Improving Creativity Performance Assessment: A Rater Effect Examination with Many Facet Rasch Model

Su-Pin Hung, Po-Hsi Chen & Hsueh-Chih Chen

National Taiwan Normal University

Published online: 16 Nov 2012.

To cite this article: Su-Pin Hung, Po-Hsi Chen & Hsueh-Chih Chen (2012) Improving Creativity Performance Assessment: A Rater Effect Examination with Many Facet Rasch Model, Creativity Research Journal, 24:4, 345-357, DOI: 10.1080/10400419.2012.730331

To link to this article: http://dx.doi.org/10.1080/10400419.2012.730331
Improving Creativity Performance Assessment: A Rater Effect Examination with Many Facet Rasch Model

Su-Pin Hung, Po-Hsi Chen, and Hsueh-Chih Chen
National Taiwan Normal University

Product assessment is widely applied in creative studies, typically as an important dependent measure. Within this context, this study had 2 purposes. First, the focus of this research was on methods for investigating possible rater effects, an issue that has not received a great deal of attention in past creativity studies. Second, the substantive question of whether restrictions on materials used and differences in instructions provided would influence outcomes on measures of creativity was considered. The many-facet Rasch model was used to investigate possible sources of rater bias, including the leniency/severity effect, central tendency effect, halo effect and randomness effect. No indications were found that these potential sources of bias strongly influenced the ratings. The result indicated that the examinees could be reliably differentiated in terms of their performance. Analysis of rater-criterion interactions depicted rater behavior more clearly and, it is suggested, can be of use as a tool for rater training in future studies. In terms of the substantive questions posed, 2 × 2 experimental instructions were manipulated and it was found that different instructions did not affect creative performance. The implications of these findings are discussed.

A number of studies, both experimental and non-experimental, have sought to identify factors which influence creative performance. The research reported in this article examined the role of instructions to participants and task constraints in the form of restrictions on the stimuli objects assessed in creative performances. Additionally, the focus of this research was on an area which has not received a great deal of attention in the field of creativity research: the evaluation of the quality of product and performance ratings. Rater reliability often used to assure rating quality in performance assessment however, different reliability indices may have different meanings. Furthermore, methods of calculation chosen may influence how to calculate the summary scores for participants and the subsequent analysis. These issues are addressed in this study. In particular, the many-facet Rasch model (MFRM) for this purpose is introduced and illustrated. After demonstrating the use of this model using the ratings produced in our own study, the authors move on to a consideration of the substantive questions posed.

STIMULI SETS AND TASK INSTRUCTIONS

Research focused on restricting the stimuli used for creative performances has produced conflicting results. In a series of studies, Finke (1990; Finke, Ward & Smith, 1992) classified three categories of stimuli sets and compared the creative performance of restricted and non-restricted groups. The restricted group was required to select one object from three categories, so that the resulting product combined at least categories. In the non-restricted group, examinees were allowed to choose objects from any category. The restricted group was found to outperform the non-restricted group and it was concluded that more highly constrained conditions result in a greater number of creative inventions, suggesting that...
restricting stimuli elements and components helps to stimulate creative thinking. However, the variability of chosen categories may have confounded the results. Other studies, however, have shown that participants allowed more flexibility in choosing task resources were rated as more creative (Amabile & Gitomer, 1984; Greenberg, 1992; Iyengar & Chua, 2008). Amabile (1996) reviewed a series of studies on the topic and concluded that restricted choice or constraint control may impede creativity and that people who feel a sense of control and have access to sufficient resources may experience enhanced creative abilities. Again, the results agree with the authors’ inference that the variability of chosen categories positively influences creative performance. Consequently, it is reasonable to question the inconsistent conclusions of the previous studies via a simplified experimental condition (i.e., by holding constant the variable of chosen variety). Investigating the effect of task constraints can be expected to provide substantial assistance to practitioners in preparing task environment to encourage creative performance. Accordingly, one purpose of this research is to ask whether creative performance is hindered or enhanced by task constraints. Finke (1990) proposed the creative mental synthesis task which was used to connect imagery and creativity (e.g., Barquero & Logie, 1999; Eardley, & Pring, 2007), and the outcomes can be interpreted as new ideas for creative inventions (Finke et al., 1992). In this procedure, participants were asked to choose three three-dimensional basic object parts from 15 made available to them and to use these objects to invent a new design. Finke (1990, p. 60) noted that there was nothing inherently important in the particular objects used and suggested that the 36 geons proposed by Biederman (1987) might work equally well. However, to date, neither Finke nor other studies have demonstrated that stimuli sets are, in fact, interchangeable. Consequently, it is necessary to clarify whether a wider range of choices will contribute to performance. This issue is addressed in this study.

A second issue is the influence of instructions on task performance. Some research has shown that the way in which instructions are presented (e.g., via video, audio or textual channels) may influence performance (Runco & Pezdek, 1984). A second question posed in this study, then, is whether differences in how the task is presented and in the instructions given to participants influence task performance. Hence, the first purpose of this study is to examine how the two factors (task constraints and types of instructions) will influence creative performance on a creative invention task when using different stimuli sets.

EVALUATING RATER EFFECTS

Two methodological problems face any investigator in the area of creativity research: the choice of rating criteria and the threat to validity posed by rater bias. Considerable attention has been focused on the first issue, resulting in near consensus in the field. It is now widely accepted that to assess a creative product both the novelty and appropriateness of the assessed product should be included in the rating criteria (see the review of Besemer & Treffinger, 1981; Amabile, 1996). Plucker and Renzulli (1999) stated that techniques for evaluating creative products could be sorted by degree of complexity. According to their classification, the most straightforward way is by rating scale (e.g., Creative Product Semantic Scale, CPSS; Besemer & O’Quin, 1986; Scale of Creative Culinary Products; Horng & Lin, 2009). Each of these scales is associated with evidence of reliability and validity (Besemer, 1998; Besemer & O’Quin, 1999; Horng & Lin, 2009; O’Quin & Besemer, 1989). In addition, some researchers provide raters with rating categories that are usually derived from previous research and suitable for the particular task being assessed (e.g., the evaluation of design tasks, Demirkan & Hasirci, 2009; the evaluation of works of art, Kozbelt & Serafin, 2009).

However, the problem of rater effects or rater bias has received far less attention. All techniques for analyzing creative products require judges to evaluate the products created by the participants. This necessity introduces threats to the validity of the measures used, in the form construct-irrelevant variance that is associated with characteristics of the raters and their performance, and not with the performance of the participants. This variance threatens the validity of inferences made about the performances themselves and, because these ratings typically serve as dependent measures in research, therefore threatens the soundness of any findings. Appropriate rating circumstances and the use of multiple raters may help to reduce these validity threats. Nonetheless, neither perfect rating circumstance nor the use of multiple raters can guarantee that systematic error (e.g., leniency or severity error) will not be present in the results. Thus, even when applying a complicated procedure (e.g., the consensual assessment technique, CAT, Amabile, 1983, 1996) designed to minimize rating error, it is still essential to investigate whether the results were significantly distorted by rater bias.

The common standard for evaluating rating quality has been intrarater reliability, on the assumption that the shared understanding of the construct implied by high intrarater reliability is an indication that the criteria have been appropriately applied during the evaluations. The choice of calculation methods depends on the purpose for which the data are being collected and the philosophy underlying the rating process (i.e., whether raters are being treated as “rating machines” or “independent experts”). Stemler (2004) has offered a useful three-fold categorization of different methods of
calculating interrater reliability, including methods that previous creative products analysis studies have used.

1. Consensus estimates: The main purpose of consensus estimates is to demonstrate exact agreement among independent judges. Such estimates are generally appropriate for situations in which raters are treated as rating machines, expected to be in near-perfect agreement. Statistics such as percent agreement and Cohen’s kappa can be used to estimate consensus among judges. The advantages of the consensus approach are that interrater reliability is easy to compute, the techniques are able to deal with nominal variables and they can be useful in diagnosing problems with judges’ interpretations of how to apply the rating scale. Furthermore, if the ratings achieve high exact agreement, the summary score can be calculated by simply taking the score from one of the judges or by averaging the scores. However, one disadvantage is that to come to exact agreement, it is often necessary to spend considerable time and effort in training the judges. Further, training judges to the point of exact agreement actually reduces the statistical independence of the ratings, and may thus threaten the validity of the resulting scores (Linacre, 2002). Previous creative product analysis studies have typically only applied consensus estimates when the construct under investigation has some objective meaning (e.g., Ward, Patterson & Sifonis, 2004).

2. Consistency estimates: The main purpose of consistency estimates is to get judges to consistently apply a scoring rubric. Because it is not necessary for two judges to share a common meaning of the rating scale, this approach may be more appropriate when raters are treated as independent experts and the intent is to determine whether raters are consistently applying the rubric. The most frequently used indicators of consistency include the Pearson correlation coefficient, Spearman rank coefficient and Cronbach’s coefficient alpha. The major advantage to using consistency estimates of interrater reliability is that they place fewer demands in terms of rater training, the method is easy to compute, and it can effectively handle ratings from multiple raters by simultaneously computing estimates across all the task or items that were rated. However, the computations involved are more complicated than for the other approaches, and some methods rely on particular software packages (e.g., ConQuest or Facets for the MFRM). Only a few creativity studies have applied this approach (e.g., Generalizability theory, and the many facet Rasch model (MFRM). This approach comes with a number of advantages, including that it can take into account errors at the level of each judge, it does not require all judges to rate all tasks to arrive at an estimate of interrater reliability and it can effectively handle ratings from multiple raters by simultaneously computing estimates across all the task or items that were rated. However, the computations involved are more complicated than for the other approaches, and some methods rely on particular software packages (e.g., ConQuest or Facets for the MFRM). Only a few creativity studies have applied this approach (e.g., Generalizability theory was applied in Amabile, 1986, using a formula proposed by Winer, 1971, p. 288; to analyze between and within variance in judge ratings; Liu, Cheng and Wang, 2010, used the MFRM).

There are two problems with use of interrater reliability indices, however. First, although reliability is a requirement for valid ratings, it is not a guarantee that valid ratings have been achieved. Interrater consistency is a necessary but not a sufficient condition for valid ratings (Bond & Fox, 2007). Second, with the profusion of interrater reliability indices, different indices may have very different meanings. Of the techniques mentioned, consistency estimates have been used most often by creativity researchers (Amabile, 1982, 1983, 1996; Baer, Kaufman, & Gentile, 2004; Besemer, 1998; Besemer & O’Quin, 1999; Demirkan & Hasirci, 2009; Horng & Lin, 2009; Kaufman, Baer, Cole & Sexton, 2008; Kozbelt & Serafin, 2009; O’Quin & Besemer, 1989, 2006). However, there are a number of unresolved issues associated with consistency estimates (e.g., some judges tend to use a specific subset of categories and other judges tend to use most or all of the categories; the uncertain influence of different numbers of categories—such as a 4-point vs. a 7-point scale).

Furthermore, although consistency estimates makes fewer demands in terms of rater training, the method of calculation is highly sensitive to the distribution of
the observed data. In addition, even though some consistency indices are able to measure consistency across multiple raters, they still require every judge to rate every examinee, thus limiting their application. For example, according to the definition of Cronbach’s alpha, the most widely used calculation method, the variance of all products rated by each rater and the total variance of all products rated by all raters are treated as factors affecting interrater reliability. However, there are at least three sources (e.g., examinee ability, task difficulty and the ratings of the raters) which will influence the variance of each examinee’s score. Besides, the interaction between raters and examinees or raters and criteria may affect the reliability of rating. All of this suggests that more facets need to be considered in rating.

In some creativity studies (Cheng, Wang, Liu, & Chen, 2010; Friedrich & Mumford, 2009; Rastogi & Sharma, 2010; Yuan & Zhou, 2008), ratings are combined to calculate a summary score for each participant. In such cases, the method of calculation chosen may influence the results. When the consensus approach is applied, for example, the summary scores may be calculated by simply taking the score from one of the judges or by averaging the scores across judges because high inter-rater reliability indicates that the judges agree about how to apply the rating scale. In this case, the summary score needs no extra adjustment. However, when the consistency approach is applied, researcher need to be aware of whether systematic difference occurred in the rating process (e.g., if Judge A consistently gave ratings that were two points lower than those of Judge B, then adding two points to each of Judge A’s ratings would provide an equitable adjustment to the raw score). Similarly, if most of the ratings fall into a small number of categories, the correlation coefficient will be deflated due to the restricted variability and the ratings may come to a high degree of exact agreement. As Stemler (2004) pointed out, if a researcher reported a high consistency and low consensus estimate of interrater reliability, it would be necessary to take some steps to correct for the discrepancy when summarizing the data. Otherwise, systematic bias could be introduced into the results, possibly invalidating inferences for subsequent analyses. Nonetheless, few researchers have reported whether systematic differences exist in their rating data or reported any adjustment for the summary scores. Moreover, a mean score adjustment may not effectively correct for cases in which two judges differ on the variability of their scores (e.g., where Judge A’s ratings all fall into one or two categories and Judge B’s cover a wider range).

Unlike consistency estimates, measurement estimates preserve more information on judges’ ratings and effectively handle ratings from multiple judges by simultaneously computing estimates across all rated items. Consequently, summary scores generated from measurement estimates tend to more accurately represent the underlying construct of interest, compared to the summary scores generated from the simple raw score ratings of the judges. However, given the strong assumptions of the PCA method (e.g., the assumption that ratings were assigned without error by the judges) and the restrictions associated with generalizability theory, which was developed for use in contexts in which only estimates of population parameters are of interest to researchers, the MFRM may be the more feasible choice in many settings.

The analysis of rating data with the MFRM does not require all judges to rate all items to arrive at an estimate of interrater reliability. It is still possible to directly compare all judges, as long as there is sufficient ‘connect-edness’ across the judges and the ratings (i.e., there is overlap between the examinees rated by Judges A and B). In addition, certain software programs which implement the MFRM (e.g., Facets, ConQuest) automatically calculate a measure of the severity of each rater as well as the degree of exact agreement between judges.

Additionally, the MFRM can be used to evaluate a number of specific hypotheses related to potential sources of rater bias that represent threats to the validity of ratings. Aiken and Groth-Marnat (2006) provided a summary of the various types of errors that may bias ratings, including constant errors, the halo effect, contrast errors and the proximity error. They also discuss various factors that may cause these errors, including observational ability, amount of experience, rater personality, interpersonal skills, perceptiveness, and freedom from judgment bias.

As noted, creativity research usually requires that raters use well-developed instruments or criteria to evaluate creative products or inventions. Nonetheless, even when well-established rubrics are consistently applied by raters, the severity of different raters may affect the ratings. Suppose two judges are asked to grade student products. Even if the correlation between these two judges shows high consistency, there may be large discrepancies in the ratings due to differences in rater severity. For example, it may be that one judge always gives very tough ratings, and the other is always very generous. Such discrepant behaviors have been shown to affect scores (Wang & Cheng, 2001; Wu, Adams, & Wilson, 1998). Though an ANOVA-based approach can help to demonstrate the interaction between judges and examinees, the MFRM approach goes beyond an ANOVA-based approach in dealing with individual-level rater effects (Myford & Wolfe, 2004a).

THE MANY-FACET RASCH MODEL

The MFRM is an extension of the two facets model proposed by Rasch (1960). Rasch introduced his two facets
model based on a probabilistic relation between person ability and item difficulty. This mathematical model can be formulated as follows:

\[
\log\left(\frac{P_{ni}}{P_{n0}}\right) = B_n - D_i \ldots
\]

where \(B_n\) is the latent trait of examinee \(n\), \(D_i\) is the difficulty of item \(i\), and \(P_{ni}\) and \(P_{n0}\) are the probabilities of examinee \(n\) on item \(i\) scoring 1 and 0, respectively. This formula allows for the simultaneous estimation of examinee ability and item or task difficulty. The original Rasch model has been expanded to allow for polytomous cases (Andrich, 1978; Masters, 1982). In a further extension of the model, Linacre (1989) recognized that factors such as raters and the criteria used could also influence assessment outcomes. In other words, in addition to the examinee and item facets, certain assessment contexts require that other facets (raters, criteria) be considered. Thus, the Rasch model was further extended into the MFRM which is able to accommodate such situations. A typical facets model for a three-facet situation can be formulated as follows:

\[
\log\left(\frac{P_{nijk}}{P_{ni(k-1)}}\right) = B_n - D_i - C_j - F_k
\]

where \(B_n\) is the latent trait of examinee \(n\), \(D_i\) is the difficulty of item \(i\), \(C_j\) is the severity of judge \(j\) and \(F_k\) is the difficulty of observing category \(k\) relative to category \(k - 1\) (note that \(F\) is not a separate facet; rather it is part of the item facet). \(P_{nijk}\) and \(P_{nijk-1}\) are the probabilities of examinee \(n\) being graded on item \(i\) by judge \(j\), with a rating of category \(k\) and \(k - 1\), respectively. In this model, the rating scale is the same for all judges and all criteria. However, it is also possible to use a model in which each criterion has its own rating scale. This may be desirable if it is expected that the use of the scale categories is expected to differ between different criteria. This model can be formulated as follows:

\[
\log\left(\frac{P_{nijk}}{P_{nij(k-1)}}\right) = B_n - D_i - C_j - F_{jk}
\]

In other situations, one may want the scale to remain the same across criteria but to differ across judges. This can also be modeled using the following formula:

\[
\log\left(\frac{P_{nijk}}{P_{nij(k-1)}}\right) = B_n - D_i - C_j - F_{jk}
\]

The three-subscript term \(F_{ijk}\) is the difficulty of observing category \(k\) used by judge \(j\) on item \(i\). Using this model will allow researchers to compare how the individual raters interpret the various categories in each criterion scale.

Different MFRM models can be used to answer different questions about rating data, making it a useful tool for improving creativity performance assessment.

**TABLE 1**

<table>
<thead>
<tr>
<th>Types of Rater Misbehavior</th>
<th>Definition</th>
<th>How To Identify</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leniency/Severity/Generosity</td>
<td>Tending toward ratings in a lenient or severe manner.</td>
<td>This is usually parameterized directly in the “Rater” facet, and measures are automatically adjusted for it.</td>
</tr>
<tr>
<td>Extremism/Central Tendency</td>
<td>Tending to award ratings in the extreme, or in the central, categories</td>
<td>This can be identified by modeling each rater to have a separate rating scale (or partial credit). Anchor all items at the same difficulty, usually 0. Raters who best fit this situation are most likely to be exhibiting “halo” effect.</td>
</tr>
<tr>
<td>Halo/”Carry Over” Effects</td>
<td>One attributes bias ratings with respect to other attributes.</td>
<td>Anchor all persons at the same ability, usually 0. Raters who best fit this situation are most likely to be exhibiting response sets.</td>
</tr>
<tr>
<td>Response Sets/randomness effect</td>
<td>The ratings are not related to the ability of the participants.</td>
<td>Specify the Interrater facet and monitor the “agreement” percentage. The rater also tends to overfit. Include the “situation” as a dummy facet (e.g., rating session), and investigate rater-situation interactions.</td>
</tr>
<tr>
<td>Playing it safe</td>
<td>The rater attempts to give a rating near the other raters, rather than independently.</td>
<td></td>
</tr>
<tr>
<td>Instability</td>
<td>Rater leniency changes from situation to situation.</td>
<td></td>
</tr>
</tbody>
</table>

**Note:** These materials were adopted from the Facet online manual and we only placed it in a table format.
tool in evaluating the quality of ratings. Engelhard (1994) discussed how the MFRM can be used to investigate particular threats, such as rater severity, the halo effect, central tendency and restriction of range (Engelhard, 1994). Strategies for investigating particular sources of rater bias, outlined in the Facets manual (Linacre, 2010), are summarized in Table 1.

Additionally, the Facets software places the measures of the different facet elements (ability of examinee, rater severity, task difficulty, etc.) onto the same logit scale, facilitating direct comparison between elements. Given the power and versatility of the MFRM, and its use in other fields (Bond & Fox, 2007), it is somewhat surprising that it has not been used more widely in the area of creativity research. Only a few studies have applied the two-facet Rasch model in dealing with creativity data in which the model is used to reflect the underlying unidimensional construct of artistic quality (e.g., Kozbelt & Serafin, 2009) or fostering creativity (e.g., Teo & Waugh, 2010). Liu et al. (2010) used the three-facet model to assess the validity of consensus assessment and found that if the effect of severity is neglected, it will lead to biased performance scores. Aside from these exceptions, however, few studies in the field have addressed the issue of rater effects. This study seeks to address this gap, by illustrating the application of the MFRM in an investigation of rater effects in the evaluation of newly designed products in a creativity study. Thus, the second purpose of this study was to evaluate the performance of judges in assessing the level of creativity expressed in the design of new products.

METHOD

Participants

A total of 113 first- to third-year undergraduate students majoring in design at National Taipei University of Technology (53 men and 60 women, \( M_{age} = 20.5 \) years), participated as product designers as part of a design seminar. Participants were randomly allocated to the four experimental groups using a stratified design, so that each group contained equal numbers of first, second and third year students. One of these participants did not follow the task instructions, and the resulting product was not included in the subsequent analysis.

Judges

There were six judges. Three of the judges were studying for a master’s degree in the art department and had at least 6–7 years of art-related training, as well as some experience with evaluating works of art. The remaining three judges were majoring in educational psychology and all had some experience with creativity research, including experience with evaluating creative works and familiarity with the rating criteria used for this purpose. Four judges rated each design, with two art majors and two educational psychology majors randomly assigned to the designs.

Material and Procedure

The experimental procedure was based on a widely used procedure to induce mental synthesis (Finke, 1990). Participants were asked to choose three-dimensional basic object parts from the presented stimulus sets and then to use these objects to design a ‘creative invention’. The stimulus sets were 36 three-dimensional basic object parts (geons) proposed by Biederman (1987).

The experiment began with participants being shown the set of 36 objects. A 2\( \times \)2 factorial design was used, with two different sets of instructions (examples and no examples) and two task constraint conditions (use of only three of the 36 objects or unlimited use of all 36 objects). The four conditions were thus as follows: (a) no examples/limited objects (NL)—in this group, the participants were given instructions but were not shown examples of completed products, and they were allowed to choose only three out of the 36 objects to design their product; (b) no examples/no limit on objects (NN)—participants were given instructions without examples, but were allowed to freely use any of the 36 objects; (c) taught with examples/limited objects (TL)—participants were given both instructions and examples on PowerPoint slides, but were limited to the use of only three of the 36 objects; and (d) taught with examples/no limit on objects (TN)—participants were given both instructions and examples on PowerPoint slides and were allowed to use any of the 36 objects. Each student was given 20 min to complete the task (preliminary research had determined that this was ample time for task performance). Completed designs were photographed, scanned and uploaded to a Web site.

The six judges first underwent training to teach them to apply the criteria used in the study. After completion of the training, four judges—two art majors and two educational psychology majors—were assigned to evaluate each completed product. These evaluations were performed independently and online.

Criteria

The scoring criteria used were based on the aforementioned literature (the CPAM framework, in particular). Based on the purpose of this task, the criteria proposed by Hasirci and Demirkan (2007) were used to assess creative products. Eight scoring criteria, each on a 4-point scale, were presented in a fixed order for individual raters. The criteria included were the following: Appearance
originality—the degree of the appearance originality of the design (1 = very common to 4 = very original); functional originality—the degree of the functional originality of the design (1 = very common to 4 = very original); elaboration—the elaborateness of the overall appearance (1 = not elaborate to 4 = very elaborate); balance—the extent to which the overall design looks balanced (1 = not balanced to 4 = very balanced); multiple functions—the degree to which the design appears to support multiple functions (1 = single function to 4 = multiple functions); reasonableness—the degree to which the design seems reasonable or feasible (1 = not reasonable to 4 = very reasonable); and basic function—the degree to which the design satisfies its basic function (1 = does not satisfy its basic function to 4 = satisfies its basic function very well).

Statistical Techniques

To examine which facets affected the ratings, a sequence of MFRMs was applied. The Facets software program was used to estimate the MFRMs, using its default estimation algorithm, joint maximum likelihood estimation. Reliability and chi-square statistics were reported. In addition, the infit mean-square (weighted fit mean-square) influenced by response patterns and the outfit mean-square (unweighted fit mean-square) influenced by outliers were also reported. These two values were used as model fit indices. The infit and outfit mean-square indices have an expected value of 1 and can range from 0 to infinity. Linacre and Wright (1994) suggested that for the rating scale type, a reasonable range is 0.6–1.4, with a mean-square larger than two indicating serious distortion or degradation of measurement. Facets also provides several other indices that can be used in assessing rating quality. The point-biserial correlation is a many-facet version of the point-biserial correlation between responses and total score; for raters, it is the correlation between a single rater and all other raters (SR/ROR), across all facets. This statistic is highly data-design dependent, but is useful in flagging potential problems (e.g., negative ratings suggesting possible data entry errors). Exact agreement offers a comparison of the number of exact agreements modeled versus the number actually observed, and parallels Cohen’s Kappa, but with an expected value of 0. The separation reliability statistic is the signal-to-noise ratio based on the ratio of “true” to error variance. It is equivalent to the KR-20 or Cronbach’s Alpha statistic. This shows how different the measures are, which may or may not indicate how good the test is. High (near 1.0) person and item reliabilities are preferred. The statistic is somewhat the opposite of intrarater reliability, so low (near 0.0) judge and rater reliabilities are preferred, that is, a 1-separation reliability.

RESULTS

Detecting Rater Effects with the MFRM

To search for a model that provides the most parsimonious fit to the rating data, a sequence of MFRM models were applied to determine if any rater effects were present in the dataset. Assuming that judges shared the same rating scale, the three-facet rating scale model (i.e., formula 2) was applied as the baseline model and results were compared with results from the three-facet partial credit model (i.e., formula 3), which modeled the rating scale for each criterion to have its own category structure. Global data-model fit was assessed by examining the responses that were unexpected given the assumptions of the model. Linacre (2010) suggested that satisfactory model fit is indicated when about 5% or less of the (absolute) standardized residuals are equal to or greater than 2, and about 1% are equal to or greater than 3. The general data-model fit of the two models were examined separately. The results indicate the two models fit our data equally well, with less than 5% of the standardized residuals equal to or greater than 2 for the rating scale model (3.64%) and partial credit model (3.54%) respectively. Because the three-facet rating scale model is more parsimonious than the partial credit model, the results from the three-facet rating scale model were used to interpret the estimation results in detail.

Figure 1 presents the variable map. The first column labeled “Logit” is the logit scale, ranging from +2 to –3. All the other columns are discussed in reference to the logit scale. In the third column participants are represented by asterisks and plus signs (+) with each asterisk representing 2 participants and each ‘+’ representing 1 participant at that location on the logit scale. Raters at the top of the variable map tended to assign lower grades overall (stricter grading). Raters at the bottom of the column tended to assign a greater number of higher final grades overall (more lenient grading). The final column indicates the location of the category coefficient estimates for the four-category rating scale. In addition, several hybrid MFRMs were applied to answer a series of particular questions, each discussed in turn in the following.

Leniency/Severity Effect

Within the context of an MFRM analysis, leniency and severity effects are defined as a rater’s tendency to assign ratings that are, on average, higher or lower than those that other raters assign (Myford & Wolfe, 2004b). Table 2 provides a summary of the measurement report illustrated in the variable map. A fixed effect chi-square test was used to test the hypothesis that all raters were equally severe (or equally lenient). The results show a statistically significant difference between at least two raters’
ratings, $\chi^2(5) = 27.9, p < .001$. However, the rater fixed chi-square test is very sensitive to sample size. Further, the rater severity estimates was examined. As shown in Table 2, the mean and range of rater severity estimates were $-0.27$ and $0.34$ logit respectively. In addition, the variance of rater severity estimates among raters was rather small ($SD = 0.13$). These results indicate none of the six raters exhibited extreme leniency or severity.

Nevertheless, the separation statistics revealed that rater severity measures were far from being homogeneous. The separation reliability for raters was $0.78$, thus indicating raters were not behaving interchangeably when assigning final grades. The rater separation index showed that within this group of raters there were about three statistically distinct strata of severity. The consensus estimates provided by the software also indicated that the percentage of exact agreements for each rater was $54.0\%$. These results are not surprising. Raters were not required to behave as ‘rating machines’ in this study. Raters were only asked to use the rating scale consistently while rating. The mean infit and outfit statistics shown in Table 2 suggested that the rating behaviors of raters were very close to performances predicted by the model.

In sum, none of the six raters was identified as being extremely lenient or severe in the presented dataset.

### Central Tendency Effect

The central tendency effect is defined as the overuse of the middle categories of a rating scale. To detect the central tendency effect in a group, the fixed chi-square test for examinee and some other group level statistical indicators such as the examinee separation ratio, the examinee separation index, and the reliability of the examinee separation index are considered. Lower values indicate that a central tendency effect at the group level can be found in the dataset. The null hypothesis that all examinees exhibited the same performance measure was examined, after accounting for measurement error. As shown in Table 2, the fixed chi-square test for examinees was statistically significant, $\chi^2(111) = 752.1, p = .00$, indicating that no group-level central tendency effect was present in this study.

In addition, the examinee separation ratio of $2.54$ indicates that the spread of examinee performance measures was over two times larger than the precision of those measures. An examinee separation index of $3.72$, $(3.72^2 + 1)/3 = 3.72$ suggests that there were about four distinct strata of examinee performances in the sample of examinees. The reliability of the examinee separation index was $0.87$, implying that raters could reliably distinguish among the examinees in terms of their performance. Myford and Wolfe (2004b) suggested that the overfit and misfit of raters’ mean square could be a signal that raters exhibited a central tendency effect. Additionally, if the fit indices for the criterion are significantly less than $1$, this could be a signal that the raters as a group may be oversusing one or more categories on the scale. As shown in Table 2, the fit indices of the criterion did not exhibit any overfit or misfit. These results

### TABLE 2
Summary Statistics for Raters, Criteria and Examinees Facets

<table>
<thead>
<tr>
<th>Measures</th>
<th>Raters</th>
<th>Criteria</th>
<th>Examinees</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>-0.27</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SD</td>
<td>0.13</td>
<td>1.15</td>
<td>0.72</td>
</tr>
<tr>
<td>N</td>
<td>6</td>
<td>7</td>
<td>113</td>
</tr>
<tr>
<td>Range</td>
<td>0.34</td>
<td>2.89</td>
<td>3.95</td>
</tr>
<tr>
<td>Infit Mean</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>SD</td>
<td>0.12</td>
<td>0.12</td>
<td>0.42</td>
</tr>
<tr>
<td>Outfit Mean</td>
<td>1.02</td>
<td>1.02</td>
<td>1.02</td>
</tr>
<tr>
<td>SD</td>
<td>0.11</td>
<td>0.14</td>
<td>0.48</td>
</tr>
<tr>
<td>Reliability of separation</td>
<td>.78</td>
<td>.71</td>
<td>.87</td>
</tr>
<tr>
<td>Separation ratio</td>
<td>1.91</td>
<td>17.21</td>
<td>2.54</td>
</tr>
<tr>
<td>Chi-square statistic</td>
<td>27.9*</td>
<td>1944.9*</td>
<td>752.1*</td>
</tr>
</tbody>
</table>

*p < .001.
indicated that at the group level raters did not exhibit a central tendency effect.

However, these results cannot indicate how the criterion scale was qualitatively used. To pinpoint which particular raters or which particular criteria were most problematic, formulas 3, 4, and 5 were implemented. The percentage of ratings using categories 2 and 3 on the four-point scale across all criteria were 77% for appearance originality; 53% for functional originality; 75% for elaboration; 44% for balance; 54% for multiple function; 58% for reasonable and 41% for basic function. The results indicate that the distribution of ratings were spread out across all rating scale categories and did not exhibit a central tendency effect at the group-level.

Thus, formula 4 and formula 5 were applied to gain an understanding of how each rater used each category on each criterion scale. Table 3 presents the scale category statistics from an analysis using formula 4. The result shows that on the individual level, none of the rating scale category thresholds were reversed; they all increased monotonically for all raters. However, rater 6 used two of the categories in the scale infrequently, assigning 32% ratings in category 1 and 4. By contrast, over 40% of the other raters’ ratings were assigned to these outer categories. The high outfit index (1.4) for rater 6’s uses of category 4 also suggest that rater 6 might have had problem using this scale category. In addition, the distance between the rating category thresholds is greater for rater 6 than for the other raters. The result indicated that rater 6 included a comparatively wider range of examinee performance levels in each of the four rating categories employed.

The outfit mean-square index for each category was close to the expected value of 1 which means that the observed examinee performance measure and the expected examinee performance measure were very close. The result indicates that the raters were not exhibiting a central tendency effect.

### Halo Effect

Linacre (2010) stated that the halo effect can be investigated using Facets by constraining criterion difficulties to be the same and examining rater fit indices. Raters with the best-fit indices in this situation are most likely to be exhibiting the halo effect. Thus, all the criteria were anchored to the same difficulty and reanalyzed using formula 2. The results indicated that none of raters best fit the model. In addition, the fixed chi-square test for all raters was statistically significant, $\chi^2(5) = 15.3$, $p = .01$, indicating that there was no group-level halo effect present in the data. Besides, the fixed chi-square test of the hypothesis that all criteria are of the same calibrated level of difficulty was statistically significant, $\chi^2(6) = 1944.9$, $p = .00$. The high degree of criterion separation reliability (1.00) implies that raters could reliably distinguish among the criteria and indicates that the raters did not assign similar ratings to many examinees across a number of criteria. The criterion separation ratio of 17.21 indicates that the spread of criterion difficulty measures is about 17 times larger than the precision of those measures. In sum, these indicators did not suggest a group-level halo effect.

In addition, another strategy to detect halo effect is examining the rating of raters’ infit and outfit mean square indices. Myford and Wolfe (2004b) suggested that when criterion difficulties vary, the ratings of raters who exhibit halo effects would be very different from the expected ratings. This will result in rater infit and outfit mean square indices greater than one. As shown in Table 2, the rating of raters’ infit and outfit mean square indices are all close to one. This indicates that there was no halo effect at the individual level.

### Response Sets Effect

Response set effects may be due to raters who develop a different interpretation of the meaning of one or more of
the criterion scales. Insufficient training or rater background may also result in the rater being less capable of making finer distinctions. Therefore, to detect if raters tended to apply one or more of the trait scales in a manner inconsistent with the way that the other raters applied the same scales, the strategy of anchoring all examinees to the same level of performance and conducting a chi-square test was used (Linacre, 2010). Statistically significant chi square results would indicate that group-level response sets effects cannot be found in the data set. Results show a chi-square value of 747.8 with 111 degree of freedom, which was statistically significant \((p = .00)\), suggesting that there were probably no group-level response sets effect present in this data.

The high degree of examinee separation reliability \((0.87)\) also supports the conclusion that raters could distinguish among the examinees in terms of their performance. The rater fit mean-square indices were very close to the expected value of 1.00, implying that the raters did not exhibit randomness in the ratings assigned. Additionally, the SR/ROR correlation ranged from 0.43 to 0.51 which is considered an acceptable range. According to Myford and Wolfe (2004a), a point-biserial correlation of less than 0.3 is somewhat low; correlations higher than 0.70 are considered high for a rating scale composed of several categories. The median SR/ROR correlation also implies that there was no individual response sets effect.

**Interaction Analysis**

The most commonly used methods for assessing rater performance focus only on consensus or consistency estimates. In this study, rater agreement and SR/ROR correlations indicated an acceptable degree of rater consensus. However, findings from consensus and consistency estimates cannot wholly depict rating behavior. Thus, to investigate whether each rater maintained a uniform level of severity across examinees, or whether particular raters scored some examinees on a particular criterion more harshly or leniently than expected, two-way interaction analyses, for rater \(\times\) criterion and rater \(\times\) examinee, were conducted.

The expected rating score of each rater were compared to their observed rating scores. Thus, the bias size of each rater’s rating can be computed which indicates the size of the discrepancy between the observed score and the expected score of each rater for every rating criterion. A positive value means that the observed rating score was higher than the expected rating score. A bias size over .2 means that the observed rating is meaningfully different from the expected rating. Figure 2 depicts the relative measures for the raters’ rating on each criterion, indicating that raters may have used their own understanding in rating, but that they had different severity levels for each particular criterion. For example, rater 3 was more severe than expected when rating functional originality \((z = 4.28)\) but more lenient than expected when rating multiple function \((z = –2.23)\). Rater 2 was more severe than expected when rating balance \((z = 2.39)\) but more lenient than expected when rating functional originality \((z = –2.14)\) and reasonable \((-2.49)\).

In addition, results from the rater and criterion bias interaction analysis can also help to identify whether there are any individual-level halo effects in the dataset. None of the raters exhibited an individual halo effect in this study.

Analyses of interactions between rater and examinee were also conducted. Table 4 presents the interaction analysis for rater \(\times\) examinee compared to the analysis of rater \(\times\) criterion. The results show that raters applied the same level of severity; however, a small number of raters seemed more severe than expected when rating particular examinees. However, this seems to have occurred more frequently when raters rated on a particular criterion for which their severity or leniency exceeded the expected value \((28.57\%)\).

**TABLE 4**  
**Summary Statistics for the Interaction Analysis**

<table>
<thead>
<tr>
<th>Statistics</th>
<th>Type of Interaction</th>
</tr>
</thead>
<tbody>
<tr>
<td>N count</td>
<td>Rater (\times) Examinee</td>
</tr>
<tr>
<td>% large Z scores&lt;sup&gt;a&lt;/sup&gt;</td>
<td>1.56</td>
</tr>
<tr>
<td>Minimum Z</td>
<td>–2.58</td>
</tr>
<tr>
<td>Maximum Z</td>
<td>2.74</td>
</tr>
<tr>
<td>M</td>
<td>0.00</td>
</tr>
<tr>
<td>SD</td>
<td>0.88</td>
</tr>
</tbody>
</table>

*Note.* A percentage of absolute \(Z\) scores (standardized bias scores) equal to or greater than 2.
Table 5: Means, Standard Deviations, and Analysis of Variance Results for Task Constraint Condition and Instruction Condition on Creative Performance

<table>
<thead>
<tr>
<th>Task Constraint</th>
<th>ANOVA F(1, 108)</th>
<th>Instructions</th>
<th>M</th>
<th>SD</th>
<th>M</th>
<th>SD</th>
<th>Instruction</th>
<th>Constraint</th>
<th>I x C</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Limited</td>
<td></td>
<td>Creative</td>
<td>74.43</td>
<td>9.22</td>
<td>75.82</td>
<td>9.75</td>
<td>3.02</td>
<td>0.18</td>
<td>1.30</td>
</tr>
<tr>
<td>Limited</td>
<td></td>
<td>Example</td>
<td>73.26</td>
<td>12.55</td>
<td>70.21</td>
<td>9.53</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Via PPT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>No example</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The Effect of Instruction

After determining that no rater misbehaviors could be shown to exist in this study and raters were able to reliably distinguish examinee performances, the impact of different instruction sets and task constraints on creative performance were then investigated. Performance in the four different conditions (NL, NN, TN, and TL) were analyzed. A two-way ANOVA was performed to examine whether there were any interactions between the different conditions. The test of homogeneity of variances was not significant $F(3, 108) = 0.72, p = 0.54$, indicating that there was no violation of the homogeneity hypothesis. Descriptive statistics are presented in Table 5. Examinees in the example group scored slightly higher than the examinees in the no example group but the difference was not significant $F(1, 108) = 3.02, p = .09$. Likewise, no significant difference in performance were found based on whether or not examinees were limited in the number of geons they could use $F(1, 108) = 0.18, p = .67$. Finally, there was no interaction between the two factors $F(1,108) = 1.30, p = .257$.

Discussion and Conclusions

Although studies of creative products have been commonplace for decades, to date, studies addressing the quality of product ratings are still rare (Plucker & Renzulli, 1999).

The majority of creative performance rating studies to date have not questioned the assumption that raters perform as expected (Baer et al., 2004; Hennessey, 1994; Kaufman et al., 2008). However, given the substantial body of research documenting problems observed in rated performances, this cannot simply be assumed. Though the widely used consistency approach may help to assure consistency among raters, it does not preserve full information from the rating process and thus cannot be used to investigate whether rater effects affected the scores for a given study.

In this study, the measurement approach was adopted and the MFRM was applied to detect rater effects in advance to assure the quality of the ratings. Fortunately, no rater effects were found in the dataset, nor did raters differ widely in the severity with which they rated examinees. Raters were also fairly consistent in their overall ratings and reliably differentiated examinees in terms of their performance. However, in this study, every rater was asked to rate only about 74 tasks and raters received thorough training before being asked to perform their evaluations. It is conceivable that the quality of ratings might be lower in situations involving less training or a larger number of ratings.

The performance of the criteria was analyzed and the results did not reveal any dependencies between the criteria, and indicated that the criteria were appropriate and efficient for the evaluation of creative performances or products. For the examinees, the least difficult criterion to perform well on was basic function and the most difficult was multiple functions. That is, for the examinees, the criteria related to creativity were more difficult to attain than the criteria related to functional appropriateness. This finding suggests an explanation for why it is not easy for everyone to do things creatively rather than just appropriately, when doing design work.

The MFRM allows investigation of the interaction between facets. In this study, the interaction analysis is demonstrated to describe rating behavior in detail. The results of the interaction analysis indicated that an interaction between rater and criterion does exist in our data set. That is, some raters exhibited rating behavior that was more severe than expected in rating specific criteria. There are two plausible explanations for this finding. One is that judges differed in their interpretation of the rating scale, perhaps due to their not having a sufficient understanding of the relevant criteria. Another possible explanation is that, although judges understood the criteria, they nonetheless differed in the severity with which they bring to different criteria.

Deciding whether and how to remedy this may depend on one’s philosophy (i.e., whether one is inclined to treat raters as ‘scoring machines’ or independent experts or witnesses). If raters are treated as scoring machines, researchers may need to offer concrete scoring rubrics and accompany them with scoring examples in the training session. This may improve the rating consensus and increase the percentage of exact agreement. However, in this study, we followed the same thinking as most creativity researchers (e.g., Amable, 1996) in treating raters as independent witnesses and expected only that raters consistently classify performances according to their own understanding of the criteria.

Creativity studies typically use performance assessments as dependent measures and the purpose of ratings is to generate a summary score for each participant. In
this study, the MFRM was used to generate the summary scores. Researchers using different methods to evaluate rater quality may have different means of calculating summary scores. In such situations, it is important to bear in mind that adjustment to such scores may be called for in situations characterized by high consistency but low consensus. As Stemler (2004) noted, simply averaging scores across judges could lead to spurious results in subsequent analyses. For example, if a systematic difference has been overlooked then examinees scores are likely to be distorted. Therefore, for studies which aim to get summary scores for specific traits, it is recommended that reports should not consist solely of either consistence or consensus estimates. Moreover, the absence of rater bias must be demonstrated before calculating a summary score.

Finally, the effects of different experimental conditions on participant performance were examined. Our results indicated that neither differences in the instructions offered nor the presence or absence of task constraints significantly influenced performance. This result is not consistent with previous studies that indicated that participants allowed more flexibility in choosing task resources were rated as more creative (Amabile & Gitomer, 1984; Greenberg, 1992; Iyengar & Chua, 2008) nor with studies that indicated highly constrained conditions resulted in a greater number of creative inventions (Finke, 1990; Finke et al., 1992). Further, contrary to expectations, participants who were taught with examples via PowerPoint were not found to perform differently. This finding may be due to the very weakness of a ‘one shot’ treatment. It may also be necessary to consider whether an instruction effect may come from the informational content the instructions convey rather than the way in which instructions are presented. Further research might fruitfully investigate whether different methods of teaching creative thinking strategies (e.g., SCAMPER: Eberle, 1971) influence performance on creative design tasks.

One limitation of this study is that all of our raters were of the same gender (female). Individual differences among raters is an issue in need of further research (Hennessey, 1994) and further studies addressing the issue of differential rater functioning are recommended.

REFERENCE


