item, rater, criteria, rater*item, rater*criteria, item*criteria and item*rater*criteria*step. The first three are main effects and correspond to a set of item difficulty parameters, a set of rater severity parameters, and a set of criteria difficulty parameters. The next three are two-way interactions between the facets. The first of the interactions allows us to model a variation in rater harshness across items (or variation in item difficulty across the raters); the second interaction models a variation in rater harshness across the criteria, and the third interaction models variation in the item difficulty across the criteria. The final term represents the step structure of the responses. The step structure is modeled as varying across all combinations of items, raters, criteria:

$$\log(P_{nijk}/P_{nij(k-1)}) = B_n - D_i - C_j - E_c + DC_{ij} + EC_{cj} + DE_{ic} - F_k, \quad (9)$$

where

P_{nijk} = the probability of student n being given a rating of k on task i by rater j,

$P_{nij(k-1)}$ = the probability of student n being given a rating of k–1 on task i by rater j,

B_n = the ability of student n,

D_i = the difficulty of item i,

C_j = the severity of rater j,

E_c = the difficulty of rating criterion c,

DC_{ij} = the interaction between item i and rater j,

EC_{cj} = the interaction between rater j and criterion c,

DE_{ic} = the interaction between item i and criterion c, and,

F_k = the difficulty of being rated in category k rather than category k–1.

The model described above provides a frame of reference for understanding the relationship of the facets of the performance assessment. It makes it possible to observe estimated student ability from the highest to lowest, estimated task difficulty from most to least difficult, and estimated judge severity from most to least severe. Here the analyses will particularly focus on estimating the parameters for each two-way interaction. Also, a chi-square test for the model fit is conducted.

Results

G study

Table 1 presents the results of the G study including the estimated variance components and proportions. The variance component for universe scores, student score (σ^2_p), was relatively small (.1887, 11.4%). However, the variation due to items (σ^2_{pi}) was large (.487, 29.4%) relative to the variation due to raters (σ^2_{pri}; .015, 0.9%). This shows that the performance difference between items is much greater than that between raters. In other words, the generalizability of scorings is substantially influenced by item difficulty. The magnitude of variance of the interaction between student and rater (σ^2_{pri}; .159, 9.6%) indicates that even though the raters used the same standards overall, they disagreed somewhat in the relative standing of students. Also, the variation due to interaction between student and criteria within item (σ^2_{pci}) is relatively very large (.496, 29.9%), This indicates that the relative standing of students differed from one criterion to another.

We can investigate whether the generalizability coefficient can be improved by trade-offs between the number of conditions of each facet. For example, increasing the number of rating criteria might be more effective than increasing the number of raters. The results also showed that the effect of raters was less than that of item or scoring criteria (although the interactions were large), therefore, one way to improve the generalizability coefficient would probably be based on having different combinations of fewer raters and more items.

Table 2 presents the results of the D study. It shows the changes in each estimated error variance and generalizability coefficient that results from changing the number of items, raters and criteria. Increasing the number of items or criteria produced a better generalizability coefficient more quickly than increasing the number of raters. The G coefficient increased considerably as the number of criteria increased. A level of generalizability of about .70 was obtained with each of the following combinations: 3 items, 2 raters, 5 criteria; 4 items, 3 raters, 3 criteria; 4 items, 4 raters, 5 criteria.

MFRM

As described above, the model contained seven sets of parameters: rater, item, criteria, interactions of rater and item, rater and criteria, item and criteria, and rater by item by criteria by

Table 1

Results of G study p × ((c × r) : i), 3 facets design

Effect	Degrees of Freedom	Sums of Squares for Score Effects	Mean Squares	*F* Statistic	Variance Component	Proportion (%)
P	228	914.75	4.01	2.34	.1887	11.39
I	1	695.02	695.02		.4866	29.38
R:I	2	23.50	11.75	6.57	.0145	0.87
C:I	4	63.69	15.92	6.92	.0297	1.79
PI	228	390.64	1.71	.98	(0.0)	0.0
PR:I	456	343.83	.75	2.83	.1590	9.62
PC:I	912	1157.97	1.27	4.59	.4964	29.97
RC:I	4	5.23	1.31	4.73	.0045	0.27
PRC:I	912	252.43	.28		.2768	16.7
TOTAL	2747		3847.09			

Note: n_p:229, n_i:2, n_r:2, n_c:3

Table 2

Summary of D Study results for ((R x C): I) x p

D Study Design No	Sample Sizes $P Inf.	I Inf.	R Inf.	C Inf.	Variances Expected Universe Score	Relative Observed Score	Absolute Error Var	Error Var	Mean	Gen. Coef.	Phi Coef.
1	229	1	1	1	.1887	1.1211	.9323	1.4667	.5403	.1684	.1139
2	229	1	2	7	.1887	.3589	.1702	.6687	.5000	.5258	.2201
3	229	2	1	3	.1887	.3972	.2084	.4647	.2580	.4752	.2889
4	229	2	3	7	.1887	.2573	.0686	.3165	.2491	.7335	.3736
5	229	3	2	3	.1887	.2858	.0970	.2652	.1694	.6604	.4158
6	229	3	2	5	.1887	.2576	.0688	.2356	.1679	.7328	.4448
7	229	4	3	3	.1887	.2511	.0623	.1878	.1266	.7518	.5013
8	229	4	4	5	.1887	.2269	.0382	.1623	.1251	.8316	.5376
9	229	4	4	7	.1887	.2189	.0301	.1538	.1246	.8623	.5510
10	229	5	3	3	.1887	.2386	.0498	.1502	.1014	.7911	.5568
11	229	5	4	5	.1887	.2153	.0306	.1299	.1002	.8606	.5924

step. Each parameter and its significance were estimated by the ConQuest program (Wu, Adams and Wilson, 1998) and the results are shown in Table 3. The results include the main effects (rater severity, item difficulty and criteria difficulty), and the interaction effects between the facets (rater harshness across the items (or variation in item difficulty across the raters), variation in rater severity across the criteria, and variation in the item difficulty across the criteria). The final term, step structure, across all combinations of raters, items, and criteria, is not shown the Table.

Notice that, consistent with the G theory results, the overall differences between the rater parameter estimates are not significant. The rater severities ranges from –0.074 (rater 3) to 0.143 (rater 4). On the other hand, the item parameter estimates are significantly different, indicating item 2 (.506) is much more difficult than item 1 (–.506). The overall differences between criteria are also significant. The difficulty level of criteria range from 0.110 logits. for criteria 3 to –0.190 logits for criterion 1. Consistently with the G study results, the rater by item effects were significant.

Table 3

Results of the Parameter estimates
*(rater +item + criteria + rater * item + rater * criteria + item * criteria + rater * item * criteria * step)*

Effect		Estimate	Error*	Weighted Fit* MS	*t*-value	chi-square	*p*
Rater	1	–0.052	0.058	1.69	6.2	2.525	0.471
	2	–0.017	0.041	1.25	2.5		
	3	–0.074	0.060	0.93	–0.7		
	4	0.143					
Item	1	–0.506	0.032	1.94	8.0		
	2	0.506					
Criteria	1	–0.190	0.050	1.86	7.5	17.332	0.000
	2	0.080	0.046	1.57	5.3		
	3	0.110					
Rater * item	1 1	0.022	0.072	0.74	–3.0	37.593	0.000
	2 1	0.325	0.063	1.02	0.2		
	3 1	–0.277	0.082	0.93	–0.7		
	4 1	–0.070					
	1 2	–0.022					
	2 2	–0.325					
	3 2	0.277					
	4 2	0.070					
Rater *criteria	1 1	–0.107	0.086	0.79	–2.4	33.393	0.000
	2 1	0.253	0.072	0.92	–0.9		
	3 1	–0.233	0.105	0.34	–9.6		
	4 1	0.087					
	1 2	0.036	0.076	0.85	–1.6		
	2 2	–0.224	0.067	1.14	1.5		
	3 2	0.159	0.088	0.37	–8.8		
	4 2	0.029					
	1 3	0.071					
	2 3	–0.029					
	3 3	0.074					
	4 3	–0.116					
Item *criteria	1 1	–0.099	0.04	1.22	2.3	27.228	0.000
	2 1	0.099	6.00				
	1 2	0.216		1.43	4.1		
	2 2	–0.216	0.04				
	1 3	–0.117	5.00				
	1 3	0.117					

*Note: that the values missing in the "Error" and "Fit" columns are for estimates that are constrained by the model.

In addition, the results show some significant interactions between raters and items, and raters and criteria. In other words, there are variations in item difficulty across the raters and in rater severity across the criteria. Based on the estimated logits of the rater*item interaction and rater*criteria interaction, rater 2 and rater 3 tend to rate the items differently and use criteria 1 and 2 with different standards. On the other hand, there is a little variation in the item difficulties across the criteria. In other words, there is some evidence that raters are using the criteria in an inconsistent fashion even though the rater severities are not significantly different from each other and there is little variation in the item difficulties across the criteria.

The fit statistics for many of the raters and criteria are larger than desired by a substantial amount. The fit of the criteria parameters are poor, suggesting the need to allow further interactions. This indicates that it might be that we should explore a model involving a parameter for every rater*item*criteria combination.

The MFRM analysis permitted the comparison of the three different facets on a common logits scale. The results of the parameter estimates of each effect and interactions are visually summarized in Figure 2. The Wright map illustrates how the logits of each main effect and their interaction effects are spread out relative to one another. The effects mentioned in the previous paragraph can be seen in this map. Although there are no statistically significant differences in the severities of the raters, item 2 is much more difficult than item 1. Also, there is noticeable variation in using the rating criteria. Based on the results concerning the interactions between rater and item, and rater and criteria, the plots show rater 2 and rater 3 tend to rate the items differently and use criteria 1 and 2 with different standards. That is the part of reason why the interaction effects are significant.

The MML reliability of the model (Mislevy et al., 1992) turned out 0.74 which is comparable with the findings for the generalizability coefficient.

Comparing the results

A comparison table of the results of the two analyses is given in Table 4. Since the G study is a partially nested design, the effect of the two approaches are expressed somewhat differently. Both the G theory and MFRM approaches showed a significant difference in difficulty between the two items. Also, both found that the rater severity main effect is not significant. The MFRM map (Figure 2) suggests that the raters were relatively homogenous, as does the G theory result of only .87% of the variance being attributable to rater difference. However, G theory results found a nonnegligible variance component for the interaction between student and raters (9.62%); this suggests that the relative standing of students differed somewhat from one rater to another. The variance component for rater-criteria interaction is small; raters were well calibrated in terms of the differences between criteria across rater. However, the interaction between students and criteria within items is large (29.97 %). Even though raters used the same standards overall, relative performance of students differ between criteria. The MFRM results provide the individual parameter estimates of each effect and each interaction, therefore, it informed how each rater rated each student's composition according to each item based on each criterion. The results indicated that each of the interaction effects in this model is statistically significant. G theory results only provide the relative influence of each source of error on the score, on the other hand, MFRM results show how each item was rated differently by each rater based on each criterion.

Discussion of Results

Performance assessment requires a rater to evaluate the quality of an examinee's performance on a task or item. The rater's evaluation is expressed as qualitatively based ratings on a rating scale—the numerical values assigned to the qualitative ratings obtained by each examinee's performance level for decision-makers. But the observed score is complicated, to a greater or lesser extent, by variation in rater severity and

item difficulty, and also by inconsistencies of raters over criteria and items. If people use different tasks and different raters and different occasions for the scores, there are many distinguishable influences on the score, such as occasion, task, rater, and their interactions. A researcher has to assess whether the performance of a student generalizes broadly over rater, tasks, and occasions. The G theory and MFRM methods in terms of the data collection designs are slightly different to each

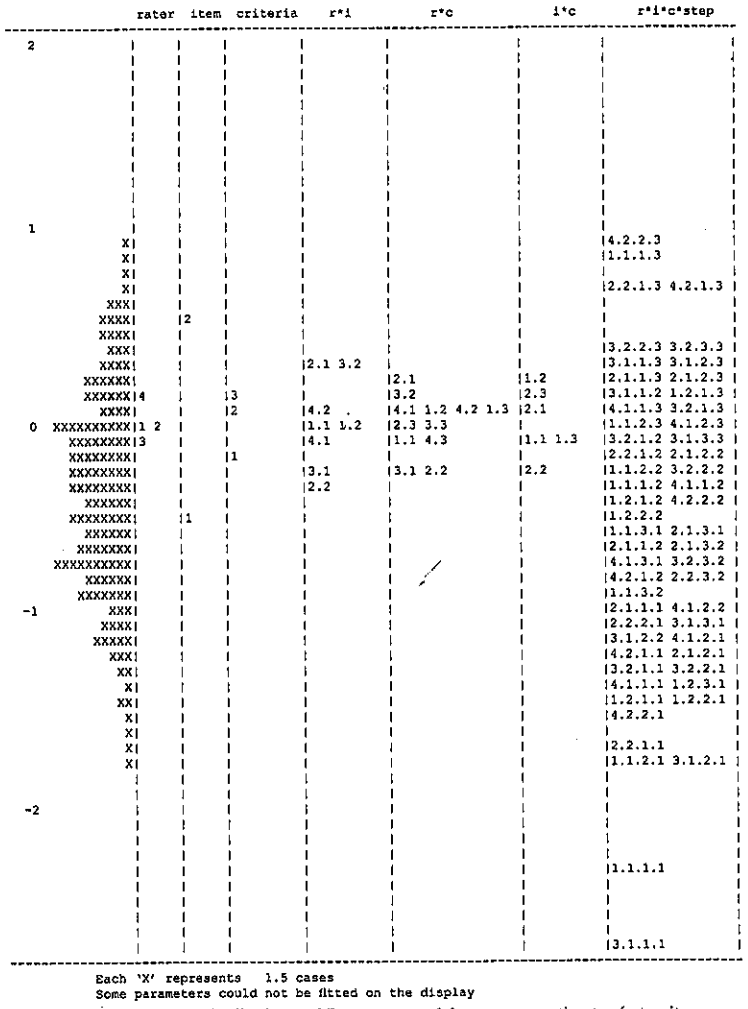

Figure 2. Map of latent Distribution and Response model parameter estimates (rater, item, criteria, rater * item, rater * criteria, item * criteria, rater * criteria * step)

other. In G theory, for example, when the different raters within one occasion or within one item scored all students' performances, the G study design called the raters "nested within occasions or items." However, in the MFRM, it is only possible to model and to estimate parameters when the rater and items are linked. Unlike G theory, however, MFRM does not require a precisely formulated experimental design; in G theory, whether the data layouts are crossed or nested and whether each effect is random or fixed is crucially important. The constraints on MFRM data collection designs are only that there are observations to link between disjointed subsets of observations. There are other considerations in the MFRM framework, however, and mis-specifications in the design can lead to over-estimation of the reliability (Wilson and Hoskens, 2001).

The results of both the G theory and MFRM analyses showed that the scoring of written compositions in the annual teacher training program turned out consistently from rater to rater across items despite variation in item difficulty and rater severity. However, the relatively large variance components of the item-criteria interaction indicated that items were rated differently across criteria. The MFRM analysis results also showed that the parameter estimates of item difficulties and the interaction of item and criteria are somewhat different. This finding suggests that, if a student's written composition is being assessed, generalizability of scores was enhanced by combining ratings from more than one rater. Some studies (Shavelson, Mayberry. and Webb, 1990; Kim, 1996) also found that interrater reliability is not a problem, but that task-sampling variability can be important.

The G study presented the relative importance of raters, items, rating criteria, and their interactions in estimating the dependability of scores. The results of the G theory approach showed relatively large error variances for the interaction between student and rating criteria within item, and the interaction between student and rater within item. Based on the relative size of each error variance, the study examined possible combinations of the conditions of each facet in order to determine the number of raters, items, rating criteria that are needed to obtain an acceptable level generalizability for a measure. For these ratings of written composition, the generalizability coefficient increased considerably as the number of items increased. Also, because the raters within item component and rater-criteria interaction variance components were small, increasing the number of raters could have little effect improving the G coefficient.

Both generalizability theory and the many-facet Rasch model can be useful for producing more reliable performance assessment measures.

Table 4

Comparison of Effect Differences

	G theory results			MFRM results	
Effect	VC	(%)	Effect	chi-square	logit range
Student	.1887	(11.39%)	NA	NA	NA
Rater	.0145	(.87%)	rater	2.526	−.074~.143
Item	.4866	(29.38%)	item		−.506~.506
student*item	0		NA	NA	NA
Criteria	.297	(1.79%)	criteria	17.332*	−.190~.110
student*rater:item	.1591	(9.62%)	rater*item	37.593*	−.325~.325
rater*criteria:item	.0045	(.27%)	rater*criteria	33.393*	−.116~.253
student*criteria:item	.4964	(29.97%)	item*criteria	27.228*	−.216~.216
	NA	NA	item*rater* criteria*step		−2.824~.967
Error	.2768	(16.7%)	NA	NA	NA

* indicates significant difference at .001 level

However, they have relative strengths and weaknesses in producing more reliable scores. On the one hand, G theory is useful in detecting various error components simultaneously in a complicated rating system. A major contribution of G theory is that it allows the researcher to estimate the influence of multiple sources of measurement error and to increase the number of conditions of each facet so that the variations decrease to an acceptable level. In other words, the researcher can compare the relative influence of each facet on a measure of the target assessment and estimate how many conditions of each facet are needed to attain a certain level of generalizability.

On the other hand, MFRM analysis makes it possible to investigate individual scores after controlling the facets. MFRM allows one to investigate whether there are significant differences in rater severity, item and criteria difficulty. The MFRM is only one member of a large family of item response models that can be used to comprehensively investigate reliability and validity issues (Wilson, 2005). G theory is a comprehensive method for designing, assessing, and improving the dependability of measurement procedures. G analysis is also useful for determining what modifications can or should be made to a measurement procedure. In response to the criticism about the limitation of G theory, Marcoulides and Drezner (1997) introduced an extension to G theory that can be used to provide specific information at the individual ability estimates. The issue of fit of such models is still to be investigated.

The criteria for choosing between the two analytic approaches are clear. If it is important to estimate the similarity between the observed raw scores of the group of students and the raw scores that similar groups of students might obtain under identical circumstances, G theory may be helpful. If it is important to estimate for each student a measure as free as possible of the particularities of the facets that generated the raw score, then MFRM is highly desirable. Another problem of the performance assessment data analyzed was that different sets of raters scored different groups of examinees in a real-life, large, sparse data set. The variation due to different groups of examinees may be explained by variation due to different groups of raters. The above discussion considered sources of error in scoring compositions, and was based on the assumption that a good grading system should consist of developing a scoring guide and specific rater training process. When interpreting student results, the analysis method used should be acknowledged, because it has a significant impact on the conclusions that may be drawn about the quality of a student's performance.

References

Baxter, G. P., Shavelson, R. J., Goldman, S. R., and Pine, J. (1992). Evaluation of procedure-based scoring for hands-on science assessment. *Journal of Educational Measurement, 29*, 1-17.

Brennan, R. (1992). *Elements of generalizability theory* (rev. ed.) Iowa City, IA: American College Testing.

Brennan, R. (2001). *Generalizability theory.* New York: Springer-Verlag.

Crick, J. E., and Brennan, R. L. (1983). *Manual for GENOVA: A generalized analysis of variance system.* (Technical Bulletin No. 43). Iowa City, IA: American College Testing.

De Boeck, P., and Wilson, M. (Eds.). (2004). *Explanatory item response models: A generalized linear and nonlinear approach.* New York: Springer-Verlag.

Du, M., and Wright, B. (1997). Effects of item characteristics in a large-scale direct writing assessment. In M. Wilson, K. Draney, and G. Engelhard, Jr. (Eds.), *Objective measurement: Theory into practice* (Vol. 4, pp. 1-24), Norwood, NJ: Ablex.

Engelhard, G. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement, 31*(2), 93-112.

Fischer, G. H. (1983). Logistic latent trait models with linear constraints. *Psychometrika, 48*, 3-26.

Kim, S. (1996, April). *Optimizing the generalizability in scoring essay items for the college entrance examination.* Paper presented at the annual meeting of the American Educational Research Association, New York.

Kim, S. (2000, April). *Investigating the generalizability of scores from different rating system in performance assessment.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans,

Lane, S., Liu, M., Ankenmann, R. D., and Stone, C. A. (1996). Generalizability and validity of a mathematics performance assessment. *Journal of Educational Measurement, 33*(1), 71-92.

Lehmann, R. H. (1990). Evaluation studies: Reliability and generalizability of ratings of compositions. *Studies in Educational Evaluation, 16*, 501-512.

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

Linacre, J. M. (1995). Generalizability theory and many-facet Rasch measurement. In M. Wilson, K. Draney, and G. Engelhard, Jr. (Eds.), *Objective measurement: Theory into practice* (pp. 85-98). Norwood, NJ: Ablex.

Lunz, M. E., and Schumacker, R. E. (1997). Scoring and analysis of performance examinations: A comparison of methods and interpretations. *Journal of Outcome Measurement, 1*(3), 219-238.

MacMillan, P. D. (2000). Classical, generalizability, and multifaceted Rasch detection of interrater variability in large, sparse data sets. *The Journal of Experimental Education, 68*(2), 167-190.

Marcoulides, G. A. (1997). A method for analyzing performance assessments. In M. Wilson, K. Draney, and G. Engelhard, Jr. (Eds.), *Objective measurement: Theory into practice* (pp. 261-277). Norwood, NJ: Abex.

Marcoulides, G. A. (1999). Generalizability theory: Picking up where the Rasch IRT model leaves off? In S. E. Embretson and S. L. Hershberger (Eds.), *The new rules of measurement: what every psychologist and educator should know* (pp. 129-152). Mahwah, NJ: Lawrence Erlbaum.

Mislevy, R. J., Beaton, A., Kaplan, B., and Shachan, K. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement, 29*(2), 133-161.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen, Denmark: Danish Institute for Educational Research. (Expanded edition, 1980. Chicago: University of Chicago Press.)

Shavelson, R., and Webb, N. M. (1991). *Generalizability theory: A primer.* Newbury Park, CA: Sage.

Stahl, J. A. (1994). What does generalizability theory offer that many-facet Rasch measurement cannot duplicate? *Rasch Measurement, 8*(1), 342-343.

Wilson, M., and Hoskens, M. (2001). The rater bundle model. *Journal of Educational and Behavioral Statistics, 26*(3), 283-306.

Wilson, M. (2005). *Constructing measures: An item response modeling approach.* Mahwah, NJ: Lawrence Erlbaum.

Wright, B., and Masters, G. (1982). *Rating scale analysis.* Chicago: MESA Press.

Wu, M., Adams, R., and Wilson, M. (1998). ConQuest: Generalized item response modeling software [Computer program]. Camberwell, VIC, Australia: ACER Press.