JOURNAL OF APPLIED MEASUREMENT, 5(2), 189-227

Detecting and Measuring Rater Effects Using Many-Facet Rasch Measurement: Part II

Carol M. Myford
University of Illinois at Chicago

Edward W. Wolfe
Michigan State University

The purpose of this two-part paper is to introduce researchers to the many-facet Rasch measurement (MFRM) approach for detecting and measuring rater effects. In Part II of the paper, researchers will learn how to use the *Facets* (Linacre, 2001) computer program to study five effects: leniency/severity, central tendency, randomness, halo, and differential leniency/severity. As we introduce each effect, we operationally define it within the context of a MFRM approach, specify the particular measurement model(s) needed to detect it, identify group- and individual-level statistical indicators of the effect, and show output from a *Facets* analysis, pinpointing the various indicators and explaining how to interpret each one. At the close of the paper, we describe other statistical procedures that have been used to detect and measure rater effects to help researchers become aware of important and influential literature on the topic and to gain an appreciation for the diversity of psychometric perspectives that researchers bring to bear on their work. Finally, we consider future directions for research in the detection and measurement of rater effects.

Requests for reprints should be sent to Carol M. Myford, College of Education, 1040 W. Harrison St., MC 147, Chicago, IL 60607-7133, e-mail: cmyford@uic.edu.

In Part II of this paper, we turn our attention to explaining how researchers can use a MFRM approach to detect five rater effects—(1) leniency/severity, (2) central tendency, (3) randomness, (4) halo, and (5) differential leniency/severity. (We will also discuss the detection of restriction of range, but we will treat it as a special case of leniency/severity, rather than as a separate rater effect.) We will operationally define each effect within the context of a MFRM approach, specify the particular measurement models needed to enable the researcher to detect the effect, and identify statistical indicators for the effect (both group- and individual-level indicators). The group-level indicators will be presented first, followed by the individual-level indicators.

As we present each effect and the various statistical indicators of that effect, we will show actual output from a *Facets* analysis (Linacre, 2001a, 2001b), pointing out the various indicators and explaining how to interpret each one. This paper takes as its starting point the seminal work of Linacre (1989) and Engelhard (1994) who have described the use of *Facets* to detect rater effects. However, it is important to note that there are other computer programs that can be used to conduct MFRM analyses to study rater effects (see, for example, *ConQuest* (Hoskens and Wilson, 2001; Wu, Adams, and Wilson, 1997)). Although *ConQuest* can estimate parameters for the models we describe in this paper, we have chosen to focus solely on output from *Facets* because of the computer program's popularity and the wide range of useful statistical indices that *Facets* generates.

We will present a number of different statistical indicators that are useful in detecting rater effects. However, researchers need to be aware that *Facets* provides other statistical information beyond what we present here. We have chosen to focus on those indicators that, in our view, are the most relevant to the investigation of rater effects. The first column of Table 1 lists the various indicators we will discuss. Column 2 pinpoints where in *Facets* output a particular indicator can be found. Columns 3-6 show which indicators are included as part of the output when rating scale or hybrid models are used to run the analyses. As the table shows, most of the indicators are included as "standard" output when rating scale or hybrid models are specified. What will differ are the rating scale category statistics tables (i.e., Table 8 of *Facets* output). Depending upon which model the researcher specifies, the output will contain varying numbers of these tables.

Specifications for the Simulations

To prepare our illustrative examples for Part II of this paper, we analyzed six simulated data sets. Five of the data sets contained simulated ratings for 10 raters who each rated 300 ratees on a single trait. All raters rated all ratees on that one trait. We used one of these simulated data sets as a baseline (i.e., no rater effects) and the remaining four to illustrate severity, central tendency, randomness, and differential severity. To give this a "real world" context, suppose that the 10 raters are a group of English-as-a-second-language teachers who are rating 300 students' audiotapes in which they have been asked to demonstrate their English speaking ability. The teachers are rating the audiotapes on one trait—communicative language ability. The rating scale devised to measure this trait has seven separately defined categories, ranging from 0, which is defined as "no effective communication," to 6, which is defined as "communication almost always effective."

The sixth simulated data set had the 10 raters each rate 300 ratees on four traits. Again, all raters rated all ratees on all four traits (i.e., a fully crossed design—a luxury not usually afforded by most rating designs). We used this simulated data set to illustrate the halo effect. To establish a real-world context for this data set, suppose that the 10 English-as-a-second-language teachers rated the 300 students' audiotapes on four traits: (1) linguistic competence, (2) discourse competence, (3) functional competence, and (4) sociolinguistic competence. The raters used four separate rating scales, one for each of the four traits. Each rating scale had seven categories with each category

Table 1

Group- and Individual-level Statistical Indicators for Detecting Rater Effects Included in Facets *Output When Various Models Are Specified*

Group- and Individual-level Statistical Indicators	Where found in *Facets* output?	Rating Scale Model	Hybrid Model #1	Hybrid Model #2	Hybrid Model #3
All Facet Vertical "Rulers" (i.e., a variable map showing rater severity measures, ratee performance measures, trait difficulty measures, and rating scale category thresholds —all displayed on a logit scale)	Table 6	✓	✓	✓	✓
Rater severity measures, fair averages, standard errors, fit mean-square indices	Table 7	✓	✓	✓	✓
Ratee performance measures, standard errors, fit mean-square indices	Table 7	✓	✓	✓	✓
Trait difficulty measures, standard errors, fit mean-square indices	Table 7	✓	✓	✓	✓
Fixed chi-square tests for raters, ratees, and traits	Table 7	✓	✓	✓	✓
Separation ratios and reliabilities for raters, ratees, and traits	Table 7	✓	✓	✓	✓
"Single rater—rest of the raters" (SR/ROR) correlations	Table 7	✓	✓	✓	✓
Table of Misfitting Ratings	Table 4	✓	✓	✓	✓
Scale Category Statistics (frequency counts of scale category usage, rating scale category thresholds, rating scale category outfit mean-square indices)					
—for all raters across all traits (i.e., one table)	Table 8	✓			
—for each trait *across all raters* (i.e., a separate table for each trait)	Table 8		✓		
—for each rater *rating a single trait* (i.e., a separate table for each rater)OR	Table 8			✓	
—for each rater *rating a set of traits* (i.e., a separate table for each rater)					
—for each rater *for each trait* (i.e., a separate table for each rater for each trait)	Table 8				✓
Scale Category Probability Curves	Table 9	✓	✓	✓	✓
Bias Interaction terms (*z*-scores), group separation ratio, group separation reliability estimate	Table 13	When a bias interaction analysis is requested	When a bias interaction analysis is requested	When a bias interaction analysis is requested	When a bias interaction analysis is requested

Note: A researcher will need to conduct a *Facets* bias interaction analysis in order to obtain *z*-scores for interaction terms (e.g., Rater x Trait, Rater x Ratee Group), a group separation ratio, and a group separation reliability estimate.

separately defined, as in the example in the previous paragraph. (One could think of these as four "items.")

In the baseline data set, we simulated data to fit the many-facet two-parameter rating scale model (i.e., a multifaceted version of the Generalized Partial Credit Model),

$$ln[P_{nijk} / P_{nijk-1}] = E_j(B_n - D_i - C_j - F_k) \quad (1)$$

where,

E_j = a slope for the item characteristic curve associated with rater j.

(It is important to note that *Facets* does not analyze data exactly according to this model; that is, there is no E_j term, but one can parameterize an F_{jk} term.) Ratee performance (B_n) was drawn from a normal distribution with a mean of 0.00 and a standard deviation of 1.00. To control for the magnitude of the correlation between raters, we sampled a separate ratee performance distribution for each of the ten raters, so that the performance parametric distributions were correlated at about $r = .72$. We set the six rating scale category coefficients (F_k) to equal the following values: [-2.50, -1.50, -0.50, 0.50, 1.50, 2.50]. We set the trait difficulty (D_i) to equal 0.00. Rather than simulating a situation in which absolutely no rater effects exist, we simulated the data so that very small rater effects existed for all raters in the baseline data set. Specifically, rater severity was drawn from a normal distribution with mean equal to 0.00 and standard deviation equal to 0.20, with a typical range extending from -0.90 to 0.90. Rater slope, a parameter designed to alter the dispersion of the ratings that a particular rater assigns (e.g., steep rater slopes would be indicative of central tendency), was sampled from a log-normal distribution (in accord with PARSCALE (Muraki and Bock, 2003) expectations) with a mean equal to 1.00 and a standard deviation equal to 0.14, with a typical range extending from 0.54 to 1.85.

Table 2 summarizes the rater parameter estimates and the related statistics that a *Facets* analysis of the baseline data produced (Table 7 of the *Facets* output). The first two columns of this table show the sum of the ratings and the number of ratings that each simulated rater assigned, respectively. Similarly, the third column shows the average rating each rater assigned, and the fourth column shows the average expected rating for each rater (the "fair average" based on the MFRM model). Note that, consistent with the simulation procedure, the elements in the column of average ratings are similar in size (i.e., within column three). Also, note that each average rating is consistent with its expected average rating (i.e., comparing columns three and four). The rater severities, shown in the fifth column, differ from one another only slightly, with a standard deviation equal to 0.35—which was about eight percent of the standard deviation of the ratee performance measure estimates. The rater fit indices, shown in columns seven through ten, all indicate that the ratings are consistent with the

Table 2

Rater Measurement Report from an Analysis Using the Rating Scale Model—Baseline Simulation

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Avrage	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	PtBis	Nu RATERS
781	268	2.9	2.96	-.12	.11	.9	0	.9	0	.95	1 1
797	268	3.0	3.03	-.30	.11	.8	-2	.7	-3	.96	2 2
759	268	2.8	2.87	.12	.11	.9	0	.9	0	.95	3 3
721	268	2.7	2.71	.54	.11	1.1	0	1.1	1	.94	4 4
721	268	2.7	2.71	.54	.11	.9	-1	.8	-1	.95	5 5
746	268	2.8	2.82	.27	.11	.9	-1	.8	-2	.96	6 6
801	268	3.0	3.05	-.34	.11	.9	-1	.9	-1	.95	7 7
818	268	3.1	3.12	-.53	.11	.9	0	1.0	0	.95	8 8
787	268	2.9	2.99	-.19	.11	1.0	0	1.0	0	.95	9 9
770	268	2.9	2.92	.00	.11	1.0	0	1.1	0	.94	10 10
770.1	268.0	2.9	2.92	.00	.11	.9	-.9	.9	-.9	.95	Mean (Count: 10)
31.5	.0	.1	.13	.35	.00	.1	1.0	.1	1.2	.01	S.D.

RMSE (Model) .11 Adj S.D. .33 Separation 3.15 Reliability .91
Fixed (all same) chi-square: 108.9 d.f.: 9 significance: .00

MFRM model. Finally, the SR/ROR correlation, shown in the eleventh column, indicates that the ratings that these raters assigned exhibited a high level of agreement (i.e., the average interrater correlation is equal to 0.95).

The data sets that we simulated to illustrate severity, central tendency, randomness, and differential severity were all generated by replacing the simulated ratings of Rater 10 (the "effect" rater) with ratings designed to simulate the effect in question. We designated the nine other raters (whose ratings were left unchanged) as "normal" raters.

In the *severity simulation*, we modeled the "effect" rater to have a severity of 1.00—slightly outside of the typical range of the "normal" raters in the baseline data. As a result, the "effect" rater assigned ratings that were, on average, about 0.80 points lower on the rating scale than did "normal" raters in the baseline data set (M_{effect} = 2.02 versus M_{normal} = 2.91). It is important to note that, in this simulation, we modeled the data to show a single rater exhibiting a severity effect rather than a group-level severity effect.

In the *central tendency simulation*, we regressed the ratings of the "effect" rater toward the mean by specifying a slope parameter for that rater that equaled 2.00 (again, slightly outside the typical range for "normal" raters in the baseline data set). As a result, the standard deviation of the observed ratings for the "effect" rater was about 75% of the standard deviation of the observed baseline ratings that "normal" raters assigned (SD_{effect} = 1.40 versus SD_{normal} = 1.87). In this simulation, we modeled the data to show a single rater exhibiting a central tendency effect, not to show a group-level central tendency effect.

In the *randomness simulation*, we added random error to the ratings of the "effect" rater by specifying a ratee performance distribution for a portion of the ratees that the "effect" rater rated that was uncorrelated with the ratee performance distribution specified for the "normal" raters. In this case, we replaced 25% of the ratings that the "effect" rater assigned with ratings randomly selected from a distribution with the same mean and standard deviation as that of the "normal" raters. The remaining 75% of the ratings that the "effect" rater assigned were normal (i.e., taken from the baseline data set for that rater). This resulted in a SR/ROR correlation for the "effect" rater that was somewhat smaller than the SR/ROR correlation for the "normal" raters (r_{effect} = .70 versus r_{normal} = .95). In this simulation, we modeled the data to show a single rater exhibiting a randomness effect, not to show a group-level randomness effect.

In the *halo simulation*, we assigned the values of 0.00, -0.50, -1.00, and 0.50 for the trait difficulties for traits one, two, three, and four, respectively. As was true for the remaining analyses, we added a small amount of rater leniency/severity and rater central tendency to the ratings of all raters, using the same values (constant across traits) as in the baseline data set. Also, to control for the magnitude of interrater correlations, we generated a separate trait distribution for each ratee-by-rater combination, with these distributions being correlated at about r = 0.92 within traits, and interrater correlations of about .78 between traits. We simulated the "effect" rater using the same ratee trait distribution for all traits so that the inter-trait correlation was about .92 for that rater. In other words, while "normal" raters produced ratings that were consistent between raters within a trait, the ratings of the "effect" rater were consistent between traits within that rater. In this simulation, we modeled the data to show a single rater exhibiting a halo effect, not to show a group-level halo effect.

In the *differential severity* simulation, we randomly assigned gender codes (male and female) to ratees. We simulated the "normal" raters to assign comparable ratings to the two groups, while the "effect" rater assigned ratings that exhibited an even greater advantage to females ($M_{normal|male}$ = 2.70, $M_{normal|female}$ = 2.87, $M_{effect|male}$ = 2.00, $M_{effect|female}$ = 2.96).[1] This effect was achieved by increasing the "effect" rater's severity by 1.00 logit for males.

It is important to emphasize from the outset that all of the individual-level rater effect indices that we will be presenting are relative measures of these effects—that is, they compare the rating

behavior of a given rater to the behavior of other raters included in the same analysis. When we identify a rater as exhibiting a particular rater effect, the rater in question exhibits the effect *relative to other raters.* For example, if we identify a rater as exhibiting severity, that rater is rating severely in comparison to the other raters. However, we do not know whether the rater in question is rating too severely, or alternatively, whether the other raters in the analysis are rating too leniently. Since we do not have access to the "true" ratings of each ratee (i.e., the valid, accurate measure of each ratee's level of performance), it is not clear which of these explanations is correct and should be accepted. As we discuss the illustrative examples that follow, it is important to keep in mind that when we identify a rater as exhibiting aberrant rating behavior, we cannot be assured that our interpretation of that behavior is the only valid interpretation, since we do not have access to the "true" ratings we would need in order to substantiate our interpretation.

Leniency/Severity Effect

Conceptual definition. Within the context of a MFRM analysis, rater severity is traditionally defined as a rater's tendency to assign ratings that are, on average, lower than those that other raters assign, even after the performances of the particular ratees that that rater has evaluated are taken into account. According to this definition, severe raters underestimate the level of ratee performance across the entire performance continuum. They do not accurately assess the level of performance of ratees at any point along that continuum. Rather, they tend to assign ratings that are consistently lower than those that other raters would assign the same ratees.

Similarly, rater leniency is traditionally defined as a rater's tendency to assign ratings that are, on average, higher than those that other raters assign, even after the performances of the particular ratees that that rater has evaluated are taken into account. By this definition, lenient raters tend to overestimate the level of ratee performance across the entire performance continuum, assigning ratings that are consistently higher than those that other raters would assign the same ratees. When researchers use the term "leniency/severity effect," it is often with this intended meaning. However, a leniency/severity effect can present itself in other ways, some more subtle than this.

Some raters may exhibit a tendency to cluster their ratings around a particular category on a rating scale (i.e., show restriction of range in their ratings). That category may be at the high end of the scale, the low end of the scale, or in the middle of the scale. If a rater's ratings tend to cluster at the high end of the scale, then that may signal leniency. By contrast, if a rater's ratings tend to cluster at the low end of the scale, then that may signal severity. Note that in these examples the rater does not overestimate (or underestimate) ratee performance across the entire performance continuum—only along a portion of that continuum. The net effect is still detectable as rater leniency/severity, though the pattern of ratings for a rater showing restriction of range may differ somewhat from the pattern of ratings for a rater who consistently assigns higher (or lower) ratings than other raters to all ratees. However, as these examples point out, it is often difficult to differentiate clearly between restriction of range and leniency/severity as separate effects, though they are frequently portrayed as such in the rater effects literature.[2]

Finally, a rater may selectively exhibit a leniency/severity effect. That is, a rater may be differentially severe, showing a tendency to assign ratings that are lower than expected to certain groups of ratees, given the ratings that other raters assign ratees in those groups. Again, this is a more subtle form of the rater leniency/severity effect that we refer to as *differential leniency/severity.* As we shall see later, differential leniency/severity has its own special methods of detection within a MFRM framework.

Measurement models for detecting the leniency/severity effect. Researchers using a rating scale model or any of the hybrid models shown in Table 1 to analyze their rating data will obtain the group- and individual-level statistical indicators described below.

Group-level statistical indicators. The output from a MFRM rating scale analysis will contain a table (i.e., Table 8 of *Facets* output) that summarizes how the raters (as a group) used the scale categories (across all trait scales). By reviewing this table, the researcher can determine whether overall the raters exhibited a general tendency to overuse the lower scale categories (i.e., exhibit severity) or overuse the higher scale categories (i.e., exhibit leniency).

The output from a MFRM analysis using a rating scale model or a hybrid model includes four group-level statistical indicators of a group-level leniency/severity effect.

(1) A *fixed chi-square test* of the hypothesis that the rater severity measures are not significantly different (i.e., that all raters share the same severity measure, after accounting for measurement error).

> *Example*: The results from the rater fixed chi-square test are shown in the bottom line of our Table 3, the Rater Measurement Report. (Note that the Rater Measurement Report appears in *Facets* output as Table 7). The chi-square value of 702.2 with 9 degrees of freedom is statistically significant ($p <$ .005), signifying that the raters did not all exercise the same level of severity when evaluating ratees. A significant rater fixed chi-square simply means that the severity measures of at least two of the ten raters included in this analysis are significantly different. However, it is important to emphasize that the rater fixed chi-square test is very sensitive to sample size. As a result, in many applications of MFRM, the rater fixed chi-square statistic may be statistically significant even if the actual variation between raters in the levels of leniency/severity exercised is small[3].

(2) The *rater separation ratio.* This ratio is a measure of the spread of the rater severity measures relative to the precision of those measures.

> *Example*: The rater separation ratio is 8.53 (shown in the second line from the bottom of our Table 3). This means that the differences between rater severities are over eight times greater than the error with which these severities are measured.

(3) The *rater separation index.* This indicator connotes the number of statistically distinct levels of rater severity among the sample of raters included in the analysis. Specifically, this index depicts the "true" variance in "error" variance units.

> *Example*: The rater separation index of 11.71 suggests that there are about twelve statistically distinct strata of rater severity in this sample of raters. (The rater separation index does not appear as part of *Facets* output. The researcher will

Table 3

Rater Measurement Report from an Analysis Using the Rating Scale Model—Severity Simulation

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Avrage	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	PtBis	Nu RATERS
793	268	3.0	2.96	-.42	.11	1.0	0	1.0	0	.95	1 1
809	268	3.0	3.02	-.60	.11	.8	-2	.8	-2	.96	2 2
771	268	2.9	2.87	-.16	.11	1.0	0	1.0	0	.95	3 3
733	268	2.7	2.72	.29	.11	1.1	1	1.2	1	.94	4 4
733	268	2.7	2.72	.29	.11	.9	0	.9	-1	.95	5 5
758	268	2.8	2.82	.00	.11	.9	-1	.9	-1	.96	6 6
813	268	3.0	3.04	-.65	.11	.9	0	.9	0	.95	7 7
830	268	3.1	3.11	-.85	.11	1.0	0	1.0	0	.95	8 8
799	268	3.0	2.98	-.49	.11	1.0	0	1.0	0	.94	9 9
545	268	2.0	1.94	2.58	.11	1.1	0	1.0	0	.93	10 10
758.4	268.0	2.8	2.82	.00	.11	1.0	-.4	1.0	-.5	.95	Mean (Count: 10)
77.8	.0	.3	.32	.94	.00	.1	1.0	.1	1.0	.01	S.D.

RMSE (Model) .11 Adj S.D. .93 Separation 8.53 Reliability .99
Fixed (all same) chi-square: 702.2 d.f.: 9 significance: .00

need to compute this index manually using the formula (4G + 1) / 3, where G is the rater separation ratio, which is included as part of *Facets* output. In this example, [4 (8.53) + 1] / 3 = 11.71). It should be noted that the rater separation index will be large, on average, when the number of observations per rater (i.e., the combination of ratees and traits) is large. This is because the error variance, as depicted by the average standard error of the rater severity measures, will decrease as the number of ratees and traits increase (i.e., each rater is measured more precisely under these conditions). This may result in rater separation indices that indicate a greater number of statistically distinct strata than there are actual raters in the analysis. The interpretation of such a situation would be fairly straightforward—that is, the spread of the rater severity measures is considerably greater than the precision of those measures.

(4) The *reliability of the rater separation index.* This indicator provides information about how well the raters are separated in order to reliably define the rater facet. It is a measure of the spread of the rater severity measures relative to their precision, reflecting potentially unwanted variation between raters in the levels of severity exercised. For example, a reliability of rater separation index of .70 would suggest that, on average, there are discernible statistically significant differences between the severe and lenient raters. (In many situations, the most desirable result is to have a reliability of rater separation close to zero, which would suggest that the raters were interchangeable, exercising very similar levels of severity.)

Example: The rater separation reliability is shown in the second line from the bottom of our Table 3. The high degree of rater separation reliability (.99) implies that raters are differentiated in terms of the levels of severity they exercised. There is some evidence here of unwanted variation between raters in their levels of severity.

Individual-level statistical indicators. Individual-level leniency/severity effects are evident when the researcher looks at the variable map that is included in *Facets* output (as Table 6 from a *Facets* analysis, the All Facet Vertical "Rulers"). Figure 1 provides an example of a variable map. This figure shows the distribution of rater severity measures (and the distribution of ratee performance measures) from our baseline simulated data set in which none of the ten raters who evaluated ratees were modeled to exhibit extreme leniency or severity. Raters are ordered in a variable map in terms of the levels of severity each exercised. More severe raters appear at the top of column 3, while more lenient raters appear lower in the column. The rater severity measures shown in Figure 1 are tightly clustered; that is,

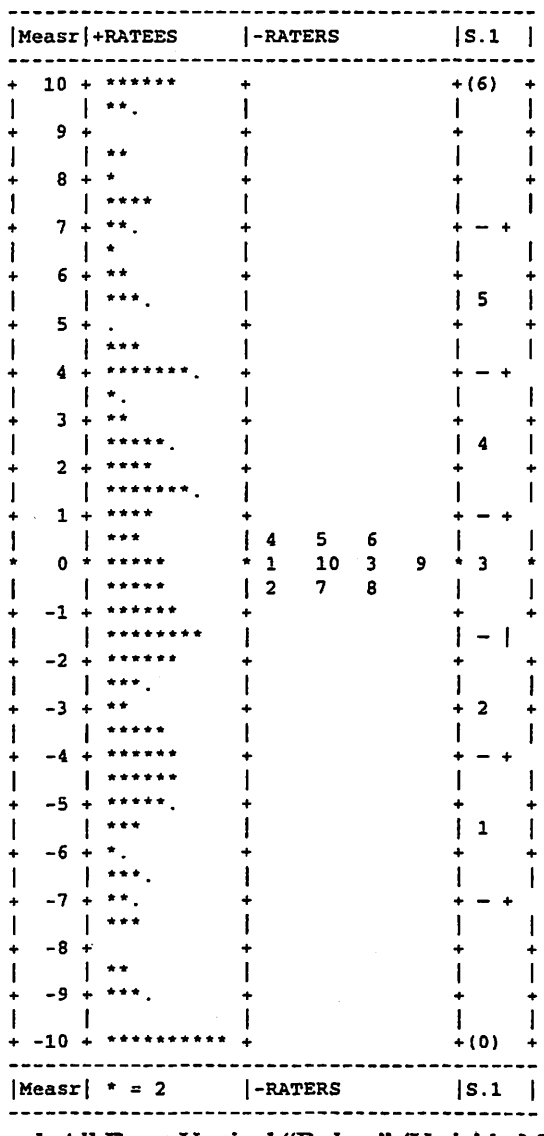

Figure 1. All Facet Vertical "Rulers" (Variable Map) for Baseline Simulation

most of the measures range from -0.50 logits to +0.50 logits. Now look at Figure 2, which shows the distribution of rater severity measures (and the distribution of ratee performance measures) for the simulated severity example. Most of the rater severity measures from this analysis range from -0.50 logits to +0.50 logits. However, Rater 10 (the "effect" rater, shown higher in the column) stands out as being more severe, having a severity measure of nearly +2.50 logits.

These figures graphically depict the manner in which rater severity and leniency are captured by MFRM analyses. *Facets* also includes as part of its output a table that provides the individual rater severity measures (in logits) and the standard error of each severity estimate, indicating the precision with which a rater's severity has been measured. Using the severity measures and their attendant standard errors, the researcher can perform *t*-tests as a follow-up to the fixed chi-square test to compare pairs of raters to determine whether their severity measures are significantly different.

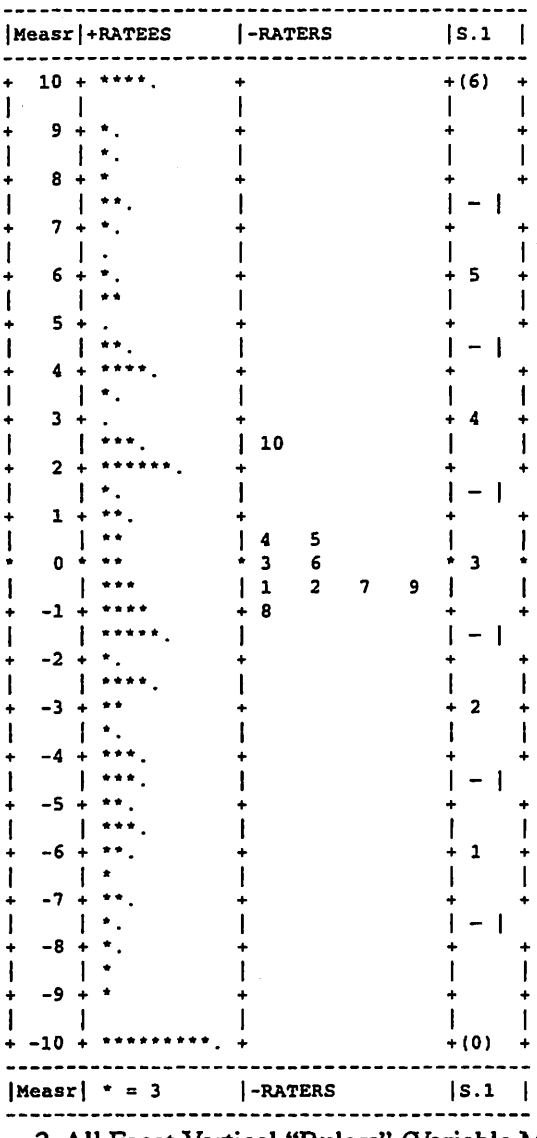

Figure 2. All Facet Vertical "Rulers" (Variable Map) for Severity Simulation

Example: The individual rater severity measures for the ten raters included in our analysis are shown in the "Measure" column of our Table 3. (Note that the rater severity measures appear in Table 7 of *Facets* output.) The larger the measure, the more severe the rater. The standard error for each severity measure appears in the "Model S. E." column. All raters, except for Rater 10, have severity measures that are relatively close to the mean severity of zero. (The mean of the rater severity measures is shown in column 5, four lines up from the bottom of our Table 3.) In fact, the severity measures for the first nine raters listed in the table are all within the range of -.85 to .29 logits. The severity measure for Rater 10 (2.58 logits) is a conspicuous outlier—over 23.5 standard errors above the mean severity of the group. (To obtain 23.5, divide 2.58 (the severity measure for Rater 10) by .11 (the model standard error).)

The researcher can get some sense of just how much more severe or lenient an individual rater is in comparison to the other raters by examining the raters' average ratings. However, if all raters do not rate all ratees, it is difficult to determine how much each rater's average rating is influenced by the particular sample of ratees that he or she evaluated. For example, if the average rating for Rater A is lower than the average rating for other raters, there are two plausible explanations for why Rater A tended to assign more low ratings than other raters. Perhaps the set of ratees that Rater A evaluated did indeed exhibit lower levels of performance than the sets of ratees that other raters evaluated. If this were the case, the fact that Rater A tended to assign an overabundance of lower ratings would have been entirely appropriate, given that the ma-

jority of ratees that this rater evaluated were lower level ratees. Chances are that other raters evaluating this same set of ratees would also have given them low ratings. According to this explanation, Rater A has a lower average rating because the ratees that Rater A evaluated were lower performing. Following this line of reasoning, if Rater A had evaluated a set of ratees that were higher performing, then Rater A would not have given an overabundance of lower ratings.

A second possible explanation is that Rater A tended to use the rating scales in a different manner than other raters, systematically assigning lower ratings than other raters. If this were the case, then when other raters evaluated the set of ratees that Rater A evaluated, these ratees would have received higher ratings than Rater A gave them. When a rater has a lower average rating than other raters, it is difficult to decide which of these explanations for the rater's behavior is the correct one.

By using a MFRM approach to analyzing the rating data, a researcher can gain needed insights to facilitate this determination. The output from a MFRM analysis contains a "fair average" rating for each rater. The fair average is the average rating for each rater once that average has been adjusted for the deviation of the ratees in each rater's sample from the overall ratee average across all raters and traits. By comparing the raters' "fair averages," the researcher can pinpoint those raters who tended to use the rating scales in a different manner than other raters (i.e., who assigned ratings that were on average lower than those that the other raters in the sample assigned, even after the particular ratees that that rater evaluated were taken into account). The "fair averages" are reported in Table 7 of *Facets* output.

> *Example*: The third column of our Table 3 ("Obsvd Average") shows each rater's average rating. The fourth column ("Fair-M Avrage") shows each rater's average rating adjusted for the deviation of the ratees in that rater's sample from the overall ratee mean. When we compare the raters' "fair averages," we see that Rater 10, the most severe rater among the ten raters included in this analysis, had a fair average of 1.94, while the most lenient rater, Rater 8, had a fair average of 3.11. This suggests that, on average, Rater 10 assigned ratings that were 1.17 raw score points lower than the ratings that Rater 8 assigned (i.e., on average, the ratings of Rater 8 tended to be over one rating scale category higher than the ratings of Rater 10).

Central Tendency Effect

Conceptual definition. Within the context of a MFRM analysis, central tendency is defined as overusing the middle categories of a rating scale. Central tendency, a special case of restriction of range, can present itself in several different ways.

In some cases, the rater who tends to overuse the middle categories of the scale may be able to accurately assess the level of performance of the very highest and lowest performing ratees (i.e., those whose performance measures fall at the extreme upper and lower ends of the performance continuum). The rater understands what constitutes really good performance and really weak performance on the trait and can use the very highest and lowest categories of the scale appropriately to assign ratings. However, the rater tends to inaccurately assess ratees whose levels of performance fall in between those extremes. Any ratee that is not really strong or really weak gets assigned a rating in the middle categories of the scale, but on an indiscriminate basis. The rater does not understand the distinctions between the middle categories of the scale, and thus is unable to use those categories in a consistent fashion to differentiate among these average-performing ratees.

Central tendency can also manifest itself as a rater's inability to differentiate among ratee performance levels along the entire performance continuum. In this situation, the rater does not understand the distinctions between any of the scale categories, and thus resorts to assigning all ratees similar "middle-of-the-road" ratings. The rater may not have sufficient background and/or training to be able to make the fine discrimina-

tions required in order to employ the scale appropriately. However, instead of using all the rating categories indiscriminately (as in the randomness effect), the rater showing central tendency tends to only assign ratings in the middle categories, rarely using the outer categories.

A rater may occasionally exhibit a central tendency effect when working in a rating operation in which raters are being carefully monitored and given feedback on their performance. If a rater is singled out for assigning too many low ratings in comparison to other raters (i.e., being too severe), or for assigning too many high ratings in comparison to other raters (i.e., being too lenient), then the rater may adopt a "play-it-safe" strategy (Myford and Mislevy, 1995; Wolfe, Chiu, and Myford, 1999). The rater may start overusing the middle categories of the rating scale in an attempt to minimize the possibility that she or he will be singled out again for assigning an overabundance of discrepant ratings (i.e., ratings that do not agree with the ratings that other raters assign the same ratees). Thus, the pattern in this situation would be for a rater to exhibit a severity or leniency effect initially, and then to try to overcompensate for that tendency by assigning more "middle-of-the-road" ratings, which would then be detectable as a central tendency effect in the rater's ratings.

If many raters show a pattern of central tendency in their ratings, then there may be a problem with the rating scale, not with the raters. Perhaps the trait scale has too many categories and is seeking to make distinctions among ratees that are too fine grained (or are not really there). In this situation, the remedy may involve revising the rating scale to include fewer categories in order to make the distinctions between the resulting categories more readily evident. Alternatively, the category definitions for the existing scale might need to be revised so that the boundaries between categories are made clearer.

Measurement models for detecting the central tendency effect. Researchers using a rating scale model or a hybrid model to analyze their rating data will obtain rater and trait fit mean-square indices. To obtain some of the other statistical indicators described below, the researcher will need to use one of the hybrid models.

Group-level statistical indicators. The output from a MFRM analysis employing a rating scale model will include a table that summarizes how the raters (as a group) used the scale categories (across all trait scales). From this table (i.e., Table 8 of *Facets* output), the researcher can determine whether the raters exhibited a general tendency to overuse the middle categories of the rating scale. However, the researcher will not be able to pinpoint which particular raters (or which particular trait scales) were most problematic. For that level of diagnosis, the researcher will need to use a hybrid model to analyze the rating data.

> *Example*: Our Table 4 shows scale category statistics from a MFRM analysis in which we used a rating scale model to analyze the central tendency simulation data set. (The information contained in this table appears in *Facets* output as Table 8). The frequency count and percentage of ratings that raters assigned in each rating scale category are shown in the second and third columns of our Table 4. For this data set, it is clear that there is not a pervasive group-level trend toward central tendency on the part of all raters. The distribution of ratings is

Table 4

Category Statistics from an Analysis Using the Rating Scale Model—Central Tendency Simulation

DATA				QUALITY CONTROL			STEP		EXPECTATION		MOST	.5 Cumul.	Cat	Obsd-Expd
Category	Counts		Cum.	Avge	Exp.	OUTFIT	CALIBRATIONS		Measure at		PROBABLE	Probabil.	PEAK	Diagnostic
Score	Used	%	%	Meas	Meas	MnSq	Measure	S.E.	Category	-0.5	from	at	Prob	Residual
0	315	11%	11%	-6.79	-6.75	.9			(-6.97)		low	low	100%	-2.1
1	407	14%	25%	-4.39	-4.38	.8	-5.84	.09	-4.73	-6.04	-5.84	-5.93	60%	
2	545	19%	44%	-2.26	-2.26	1.0	-3.59	.08	-2.38	-3.56	-3.59	-3.57	61%	1.7
3	546	19%	64%	-.09	-.11	1.1	-1.21	.07	-.02	-1.20	-1.21	-1.21	61%	2.6
4	490	17%	81%	2.15	2.13	.9	1.13	.08	2.41	1.17	1.13	1.14	64%	1.3
5	308	11%	92%	4.40	4.43	.9	3.71	.09	4.77	3.63	3.71	3.66	58%	-.8
6	239	8%	100%	6.59	6.57	.9	5.79	.10	(6.95)	6.04	5.79	5.89	100%	-2.3
									(Mean)		(Modal)	(Median)		

spread out across all rating scale categories, with a non-trivial proportion of ratings falling in the upper and lower rating scale categories (i.e., categories 0 and 6).

When most of the raters exhibit central tendency effects, there should be a lack of variation between ratees in the level of performance demonstrated. Because the raters overused the middle categories of the scale, they would not be able to distinguish reliably among ratees, since many ratees would have received similar "middle-of-the-road" ratings. (Alternatively, it is important to remember that in some situations a group of ratees may, in fact, be quite homogeneous in their performance, differing little. If this were the case, then the fact that the raters tended to use the middle categories of the scale when evaluating those ratees' performance would not be considered rater "error." Rather, their rating behavior would be entirely appropriate, given that there truly is a lack of variation between ratees in the level of performance demonstrated.)

The output from an analysis using the rating scale model or any of the hybrid models includes several group-level indicators of central tendency that focus on the measurement of the ratees:

(1) A *fixed chi-square test* of the hypothesis that all ratees exhibit the same calibrated level of performance (i.e., that all ratees share the same performance measure, after accounting for measurement error). A nonsignificant chi-square value suggests a group-level central tendency effect.

Example: The results from the ratee fixed chi-square test are shown in the bottom line of our Table 5, the Ratee Measurement Report. (Note that what is shown here is the summary report that appears at the bottom of Table 7 in *Facets* output, not the entire Table 7.) The chi-square value of 13718.5 with 284 degrees of freedom is statistically significant ($p < .005$), suggesting that there is not a group-level central tendency effect present in this simulation data set.

(2) The *ratee separation ratio*. This ratio is a measure of the spread of the ratee performance measures relative to the precision of those measures. A low ratee separation ratio suggests a group-level central tendency effect.

Example: The ratee separation ratio of 7.29 (shown in the second line from the bottom of our Table 5) indicates that the spread of the ratee performance measures is over seven times larger than the precision of those measures. This indicator does not suggest a group-level central tendency effect.

(3) The *ratee separation index*. This indicator connotes the number of statistically distinct levels of ratee performance. A low ratee separation index suggests a group-level central tendency effect.

Example: The ratee separation index of 10.1 suggests that there are about ten statistically distinct strata of ratee performance in this sample of ratees. (Note that the ratee separation index does not appear as part of *Facets* output. The

Table 5

Ratee Measurement Report Summary from an Analysis Using the Rating Scale Model—Central Tendency Simulation

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Avrage	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	PtBis	Num PERSON
28.1	10.0	2.8	2.81	-.43	.54	.9	-.3	.9	-.3	.16	Mean (Count: 300)
17.4	.0	1.7	1.74	4.13	.14	.5	1.1	.5	1.1	.29	S.D.

RMSE (Model) .56 Adj S.D. 4.09 Separation (7.29) Reliability (.98)
Fixed (all same) chi-square (13718.5) d.f.: 284 significance: (.00)

Note: In *Facets* output, the summary information appearing in this table would be included at the foot of the Ratee Measurement Report. The body of the Ratee Measurement Report would include a listing of ratees and detailed statistical information about each ratee's performance (i.e., a ratee performance measure, standard error of the measure, ratee fit statistics, single rater-rest of the raters correlation).

researcher will need to compute this index manually using the formula (4G + 1) / 3, where G is the ratee separation ratio, which is included as part of *Facets* output. In this example, [4 (7.29) + 1] / 3 = 10.1). There is no evidence here of a group-level central tendency effect.

(4) The *reliability of the ratee separation index.* This indicator is a measure of the spread of the ratee performance measures relative to their precision, showing the extent to which the raters have been able to distinguish reliably among the ratees in terms of their performance. A low ratee separation reliability index suggests a group-level central tendency effect.

Example: The ratee separation reliability is shown in the second line from the bottom of our Table 5. The high degree of ratee separation reliability (.98) implies that raters could reliably distinguish among the ratees. The ratees are well differentiated in terms of their levels of performance. Therefore, this indicator does not suggest a group-level central tendency effect in this data set.

Fit indices for the traits also provide useful information for the detection of a group-level central tendency effect. (The central tendency simulation data set we used had raters evaluating ratees on only a single trait, not multiple traits, so we will not provide an example of *Facets* output to show fit indices for traits. However, we include the following interpretive information for researchers who are analyzing data sets that include several trait scales.)

For each *trait* included in an analysis using the rating scale model or any of the hybrid models, the output from the analysis will provide measures of the consistency with which the raters used the scale to assign ratings on that particular trait (Table 7 of *Facets* output, the Trait Measurement Report). *Trait fit mean-square indices* that are significantly less than 1 indicate less variability than expected in the raters' ratings of the trait. For that particular trait, the low fit mean-square indices could be a signal that raters as a group may have overused one or more categories on the scale. In some cases, it may be that those categories were the middle categories of the scale, which would signal a central tendency effect in the ratings of the trait.

As a next step in the diagnostic process, the researcher can determine whether the raters overused the middle categories on an overfitting trait scale by conducting an analysis employing Hybrid Model #1 The output from such an analysis will include a series of tables, one for each trait, that summarize how the raters (as a group) used the categories on the scale. By reviewing these tables, the researcher can easily identify overfitting trait scales that show central tendency in the ratings (i.e., those scales in which there was a clumping of ratings at the center of the scale), but the researcher will not be able to determine which particular raters were most responsible for the central tendency effect. (The researcher would need to use Hybrid Model #3 to identify those raters.)

Individual-level statistical indicators. In some settings, ratees tend to differ a great deal in the levels of performance they demonstrate (i.e., the range of ratee performance measures is wide). In these situations, a rater whose ratings show central tendency will exhibit less variability than expected in his or her ratings, even after the particular ratees that rater rated have been taken into account. *Rater infit and outfit mean-square indices* are sometimes useful in these types of situations for detecting central tendency effects.

A short digression is warranted here because our experiences in analyzing a number of real and simulated sets of rating data have led us to be cautious when interpreting fit mean-square indices. In real data sets, we have frequently observed rater fit mean-square indices that are less than 1 for raters for whom the pattern of ratings implies central tendency (e.g., a "flat line" rating pattern such as [3,3,3,3,3,3]). However, in simulation studies, we have observed the opposite effect—fit mean-square indices that are greater than 1 for ratings simulated to exhibit a central tendency effect. We believe the explanation for this discrepancy lies in the variability of the traits.

Consider two hypothetical vectors of expected ratings, each for a single ratee who is rated on two different sets of ten traits (A and B) by a single "perfect" rater (i.e., a rater who introduces no rater effects).

$$A = [3,3,3,3,3,3,3,3,3,3]$$
$$B = [1,5,1,5,1,5,1,5,1,5]$$

Given the fact that a "perfect" rater assigned the ratings, the trait difficulties in set A must be very similar because the expected ratings are identical. On the other hand, the trait difficulties in set B must vary considerably because the expected ratings vary considerably.

Now consider two hypothetical vectors of assigned (rather than expected) ratings that two different raters assign for the ten traits in set A. In both cases, we assume that the two raters do not differ in severity.

$$A_1 = [3,3,3,3,3,3,3,3,3,3]$$
$$A_2 = [1,5,1,5,1,5,1,5,1,5]$$

Clearly, the ratings the first rater assigned (A_1) are very consistent with the vector of expected ratings for these traits (A), while the ratings the second rater assigned (A_2) are very inconsistent with that vector of expected ratings. As a result, the rater fit mean-square indices for the A_1 vector of observed ratings would be zero, indicating overfit of the observed ratings to the expected ratings. After examining the vector of ratings in A_1, we might conclude that the first rater exhibits a central tendency effect. However, that would be an incorrect conclusion to draw from this data. Rather, the first rater is actually rating extremely accurately. Hence, we would incorrectly conclude that the rater exhibits a central tendency effect. The fit mean-square indices for the second rater would be significantly greater than 1, rightly indicating misfit.

Now consider two hypothetical vectors of ratings that two different raters assign for the ten traits represented by the expected ratings in set B. Again, in both cases, we assume that the two raters do not differ in severity.

$$B_1 = [3,3,3,3,3,3,3,3,3,3]$$
$$B_2 = [1,5,1,5,1,5,1,5,1,5]$$

In this case, the ratings shown in vector B_2 are clearly accurate ratings—they perfectly match the expected ratings. As a result, the fit mean-square indices for the vector of observed ratings that the second rater assigned will equal zero. On the other hand, the first rater is clearly exhibiting a central tendency effect. In this case, the rater fit mean-square indices for this rater will probably be greater than 1. In analyses of operational rating data, central tendency frequently results in rater overfit (and thus rater fit mean-square indices that are less than one). However, it is important to remember that this will not always be the case, as we have just demonstrated. Some-

Table 6

Rater Measurement Report from an Analysis Using the Rating Scale Model—Central Tendency Simulation

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Avrage	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	PtBis	Nu	RATERS
817	285	2.9	2.87	-.09	.10	.8	-2	.8	-2	.95	1	1
833	285	2.9	2.94	-.24	.10	.7	-4	.6	-4	.96	2	2
795	285	2.8	2.79	.11	.10	.8	-2	.8	-2	.95	3	3
757	285	2.7	2.64	.46	.10	.9	-1	.9	-1	.94	4	4
757	285	2.7	2.64	.46	.10	.8	-2	.7	-3	.95	5	5
782	285	2.7	2.74	.23	.10	.8	-2	.8	-2	.96	6	6
837	285	2.9	2.95	-.28	.10	.8	-3	.7	-3	.95	7	7
854	285	3.0	3.02	-.44	.10	.7	-3	.7	-3	.95	8	8
823	285	2.9	2.90	-.15	.10	.8	-2	.7	-3	.95	9	9
814	285	2.9	2.86	-.07	.10	2.2	9	2.7	9	.84	10	10
806.9	285.0	2.8	2.83	.00	.10	.9	-1.6	.9	-1.9	.94	Mean (Count: 10)	
31.6	.0	.1	.12	.29	.00	.4	3.6	.6	3.7	.03	S.D.	

RMSE (Model) .10 Adj S.D. .28 Separation 2.86 Reliability .89
Fixed (all same) chi-square: 91.9 d.f.: 9 significance: .00

times the rater fit mean-square indices for raters exhibiting a central tendency effect will be greater than one. Therefore, we suggest that the researcher carefully examine vectors of observed ratings for all overfitting or misfitting raters before concluding that they are exhibiting a central tendency effect.

Example: The rater infit and outfit mean-square indices appear in columns 7 and 9 of our Table 6. (They are included in *Facets* output in Table 7, the Rater Measurement Report.) When we examine the fit mean-square indices for Raters 1 through 9, we see that they range from 0.6 to 0.9, which suggests that the ratings of these nine raters exhibit only small deviations from the MFRM expected ratings. (The expected value for these fit mean-square indices is 1.) By contrast, the fit mean-square indices associated with Rater 10, the rater simulated to exhibit central tendency, indicate that there are large differences between this rater's observed and expected ratings. The infit mean-square index is 2.2, while the outfit mean-square index is 2.7. However, it is important to note that the ratings that Rater 10 assigned do not exhibit randomness. Based on the "single rater—rest of the raters" (SR/ROR) correlation of .84 shown in column 11 of our Table 6, we can see that Rater 10 tends to rank order ratees in a manner that is consistent with the rank orderings of the other nine raters.[4] (Note that the SR/ROR correlation appears in the column labeled "PtBis" in *Facets* output.)

If there are individual raters whose fit mean-square indices suggest misfit, then the researcher can examine the scale category statistics tables from an analysis using Hybrid Model #2 or Hybrid Model #3 to gain an understanding of how each rater used each category on each trait scale. The scale category statistics table for a rater (Table 8 of *Facets* output) provides frequency counts showing how many times that rater used each category on the trait scale. The researcher can examine the frequency distribution for the ratings that each rater assigned on a given trait and easily identify those raters who exhibited central tendency by overusing the inner categories on the scale.

Table 7

Category Statistics from an Analysis Using Hybrid Model #2—Central Tendency Simulation

Rater 9

DATA Category Score	Counts Used	%	Cum. %	QUALITY CONTROL Avge Meas	Exp. Meas	OUTFIT MnSq	STEP CALIBRATIONS Measure	S.E.	EXPECTATION Measure at Category	-0.5	MOST PROBABLE from	.5 Cumul. Probabil. at	Cat PEAK Prob
0	29	10%	10%	-7.98	-7.95	1.0			(-7.68)		low	low	100%
1	34	12%	22%	-4.85	-4.88	.6	-6.56	.34	-5.34	-6.73	-6.56	-6.63	63%
2	58	20%	42%	-2.49	-2.39	.9	-4.12	.27	-2.67	-4.05	-4.12	-4.09	67%
3	59	21%	63%	.07	-.07	.9	-1.26	.23	.00	-1.29	-1.26	-1.28	63%
4	53	19%	82%	2.15	2.33	.9	1.25	.24	2.72	1.30	1.25	1.26	68%
5	28	10%	92%	5.14	4.92	.6	4.21	.29	5.35	4.12	4.21	4.16	61%
6	24	8%	100%	7.86	7.81	.7	6.47	.36	(7.62)	6.69	6.47	6.55	100%
									(Mean)		(Modal)	(Median)	

Rater 10

DATA Category Score	Counts Used	%	Cum. %	QUALITY CONTROL Avge Meas	Exp. Meas	OUTFIT MnSq	STEP CALIBRATIONS Measure	S.E.	EXPECTATION Measure at Category	-0.5	MOST PROBABLE from	.5 Cumul. Probabil. at	Cat PEAK Prob
0	6	2%	2%	-7.11	-9.39	2.4			(-11.35)		low	low	100%
1	37	13%	15%	-7.14*	-7.16	1.3	-10.28	.49	-8.01	-10.29	-10.28	-10.29	82%
2	69	24%	39%	-3.14	-3.40	2.0	-5.82	.29	-3.91	-5.83	-5.82	-5.83	77%
3	86	30%	69%	-.02	-.23	2.0	-2.05	.22	-.08	-2.02	-2.05	-2.04	78%
4	56	20%	89%	2.56	3.06	2.0	1.83	.23	3.68	1.81	1.83	1.81	76%
5	29	10%	99%	6.09	6.78	2.5	5.56	.33	8.12	5.59	5.56	5.56	87%
6	2	1%	100%	8.26	9.23	1.2	10.77	.78	(11.84)	10.77	10.77	10.76	100%
									(Mean)		(Modal)	(Median)	

Example: Our Table 7 presents the scale category statistics from an analysis using Hybrid Model #2 for two raters. (The summary statistics for Rater 9 are shown in the first table, while the summary statistics for Rater 10 are shown in the second table.) Notice that Rater 10 tended to use categories 2 through 4 much more frequently than the other rating categories—74% of this rater's ratings fell within this range, as shown in the third column. On the other hand, only 60% of the ratings of Rater 9 fell within this range.

There are several other pieces of information included in a rater's category statistics table from an analysis using Hybrid Models #2 or #3 that are useful for detecting an individual-level central tendency effect:

(1) *Rating scale category thresholds*. A rating scale category threshold denotes the point at which the probability curves for two adjacent scale categories cross (Linacre, 1999). Thus, the rating scale category threshold represents the point at which the probability is 50% of a ratee being rated in one or the other of these two adjacent categories, given that the ratee is in one of them (Andrich, 1998). If a rater exhibits a central tendency effect, then the rating scale category thresholds will be widely dispersed. Additionally, sometimes raters exhibiting central tendency will not use the outer categories of the trait scale. In those cases, there will be fewer category thresholds reported than when raters use all categories on the scale. Alternatively, there may also be evidence of threshold reversal in these outer categories; that is, the thresholds for those categories may not increase monotonically (i.e., the thresholds will not continue to get larger as one "goes up" the rating scale). When there are threshold reversals, one or more rating categories are observed with a very low probability.[5]

Example: Compare the rating scale category thresholds presented in our Table 7 for Raters 9 and 10. The thresholds are shown as "Measures" in column 8 in the *Step Calibrations* section of the table. (They appear in Table 8 of *Facets* output.) None of the rating scale category thresholds is reversed; they all increase monotonically for both raters. However, Rater 10 used two of the categories on the scale infrequently, assigning 8 ratings (3%) in categories 0 and 6. By contrast, 18% of Rater 9's ratings were assigned to these outer categories. In addition, the distance between the rating scale category thresholds is greater for Rater 10 than for Rater 9. For example, the average distance between category thresholds for Rater 10 is 4.21 logits, while the average distance between category thresholds for Rater 9 is 2.61 logits. These results suggest that Rater 9 included a more narrow range of ratee performance levels in each of the seven rating categories employed, while Rater 10 included a wider range of ratee performance levels in each of the rating scale categories.

(2) *Outfit mean-square indices for the rating scale categories*. For each rating category on a trait scale, *Facets* computes two ratee performance measures: 1) an "observed" ratee performance measure (i.e., the "Avge Meas"), and 2) an "expected" ratee performance measure (i.e., the "Exp. Meas"), which is the ratee performance measure the MFRM model would predict for that rating scale category if the data were to fit the model. (These two measures appear in columns 5 and 6 of our Table 7. They are included in Table 8 of *Facets* output.) When the observed ratee performance measure (the "Avge Meas") and the expected ratee performance measure (the "Exp. Meas") for a given rating scale category are close, then the outfit mean-square index for that category will be near the expected value of 1. The greater the discrepancy between the observed and expected ratee performance measures, the larger the rating scale category's outfit mean-square index will be. Consequently, when using the Hybrid models, it is important to examine the outfit mean-square indices for a rater's scale categories, since they may signal a central tendency effect.

Example: Notice that the outfit mean-square indices for Rater 10's scale categories in our Table 7 are much larger than the comparable indices for Rater 9 (shown in the *Quality Control* section of each table in column 7). The average outfit mean-square index for Rater 10 is 1.91, while the average outfit mean-square index for Rater 9 is 0.80.

The next step in understanding an individual rater's tendency to overuse the middle categories on a rating scale is to examine the Table of Misfitting Ratings that is part of the output from an analysis using the rating scale model or any of the hybrid models. The Table of Misfitting Ratings (Table 4 of *Facets* output) inventories the most surprising or unexpected ratings, based on differences between observed ratings and modeled expectancies. It pinpoints the particular ratings the rater gave that were unexpectedly high or low, taking into account that rater's overall level of severity and the ratings the ratee received from other raters.

Example: Our Table 8 presents a portion of the Table of Misfitting Ratings inventory from the central tendency simulation. (We used the rating scale version of the MFRM to run this analysis.) The expected and observed ratings for Rater 10 across eight ratees are shown. Here, we see that in each case Rater 10's observed ratings tend to be closer to the midpoint of the scale than the expected ratings for all eight ratees. Specifically, the values of the Observed—Expected ratings are positive for ratees for whom the rater gave a higher-than-expected rating, and these values are negative for ratees for whom the rater gave a lower-than-expected rating.

The researcher can also look at the category probability curves for individual raters that are included in *Facets* output (as Table 9 from a *Facets* analysis) to help in detecting central tendency effects. The researcher would need to use Hybrid Model #2 to obtain these curves if the raters are working with a single trait scale, or Hybrid Model #3 if the raters are working with multiple trait scales. The horizontal axis is the ratee performance scale; the vertical axis is the probability of observing a particular rating (from 0 to 1). A separate curve is produced for each of the rating scale categories. When we look at one of these figures, the chief concern is whether the rating scale categories are widely separated on the logit scale, or not. If the categories are widely spaced (and thus have very distinct peaks), then that may suggest that the rater was exhibiting central tendency. For example, Figure 3 shows the category probability curves from an analysis of ratings for Rater 9 who did not exhibit central tendency. Note that, although there is a separate peak for each of the categories on the trait scale, the individual category probability curves are fairly narrowly dispersed. Now compare Figure 3 to Figure 4. Figure 4 shows the category probability curves from an analysis for Rater 10 who exhibited central tendency. Note that there is a separate and distinct peak for each rating scale category, and those categories are widely dispersed

Table 8

Expected and Observed Ratings for Rater 10—Central Tendency Simulation

Ratee	Expected Rating for Rater 10	Observed Rating for Rater 10	Observed - Expected
1	0.2	2	1.8
2	1.0	3	2.0
3	1.6	4	2.4
4	2.3	5	2.7
5	3.9	2	-1.9
6	4.8	2	-2.8
7	5.4	3	-2.4
8	5.9	5	-0.9

across the ratee performance measure scale. In general, when a rater exhibits central tendency, it increases the probability of observing in the innermost categories of the scale, and that results in a wide separation of the category thresholds, particularly in the middle of the ratee performance measure distribution.

Randomness Effect

Conceptual definition. Within the context of a MFRM analysis, the randomness effect is defined as a rater's tendency to apply one or more trait scales in a manner inconsistent with the way in which the other raters apply the same scales. A rater who is experiencing randomness is overly inconsistent in the use of the scales, exhibiting more random variability than expected in his or her ratings, even after the performances of the particular ratees the rater evaluated have been taken into account. When one rater exhibits randomness and other raters do not, that rater will rank ratees in a different order than the other raters will.

The rater who exhibits randomness may have developed a different interpretation of the meaning of one or more of the trait scales (or of one or

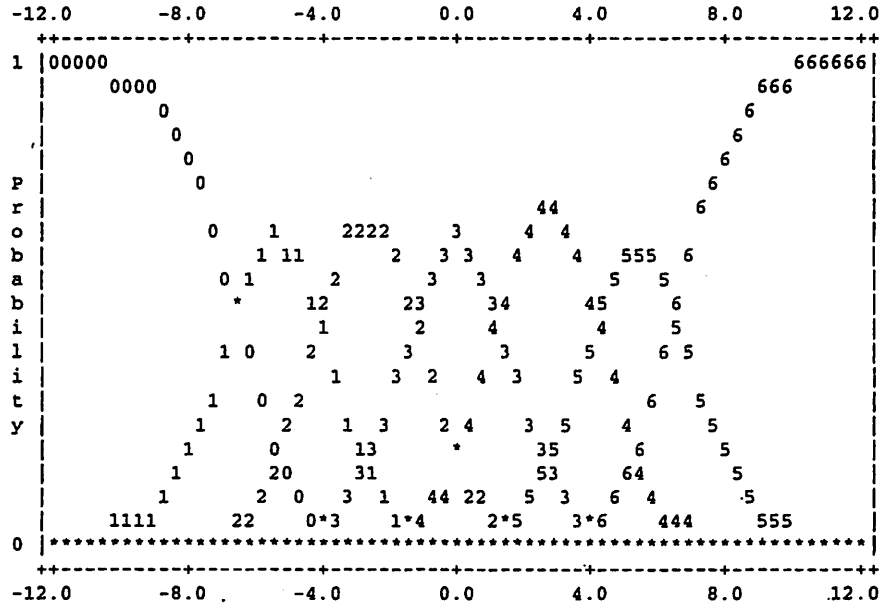

Figure 3. Category Probability Curves for "No Effect" Rater

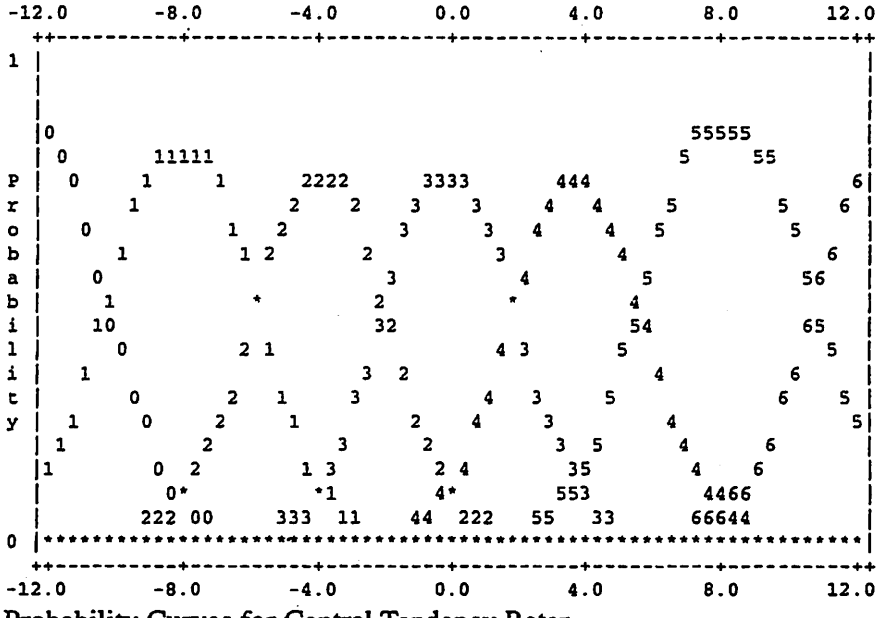

Figure 4. Category Probability Curves for Central Tendency Rater

more of the categories on a scale), causing the rater to use the scale categories in a different way than the other raters do. In some cases, the rater may not have sufficient background and/or training to be able to make the fine discriminations required in order to employ the trait scale(s) appropriately. Instead, the rater appears to assign ratings in an unreliable, haphazard fashion, using an approach to evaluating ratees that seems to bear little resemblance to the approach that other raters are employing.

Measurement models for detecting the randomness effect. Researchers using the rating scale model or any of the hybrid models to analyze their rating data will obtain the group- and individual-level statistical indicators described below.

Linacre (personal communication, January 20, 2003) has recently outlined an alternative approach to detecting randomness using *Facets*. The idea is to explicitly match the model used in the analysis to the aberrant rater behavior the researcher wants to detect. If one or more raters fit the model, then the researcher might suspect that those raters exhibit the aberrant behavior. In this case, to determine whether raters are exhibiting randomness, the researcher would anchor all ratees at the same level of performance (i.e., usually 0) and then run the analysis. Raters who show the best fit to this model are likely to be exhibiting randomness (i.e., their ratings are not related to the level of performance of the ratee).

Group-level statistical indicators. If most of the raters included in a MFRM analysis exhibit randomness in their ratings, then ratees will differ little in their levels of performance. As a result, it would be difficult to distinguish reliably among ratees. There are four group-level statistical indicators of randomness related to ratee performance that are included in the output from an analysis using the rating scale model or any of the hybrid models:

(1) A *fixed chi-square test* of the hypothesis that all ratees exhibit the same calibrated level of performance (i.e., that all ratees share the same performance measure, after accounting for measurement error). A nonsignificant chi-square value suggests a group-level randomness effect.

> *Example*: The results from the ratee fixed chi-square test are shown in the bottom line of our Table 9, the Ratee Measurement Report. (Note that what is shown here is the summary report that appears at the bottom of Table 7 in *Facets* output, not the entire Table 7.) The chi-square value of 9143.1 with 272 degrees of freedom is statistically significant ($p < .005$), suggesting that there is probably not a group-level randomness effect present in this simulation data set.

(2) The *ratee separation ratio*. This ratio is a measure of the spread of the ratee performance measures relative to the precision of those measures. A low ratee separation ratio suggests a group-level randomness effect.

> *Example*: The ratee separation ratio of 6.24 (shown in the second line from the bottom of our Table 9) indicates that the spread of the ratee performance measures is more than six times larger than the precision of those measures. This indicator does not suggest a group-level randomness effect.

(3) The *ratee separation index.* This indicator connotes the number of statistically distinct levels of ratee performance. A low ratee separation index suggests a group-level randomness effect.

> *Example*: The ratee separation index of 8.65 suggests that there are more than eight statistically distinct strata of ratee performance in this sample of ratees. (Note that the ratee separation index does not appear as part of *Facets* output. The researcher will need to compute this index manually using the formula $(4G + 1) / 3$, where G is the ratee separation ratio, which is included as part of *Facets* output. In this example, $[4 (6.24) + 1] / 3 = 8.65$). Again, there is no evidence here of a group-level randomness effect.

(4) The *reliability of the ratee separation index.* This indicator is a measure of the spread of the ratee performance measures relative to their pre-

cision, showing the extent to which the raters have been able to distinguish reliably among the ratees in terms of their performance. A low ratee separation reliability index suggests a group-level randomness effect.

> *Example*: The ratee separation reliability is shown in the second line from the bottom of our Table 9. The high degree of ratee separation reliability (.97) implies that raters could reliably distinguish among the ratees in terms of their performance. Therefore, this indicator does not suggest a group-level randomness effect in this data set.

Individual-level statistical indicators. For each rater included in an analysis using a rating scale model or any of the hybrid models, *Facets* provides measures of the consistency with which the rater has used the rating scales across traits to rate multiple ratees. A rater's fit indices will indicate the cumulative agreement between observed and expected ratings for that rater across all traits and ratees the rater evaluated. Raters showing a randomness effect in their ratings will have *rater infit and outfit mean-square indices* that are significantly greater than 1, suggesting that those raters may not have been able to differentiate reliably between ratee performances on the trait being measured. Instead, they may have assigned seemingly random ratings to many ratees.

Rater infit and outfit mean-square indices greater than 1 can signal other rater effects, as well. To eliminate other potential rater effects from consideration, the researcher may want to look for uncharacteristically low "single rater—rest of the raters" (SR/ROR) correlations for individual raters when compared to other raters' SR/ROR correlations. If a rater's SR/ROR correlation is considerably lower than the correlations for other raters, then that rater tends to rank ratees in a dif-

Table 9

Ratee Measurement Report Summary from an Analysis Using the Rating Scale Model—Randomness Simulation

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Avrage	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	PtBis	Num PERSON
28.1	10.0	2.8	2.81	-.26	.48	1.0	-.4	1.0	-.5	.15	Mean (Count: 300)
17.5	.0	1.8	1.74	3.13	.12	1.0	1.6	1.0	1.6	.29	S.D.

RMSE (Model) .50 Adj S.D. 3.09 Separation 6.24 Reliability .97
Fixed (all same) chi-square 9143.1 d.f.: 272 significance: .00

Note: In *Facets* output, the summary information appearing in this table would be included at the foot of the Ratee Measurement Report. The body of the Ratee Measurement Report would include a listing of the ratees and detailed statistical information about each ratee's performance.

Table 10

Rater Measurement Report from an Analysis Using the Rating Scale Model—Randomness Simulation

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Avrage	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	PtBis	Nu RATERS
793	273	2.9	2.91	-.07	.09	.7	-4	.7	-3	.95	1 1
809	273	3.0	2.98	-.20	.09	.6	-6	.5	-6	.96	2 2
771	273	2.8	2.82	.09	.09	.7	-4	.7	-3	.95	3 3
733	273	2.7	2.65	.38	.09	.7	-3	.8	-2	.94	4 4
733	273	2.7	2.65	.38	.09	.6	-4	.6	-4	.95	5 5
758	273	2.8	2.76	.19	.09	.7	-4	.6	-4	.96	6 6
813	273	3.0	3.00	-.23	.09	.6	-4	.7	-4	.95	7 7
830	273	3.0	3.07	-.35	.09	.7	-4	.7	-3	.94	8 8
799	273	2.9	2.94	-.12	.09	.6	-4	.7	-4	.94	9 9
792	273	2.9	2.91	-.07	.09	3.6	9	3.9	9	.70	10 10
783.1	273.0	2.9	2.87	.00	.09	.9	-3.2	1.0	-2.8	.92	Mean (Count: 10)
31.6	.0	.1	.14	.24	.00	.9	4.1	1.0	4.1	.08	S.D.

RMSE (Model) .09 Adj S.D. .22 Separation 2.55 Reliability .87
Fixed (all same) chi-square: 74.8 d.f.: 9 significance: .00

ferent order than other raters. If the rater's SR/ROR correlation is consistent with that of other raters, then another (systematic) type of rater effect might be operating (i.e., central tendency).

> *Example*: The rater infit and outfit mean-square indices appear in columns 7 and 9 of our Table 10. (They are included in *Facets* output in Table 7, the Rater Measurement Report.) When we examine these indices, we see that one rater exhibits a randomness effect. Rater 10 shows considerable misfit, having infit and outfit mean-square indices of 3.6 and 3.9, respectively. The rater's "single rater—rest of the raters" (SR/ROR) correlation (.70) provides further evidence of randomness in the ratings. (Note that in the *Facets* output, this indicator is shown in column 11 and is labeled "PtBis.") Because this rater's SR/ROR correlation is considerably lower than the other raters' SR/ROR correlations, we would conclude that not only does Rater 10 exhibit randomness in the ratings assigned, but that rater also tends to rank ratees in a different order than other raters do.

Halo Effect

Conceptual definition. Within the context of a MFRM analysis, the halo effect is defined as a rater's tendency to assign ratees similar ratings on conceptually distinct traits. A rater who exhibits halo cannot readily distinguish among those traits and thus gives a ratee similar ratings across those traits.

Measurement models for detecting the halo effect. Researchers using the rating scale model or any of the hybrid models to analyze their rating data will obtain the group- and individual-level statistical indicators described below.

Linacre (personal communication, January 20, 2003) has recently outlined an alternative approach to detecting halo using *Facets*. The idea is to explicitly match the model used in the analysis to the aberrant rater behavior the researcher wants to detect. If one or more raters fit the model, then the researcher might suspect that those raters exhibit the aberrant behavior. In this case, to determine whether raters are exhibiting a halo effect, the researcher would anchor all traits at the same difficulty (i.e., usually 0) and then run the analysis. Raters who show the best fit to this model are likely to be exhibiting halo.

Group-level statistical indicators. When most of the raters exhibit halo effects, ratings are similar across traits for ratees. As a result, the traits would appear to differ little in terms of their difficulties when the traits do, indeed, differ in their true difficulties. The apparent lack of difference in trait difficulty can be a reflection of the raters' inability to readily distinguish among the traits, which may lead them to give each ratee similar ratings across each and every trait. However, it is important to emphasize that the appearance of no difference in trait difficulty does not necessarily imply that the raters exhibited halo. Traits can be conceptually distinct but not differ in difficulty.

The output from an analysis using the rating scale model or any of the hybrid models includes several group-level indicators of halo that focus on the measurement of the traits:

(1) A *fixed chi-square test* of the hypothesis that all traits are of the same calibrated level of difficulty (i.e., they share the same difficulty measure, after accounting for measurement error). A non-significant chi-square value may suggest a pervasive trend toward halo in the ratings of all raters. (However, a non-significant chi-square value may also simply indicate that the traits are not significantly different in terms of their difficulties.)

> *Example*: The results from the fixed chi-square test for the traits are shown in the bottom line of our Table 11. (These results would appear in Table 7 of *Facets* output.) The chi-square value of 3240.5 with 3 degrees of freedom is statistically significant ($p < .005$), indicating that at least two traits are significantly different in terms of their

difficulty. These results suggest that there is not a group-level halo effect present in this simulation data set.

(2) The *trait separation ratio*. This ratio is a measure of the spread of the trait difficulty measures relative to the precision of those measures. A low trait separation ratio suggests halo in the ratings.

Example: The trait separation ratio of 28.54 (shown in the second line from the bottom of our Table 11) signals that the spread of the trait difficulty measures is about 29 times larger than the precision of those measures. This indicator does not suggest a group-level halo effect.

(3) The *trait separation index*. This indicator connotes the number of statistically distinct levels of trait difficulty among the traits included in the analysis. A low trait separation index suggests halo in the ratings.

Example: The trait separation index of 38.39 suggests that there are over 38 statistically distinct strata of trait difficulty in this sample of traits. (Note that the trait separation index does not appear as part of *Facets* output. The researcher will need to compute this index manually using the formula $(4G + 1) / 3$, where G is the trait separation ratio, which is included as part of *Facets* output. In this example, $[4 (28.54) + 1] / 3 = 38.39$). There is no evidence here of a group-level halo effect.

It should be noted that the trait separation index may be large when the number of ratees (and/or raters) is large. This is because the error variance, as depicted by the average standard error of trait difficulty measures, will decrease as the number of ratees (and/or raters) increase (i.e., each trait is measured more precisely under these conditions). This may result in trait separation indices that indicate a greater number of statistically distinct strata than there are traits in the analysis. The interpretation of such a situation would be fairly straightforward—that is, the spread of the trait difficulty measures is considerably greater than the precision of those measures.

(4) The *reliability of the trait separation index*. This indicator provides information about how well the traits are separated in terms of their difficulty, showing the extent to which the raters have been able to distinguish among the traits. A low trait separation reliability index suggests a halo effect.

Example: The trait separation reliability appears in the second line from the bottom of our Table 11. The high degree of trait separation reliability (1.00) implies that raters could reliably distinguish among the traits. Therefore, this indicator does not suggest a group-level halo effect in this data set.

Individual-level statistical indicators. For each rater included in an analysis using the rating scale model or any of the hybrid models, *Facets* provides measures of the consistency of the rater's ratings with the MFRM expected ratings. A rater's

Table 11

Trait Measurement Report from an Analysis Using the Rating Scale Model—Halo Simulation

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Avrage	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	PtBis	N TRAIT
8395	2860	2.9	2.95	-.26	.02	.85	-5.7	.89	-3.3	.65	1 1
7298	2860	2.6	2.48	.37	.02	1.05	1.7	1.01	.1	.64	2 2
6463	2860	2.3	2.12	.86	.02	1.06	2.1	1.08	1.9	.64	3 3
9631	2860	3.4	3.47	-.98	.02	1.04	1.5	1.03	.9	.63	4 4
7946.8	2860.0	2.8	2.76	.00	.02	1.00	-.1	1.00	-.1	.64	Mean (Count: 4)
1189.5	.0	.4	.51	.69	.00	.09	3.3	.07	2.0	.01	S.D.

RMSE (Model) .02 Adj S.D. .69 Separation 28.54 Reliability 1.00
Fixed (all same) chi-square: 3240.5 d.f.: 3 significance: .00

fit indices indicate the cumulative agreement between observed and expected ratings across all traits and ratees the rater evaluated.

When trait difficulties vary little, then MFRM expected ratings would also differ little. As a result, a rater who exhibits a halo effect will assign nearly identical ratings on all traits for each ratee. Consequently, the rater's ratings will exhibit little deviation from the expected ratings, even after the particular ratees the rater evaluated have been taken into account. In this situation, raters showing a halo effect in their ratings will have *rater infit and outfit mean-square indices* that are significantly less than one, suggesting that those raters may not have been able to differentiate reliably between conceptually distinct traits. Instead, they may have assigned similar ratings to many ratees across a number of traits. (Again, remember that traits can be similar in difficulty and still be conceptually distinct, so rater overfit does not necessarily signal halo in the ratings.)

Alternatively, when trait difficulties vary, then MFRM expected ratings will show greater variability. As a result, the ratings of raters who exhibit halo effects (and assign similar ratings across traits for a particular ratee) will be very different from the expected ratings. This will result in rater infit and outfit mean-square indices that are significantly greater than one. In either case, a researcher should inspect vectors of observed ratings when rater fit indices are not close to one, because interpretation of the fit indices is not straightforward, but rather context bound.

Example: As shown in our Table 12 (Table 7 of *Facets* output), it is clear that Rater 10 exhibits greater misfit when compared to the other raters. The infit and outfit mean-square indices for Rater 10 (1.22 and 1.34, respectively, as shown in columns 7 and 9) are larger than those for the remaining raters. It is also important to note that Rater 10 does not exhibit an undue amount of randomness. The "single rater—rest of the raters" (SR/ROR) correlation for Rater 10 is of about the same magnitude as the (SR/ROR) correlations for the remaining raters (i.e., 0.63 for Rater 10, as compared to 0.64-0.65 for the other raters). (Note that the SR/ROR correlations are shown in column 11 as "PtBis" in *Facets* output.)

To determine whether raters may have assigned similar ratings to many ratees across a number of traits, the researcher may want to look for patterns in each overly consistent rater's ratings to see how many times the rater gave a string of identical ratings across traits to a ratee. For example, suppose a rater evaluated each ratee on four traits using three-category rating scales. In this instance, the researcher would tally the number of times the rater assigned ratees four 3's, four 2's, or four 1's. The next step would be to determine what percent of the total number of ratees the rater evaluated these ratees represented. That is, if the rater evaluated 50 ratees, in what percent of those cases did the rater assign ratees identical ratings across all traits?

Table 12

Rater Measurement Report from an Analysis Using the Rating Scale Model—Halo Simulation

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Avrage	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	PtBis	Nu	RATERS
3196	1144	2.8	2.77	-.03	.04	.91	-2.2	.89	-2.0	.65	1	1
3253	1144	2.8	2.84	-.11	.04	.97	-.6	.93	-1.2	.65	2	2
3092	1144	2.7	2.66	.13	.04	.98	-.5	.91	-1.5	.65	3	3
3002	1144	2.6	2.57	.26	.04	1.00	.0	1.16	2.5	.64	4	4
2948	1144	2.6	2.51	.34	.04	.93	-1.7	.86	-2.3	.65	5	5
3076	1144	2.7	2.65	.15	.04	.93	-1.6	.84	-2.8	.65	6	6
3266	1144	2.9	2.85	-.13	.04	1.05	1.0	1.04	.7	.64	7	7
3332	1144	2.9	2.92	-.22	.04	1.03	.8	1.05	.9	.64	8	8
3212	1144	2.8	2.79	-.05	.04	.99	-.1	1.01	.1	.64	9	9
3410	1144	3.0	3.00	-.34	.04	1.22	4.6	1.34	5.2	.63	10	10
3178.7	1144.0	2.8	2.76	.00	.04	1.00	.0	1.00	.0	.64	Mean (Count: 10)	
139.3	.0	.1	.15	.20	.00	.08	1.9	.15	2.4	.01	S.D.	

RMSE (Model) .04 Adj S.D. .20 Separation 5.23 Reliability .96
Fixed (all same) chi-square: 283.2 d.f.: 9 significance: .00

Alternatively, a bias-interaction analysis can be performed in which a Rater x Trait interaction term is estimated. In this case, the bias interaction term would indicate the degree to which the ratings produced for a particular Rater x Trait combination deviate from the expectations produced using the model depicted in Equation 1 (see Part I of this paper). The interaction term can be standardized by dividing the Rater x Trait estimate by its standard error, and statistically significant misfit for the particular Rater x Trait combination would be indicated by absolute values of the standardized index that exceed 2.

If the resulting z-score for a given Rater x Trait interaction is greater than 2, then the rater was more severe than expected when rating that particular trait. By contrast, if the z-score is less than –2, then the rater was more lenient than expected when rating that particular trait. These indices can be used to identify individual raters who exhibit misfit from expected ratings. However, to determine whether that misfit is the result of a halo effect, the researcher still needs to examine the observed and expected ratings for those raters who are flagged by the standardized indices.

Example: Our Table 13 presents a portion of the output from the Rater x Trait bias interaction analysis for the halo simulation. (Note that the results from bias interaction analyses are reported in Table 13 of *Facets* output.) The first four lines of our Table 13 show summary statistics for the ratings of Rater 9 ("RAT 9") on the four traits ("TR 1, 2, 3, 4"). Lines 5-8 show summary statistics for the ratings of Rater 10 ("RAT 10") on the same four traits.

The eight z-score summary statistics for Rater x Trait interactions involving Raters 9 and 10 are shown in column 7 of our Table 13. Rater 9 shows no statistically significant interactions for any of the traits. (The magnitudes of the Rater x Trait z-scores for the remaining raters were similar to those observed for Rater 9. Recall that these nine raters were modeled to exhibit no rater effects.) By contrast, we modeled Rater 10 to exhibit a halo effect, and this rater's z-scores are all statistically significant.

Comparing the observed and expected scores shown in columns 1 and 2 of our Table 13, we see that Rater 9 assigned ratings that were fairly consistent with the expected ratings. By contrast, Rater 10 assigned lower-than-expected ratings for Traits 1 and 4 and higher-than-expected ratings for Traits 2 and 3.

In order to understand the implications of these findings, it is important to examine them in the context of what we have learned about the difficulties of the traits (refer to our Table 11). Trait 3 was

Table 13

Rater x Trait Bias Interaction Report from an Analysis Using the Rating Scale Model—Halo Simulation

Obsvd Score	Exp. Score	Obsvd Count	Obs-Exp Average	Bias+ Model Measure	S.E.	Z-Score	Infit MnSq	Outfit MnSq	Sq	N	TR	measr	Nu	RAT	measr
860	848.0	286	.04	-.07	.08	-.91	.9	.9	33	1	1	-.26	9	9	-.05
730	738.1	286	-.03	.05	.08	.62	1.0	1.0	34	2	2	.37	9	9	-.05
632	654.3	286	-.08	.13	.08	1.72	1.0	1.0	35	3	3	.86	9	9	-.05
990	971.6	286	.06	-.11	.08	-1.41	1.0	1.0	36	4	4	-.98	9	9	-.05
844	898.1	286	-.19	.31	.08	4.11	.9	1.2	37	1	1	-.26	10	10	-.34
853	788.0	286	.23	-.37	.08	-4.93	.9	.9	38	2	2	.37	10	10	-.34
861	703.3	286	.55	-.91	.08	-12.02	.8	.9	39	3	3	.86	10	10	-.34
852	1020.6	286	-.59	.98	.08	12.94	.8	.8	40	4	4	-.98	10	10	-.34
Obsvd Score	Exp. Score	Obsvd Count	Obs-Exp Average	Bias+ Model Measure	S.E.	Z-Score	Infit MnSq	Outfit MnSq	Sq	N	TR	measr	Nu	RAT	measr
794.7	794.7	286.0	.00	.00	.08	.01	1.0	1.0	Mean (Count: 40)						
130.7	123.9	.0	.15	.24	.00	3.19	.1	.2	S.D.						

the most difficult trait—the hardest for ratees to get high ratings on (i.e., its trait difficulty measure is 0.86 logits). While the other raters tended to assign lower ratings on average to ratees on this trait, it appears that Rater 10 did not follow suit. The results from the bias interaction analysis indicate that Rater 10 tended to assign *higher* ratings than would have been expected, given how the other raters used the trait scale. According to the results shown in our Table 11, Trait 4 was the easiest for ratees to get high ratings on (i.e., its trait difficulty measure is –0.98 logits). While the other raters tended to assign higher ratings on average to ratees on this trait, it appears from the bias interaction analysis that Rater 10 did the opposite. This rater tended to assign *lower* ratings than would have been expected, given how the other raters used this trait scale. This evidence would suggest that Rater 10 was exhibiting halo. To verify this, the researcher should compare the observed and expected ratings for raters who exhibit large z-scores, as we demonstrated in the discussion of our Table 8.

Differential Leniency/Severity Effect

Conceptual definition. Within the context of a MFRM analysis, differential rater severity is defined as a rater's tendency to assign ratings to a particular group of ratees that are, on average, lower than the measurement model would expect for that group, given other raters' ratings of the group (i.e., the rater shows bias in the ratings of the group). Similarly, differential rater leniency is defined as a rater's tendency to assign ratings to a group of ratees that are, on average, higher than the measurement model would expect for that group, given other raters' ratings of the group.

Measurement models for detecting the differential leniency/severity effect. Researchers who study differential leniency/severity effects (e.g., across various subgroups of ratees) use MFRM models in which they specify *a priori* one or more bias interaction analyses.

Linacre (personal communication, January 20, 2003) has recently outlined an alternative approach to detecting differential leniency/severity using *Facets*. The idea is to explicitly match the model used in the analysis to the aberrant rater behavior the researcher wants to detect. If one or more raters fit the model, then the researcher might suspect that those raters exhibit the aberrant behavior. In this case, to determine whether raters are exhibiting differential leniency/severity related to ratee subgroup, the researcher would include "subgroup" as a dummy facet (i.e., included for classification, not for measurement) in the analysis to investigate rater—subgroup interactions using "B" in the "Models =" statements. For example, the researcher might include gender as a dummy facet to determine whether there is evidence of rater—gender subgroup interactions. Raters who show the best fit to these models are likely to be exhibiting differential leniency/severity related to gender. That is, the level of leniency/severity the raters exercise varies, depending upon the gender of the ratee they are evaluating.

Group-level statistical indicators. The output from an analysis that employs an extension of the many-facet rating scale model or any of the hybrid models, which includes a facet for the groups to be compared, will contain limited information relevant to detecting group-level differential leniency/severity among the raters. However, listed below are several potentially useful indicators:

(1) The *fixed chi-square test* of the hypothesis that all groups of ratees are of the same calibrated level of performance (i.e., that they share the same average measure, after accounting for measurement error). In situations in which, based on past research, the researcher has prior knowledge about whether the average measures of two or more ratee groups should differ, the fixed chi-square test may be useful in determining whether the raters exhibited a group-level differential leniency/severity effect. For example, if past research suggests that the ratee groups should share

the same average measure, but the fixed chi-square test indicates that they do *not*, then there is reason to suspect that the raters exhibited a group-level differential leniency/severity effect. On the other hand, if past research suggests that the ratee groups should *not* share the same average measure, but the fixed chi-square test indicates very small differences between groups, there is also reason to suspect that the raters may have exhibited a group-level differential leniency/severity effect. Note that this index may be interpreted as an indicator of group-level rater differential leniency/severity only if the researcher has prior knowledge to indicate that there is a true difference in level of performance between ratee groups—*a priori* knowledge that may not be available to the researcher.

Example: The results from the fixed chi-square test for the two ratee groups (i.e., males = 1, females = 2) are shown in the bottom line of our Table 14. (Note that these results would be reported in Table 7 of *Facets* output.) The chi-square value of 41.3 with 1 degree of freedom is statistically significant ($p < .005$), indicating that the average measures for males and females are different. In light of the fact that the two groups were originally modeled to exhibit different average measures, these results may or may not suggest a group-level differential leniency/severity effect, depending on the magnitude of this modeled difference. If there were no difference, given that we expected one, then we would have evidence of bias. However, given that there is a difference, and that we expected one, we must compare the magnitude of that difference to the expected magnitude. In this case, the two groups were modeled to exhibit a raw score difference of about .2 points. As is evidenced by the observed average ratings (column 3 of Table 14), the observed logit difference is comparable to this effect size. Hence, we would not conclude that the observed difference is indicative of differential leniency/severity because it is comparable to the expected difference in magnitude.

(2) The *group separation index*. This indicator connotes the number of statistically distinct levels of performance among the ratee groups included in the analysis. A small group separation index would suggest differential leniency/severity in our simulation example because the ratee groups were modeled to exhibit different average measures.

Example: The group separation index of 6.24 (computed from the group separation ratio of 4.43 that is shown in the second-to-last line of our Table 14) suggests that the ratee group measures are different enough such that it is possible to distinguish between more than six distinct strata. That is, in this sample, it is possible to detect six strata, even though there are only two groups, because the spread of ratee group measures is large relative to the precision of those measures. Given that the ratee groups were modeled to be different, there is little evidence here that raters exhibited

Table 14

Group Measure Report Summary from an Analysis Using the Rating Scale Model—Differential Severity Simulation

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Avrage	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	PtBis	N	GENDER
4227	1410	3.0	3.02	-.22	.05	1.01	.2	1.02	.4	.67	1	1
3415	1250	2.7	2.88	.22	.05	.89	-2.6	.89	-2.5	.67	2	2
3821.0	1330.0	2.9	2.95	.00	.05	.95	-1.2	.95	-1.1	.67	Mean (Count:2)	
406.0	80.0	.1	.07	.22	.00	.06	1.5	.07	1.5	.00	S.D.	

RMSE (Model) .05 Adj S.D. .21 Separation 4.43 Reliability .95
Fixed (all same) chi-square: 41.3 d.f.: 1 significance: .00

a group-level differential leniency/severity effect, since that would have resulted in the two groups having average measures that were not significantly different.

(3) The *reliability of the group separation index*. This indicator provides information about how well the ratee groups are separated in terms of their performances. In this example, a low group separation reliability index suggests a differential leniency/severity effect.

Example: The group separation reliability is shown in the second line from the bottom of our Table 14. The high degree of group separation reliability (.95) implies that raters, on average, reliably distinguished between the ratee groups—as they should have, given the fact that group differences were modeled to exist. Therefore, this indicator does not suggest a group-level differential leniency/severity effect in this data set.

Individual-level statistical indicators. A starting point for detecting individual-level differential leniency/severity involves examining the rater severity measures to determine whether some raters tend to rate more severely overall than other raters. For each rater included in a MFRM analysis, the output from the analysis provides a measure (in logits) of the level of severity that each rater exercised and the standard error of each severity estimate, indicating the precision with which a rater's severity has been measured.

Example: Our Table 15 presents the rater severity measures from the differential leniency/severity analysis. (Note that these results would be reported in Table 7 of *Facets* output.) The larger the rater severity measure, the more severe the rater. Here, we see that Rater 10 is more severe than the other raters. Rater 10 has a severity measure that is 8 standard errors from the mean of the group (0.00). (To obtain 8, divide 0.88 (the severity measure for Rater 10) by 0.11 (the model standard error).) More important, however, is the fact that Rater 10 has large fit mean-square indices (about 1.4) when compared to the other nine raters in this data set (see columns 7 and 9). Given these results, we need to examine other diagnostic indicators to gain a better understanding of how this rater is performing differently from other raters. Is this rater differentially severe, or is the rater exercising a consistent level of severity across different ratee groups?

To identify raters who are differentially severe or lenient, a bias-interaction analysis can be performed in which a Rater x Group interaction term is estimated using a model similar to the

Table 15

Rater Measurement Report from an Analysis Using the Rating Scale Model—Differential Severity Simulation

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Avrage	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	PtBis	Nu	RATERS
782	266	2.9	3.02	-.20	.11	.96	-.3	.97	-.2	.67	1	1
810	266	3.0	3.12	-.51	.11	.88	-1.3	.84	-1.6	.68	2	2
752	266	2.8	2.91	.14	.11	.92	-.8	.87	-1.4	.68	3	3
718	266	2.7	2.78	.52	.11	1.00	.0	1.08	.7	.67	4	4
730	266	2.7	2.83	.38	.11	.88	-1.4	.87	-1.3	.68	5	5
770	266	2.9	2.98	-.07	.11	.91	-1.0	.86	-1.4	.68	6	6
796	266	3.0	3.07	-.36	.11	.87	-1.4	.88	-1.3	.67	7	7
821	266	3.1	3.16	-.64	.11	.80	-2.4	.85	-1.5	.68	8	8
777	266	2.9	3.00	-.14	.11	.91	-1.0	.91	-.9	.67	9	9
686	266	2.6	2.65	.88	.11	1.42	4.2	1.43	3.8	.66	10	10
764.2	266.0	2.9	2.95	.00	.11	.96	-.6	.96	-.5	.67	Mean (Count: 10)	
40.5	.0	.2	.15	.45	.00	.16	1.7	.17	1.6	.00	S.D.	

RMSE (Model) .11 Adj S.D. .44 Separation 4.17 Reliability .95
Fixed (all same) chi-square: 182.7 d.f.: 9 significance: .00

one depicted in Equation 15 (see Part I of this paper). In this case, the bias interaction term indicates the degree to which the ratings produced for a particular Rater x Group combination deviate from the expected ratings produced using the model depicted in Equation 1 (see Part I of this paper). The interaction term can be standardized by dividing the Rater x Group measure by its standard error, and statistically significant misfit for the particular Rater x Group combination can be indicated by absolute values of the standardized index that exceed 2. If the resulting z-score for a given Rater x Group interaction is greater than 2, then the rater was more severe than expected when rating that particular ratee group. By contrast, if the z-score is less than -2, then the rater was more lenient than expected when rating that particular ratee group.

Example: Our Table 16 displays the Rater x Gender Bias Interaction Report from the differential severity simulation. (Note that these results would appear in Table 13 of *Facets* output.) This table contains summary statistics for each of the 10 raters' ratings of each group (1 = males, and 2 = females). For example, line 1 shows summary statistics for the ratings of Rater 1 ("RA" 1) for Group 1 ("G" 1), and line 2 shows summary statistics for the ratings of that rater for Group 2.

The z-score summary statistics for the various Rater x Gender interactions are shown in column 7. Comparing the observed and expected scores shown in columns 1 and 2, we see that Rater 10 tended to rate males (Group 1) more leniently than expected (z-score = -7.71, $p < .01$), while Rater 10 tended to rate females (Group 2) more severely than expected (z-score = 8.25, $p < .01$). Based on these findings, we would conclude that Rater 10 displayed differential severity. It is interesting to note that Rater 6 also exhibits apparent differential severity for females (z-score = -2.13, $p < .01$), which is a Type I error, given that Rater 6 was not modeled to exhibit

Table 16

Rater x Gender Bias Interaction Report from an Analysis Using the Rating Scale Model—Differential Severity Simulation

Obsvd Score	Exp. Score	Obsvd Count	Obs-Exp Average	Bias Measure	Model S.E.	Z-Score	Infit MnSq	Outfit MnSq	Sq	N	G	measr	Nu	RA	measr
425	432.1	141	-.05	.15	.15	1.04	1.0	1.0	1	1	1	-.22	1	1	-.20
357	349.9	125	.06	-.17	.15	-1.10	.9	1.0	2	2	2	.22	1	1	-.20
441	447.0	141	-.04	.13	.15	.86	.9	.9	3	1	1	-.22	2	2	-.51
369	362.9	125	.05	-.14	.15	-.93	.8	.8	4	2	2	.22	2	2	-.51
412	416.3	141	-.03	.09	.15	.62	1.0	1.0	5	1	1	-.22	3	3	.14
340	335.8	125	.03	-.10	.15	-.64	.8	.7	6	2	2	.22	3	3	.14
389	398.2	141	-.07	.20	.15	1.35	1.0	1.1	7	1	1	-.22	4	4	.52
329	320.0	125	.07	-.22	.16	-1.40	.9	1.0	8	2	2	.22	4	4	.52
401	404.6	141	-.03	.08	.15	.52	1.0	.9	9	1	1	-.22	5	5	.38
329	325.6	125	.03	-.08	.16	-.53	.7	.8	10	2	2	.22	5	5	.38
412	425.8	141	-.10	.29	.15	2.00	.9	.9	11	1	1	-.22	6	6	-.07
358	344.3	125	.11	-.33	.15	-2.13	.8	.8	12	2	2	.22	6	6	-.07
437	439.6	141	-.02	.05	.15	.37	.8	.8	13	1	1	-.22	7	7	-.36
359	356.4	125	.02	-.06	.15	-.40	1.0	.9	14	2	2	.22	7	7	-.36
451	452.8	141	-.01	.04	.15	.26	1.0	1.1	15	1	1	-.22	8	8	-.64
370	368.1	125	.02	-.05	.15	-.29	.6	.6	16	2	2	.22	8	8	-.64
425	429.5	141	-.03	.09	.15	.65	.9	1.0	17	1	1	-.22	9	9	-.14
352	347.5	125	.04	-.11	.15	-.69	.9	.8	18	2	2	.22	9	9	-.14
434	381.3	141	.37	-1.12	.15	-7.71	1.0	.9	19	1	1	-.22	10	10	.88
252	305.0	125	-.42	1.33	.16	8.25	.9	.8	20	2	2	.22	10	10	.88
382.1	382.1	133.0	.00	.00	.15	.00	.9	.9	Mean (Count: 20)						
48.5	45.3	8.0	.14	.42	.01	2.71	.1	.1	S.D.						

Fixed (all = 0) chi-square: 146.5 d.f.: 20 significance: .00

differential severity. However, the effect is considerably smaller for Rater 6 than for Rater 10 (about 0.3 logits for Rater 6, versus about 1.2 logits for Rater 10).

Our Table 17 provides a summary listing of all the group- and individual-level statistical indicators that we have discussed for the five rater effects. The interested reader may want to keep this

Table 17

Group- and Individual-Level Statistical Indicators Obtained from Facets *for Five Rater Effects*

Rater Effect	Statistical Indicators
Leniency/Severity	**Group-level indicators:** Frequency counts indicating how many times the raters (as a group) used each scale category (across all traits) Fixed chi-square test for raters Rater separation ratio, index, reliability **Individual-level indicators:** Locations of the individual rater severity measures on the "All Facet Vertical Rulers" (i.e., variable map) Frequency counts indicating how many times each rater used each category on each trait scale Rater severity measures Rater "fair averages"
Central Tendency	**Group-level indicators:** Frequency counts indicating how many times the raters (as a group) used each scale category (across all traits) For each trait, frequency counts indicating how many times the raters (as a group) used each scale category Fixed chi-square test for ratees Ratee separation ratio, index, reliability Trait fit mean-square indices **Individual-level indicators:** Frequency counts indicating how many times each rater used each category on each trait scale Rater fit mean-square indices Rating scale category thresholds Rating scale category outfit mean-square indices Table of Misfitting Ratings Category Probability Curves for each rater
Randomness	**Group-level indicators:** Fixed chi-square test for ratees Ratee separation ratio, index, reliability **Individual-level indicators:** Rater fit mean-square indices "Single rater—rest of the raters" (SR/ROR) correlations
Halo	**Group-level indicators:** Fixed chi-square test for traits Trait separation ratio, index, reliability **Individual-level indicators:** Rater fit mean-square indices # of times a rater assigned a string of identical ratings across traits to ratees *z*-scores from a Rater x Trait bias interaction analysis
Differential Leniency/Severity	**Group-level indicators:** Fixed chi-square test for groups Group separation ratio, index, reliability **Individual-level indicators:** *z*-scores from a Rater x Group bias interaction analysis

summary table close at hand when examining *Facets* output, since the table may prove useful as a quick reminder of those particular indicators that are most important for detecting each rater effect.

Other Approaches to Detecting Rater Effects

In this section of the paper, we will briefly describe other statistical procedures that researchers have used to detect and measure rater effects. It is our hope that by becoming aware of important and influential literature on this topic, readers will gain an appreciation for the diversity of psychometric perspectives that researchers bring to bear on their work.

As we noted earlier, researchers working from a classical test theory perspective have used a variety of statistical procedures to study rater effects. Some have calculated the means and standard deviations of trait ratings, searching for evidence in these indices of various rater effects (e.g., Bernardin and Pence, 1980; Borman, 1977; Latham, Wexley, and Pursell, 1975; Murphy and Anhalt, 1992). Still others have inspected the intercorrelations among ratings across traits (e.g., Keaveny and McGann, 1975; Pulakos, Schmitt, and Ostroff, 1986). Researchers have also used confirmatory factor analysis (CFA) employing various factor analytic models, including correlated traits, correlated methods (CTCM) (Kenny and Kashy, 1992; Widaman, 1985), the correlated uniquenesses CFA method (CU-CFA)(Marsh, 1989; Marsh and Bailey, 1991; Scullen, 1999; Scullen, Mount, and Goff, 2000), and maximum-likelihood CFA methods (O'Grady and Medoff, 1991).

Researchers have employed Guilford's (1954) analysis of variance approach to examine main effects and interactions involving raters (e.g., Hedge and Kavanagh, 1988; Hill, O'Grady, and Price, 1988). Other researchers have investigated rater effects using generalizability theory (e.g., Brennan, 1983; Crocker and Algina, 1986; Cronbach, Gleser, Nanda, and Rajaratnam, 1972; Murphy and DeShon, 2000b; Shavelson, Webb, and Rowley, 1989). Generalizability theory employs a more sophisticated and versatile variance estimation approach that goes beyond Guilford's analysis of variance model, both conceptually and methodologically. (For interested readers, Linacre (1996) compares the capabilities of generalizability and many-facet Rasch measurement.) In many of the generalizability studies, the researchers' focus has been the detection of group-level rater effects (e.g., Baker, Abedi, Linn, and Niemi, 1996; Clauser, Clyman, and Swanson, 1999; Hoyt, 2000; Hoyt and Kerns, 1999; Lane, Liu, Ankenmann, and Stone, 1996; Linn, Burton, DeStefano, and Hanson, 1996; Murphy and DeShon, 2000a). However, as Lynch and McNamara (1998) noted, methodologists have recently been working to extend generalizability theory to increase its flexibility in terms of its data and design requirements and further its capabilities to provide specific information regarding the performance of individual raters, ratees, and traits (Brennan, 2000, 2001; MacMillan, 2000; Marcoulides, 1999; Marcoulides and Drezner, 1997).

Researchers have devised a variety of regression-based procedures to investigate the rater leniency/severity effect and to adjust ratee measures for the impact of this effect. Some have experimented with multivariate analysis procedures for incomplete data to impute ratings (Beale and Little, 1975; Houston, Raymond, and Svec, 1991; Little and Rubin, 1987; Raymond, 1986; Raymond and Houston, 1990). Others have proposed least-squares regression procedures (Cason and Cason, 1984; DeGruijter, 1984; Raymond and Viswesvaran, 1993; Raymond, Webb, and Houston, 1991). In some studies, ordinary least-squares approaches are used (e.g., Braun, 1988; Lance, LaPointe, and Stewart, 1994), while in other studies weighted least-squares approaches are employed (e.g., Wilson, 1988).

Researchers using least-squares regression procedures and many-facet Rasch measurement procedures adjust ratee performance measures for differences in the level of leniency/severity that raters exercise. Longford (1994a, 1994b, 1995, 1996) has proposed an alternative approach for measuring and adjusting ratee performance measures for rater behavior. Longford's additive

variance components model employs an empirical Bayes framework to estimate variances due to true scores, rater leniency/severity, and rater inconsistency. The model then adjusts ratee performance measures using these multiple sources of information it obtains about each rater. In certain settings, variance due to differences in rater consistency can be much larger than variance due to differences in rater leniency/severity and therefore needs to be taken into consideration in adjusting ratee performance measures, Longford argues.

Researchers working from an item response theory (IRT) perspective have recently proposed a variety of complex models for detecting and measuring rater effects. Several researchers (McNamara and Adams, 2000; Muraki, 1999; Patz, Junker, Johnson, and Mariano, 2002; Wilson and Hoskens, 1999) have devised IRT models to deal with possible dependency among multiple ratings (i.e., models designed to correct the standard errors of ratee performance measures when there is substantial dependence among a ratee's ratings across traits). Extensions of conventional Rasch models have also been developed to handle multidimensional rating data (Adams, Wilson, and Wang, 1997; Wang, Wilson, and Adams, 1997; Wilson and Adams, 1995) and multilevel rating data (Adams, Wilson, and Wu, 1997; Wang, 1997). Some researchers have proposed variants of two-parameter IRT models that explicitly model rater discrimination, allowing the researcher to use a data-modeling approach to identify rater effects, such as central tendency and differential leniency/severity (Patz, Wilson, and Hoskens, 1997; Rost, 1988; Wolfe, 1998).

Future Directions for Research in the Detection and Measurement of Rater Effects

An important issue that we have not addressed, one that is ripe for future research in the detection and measurement of rater effects, concerns the extent to which rater behavior fluctuates as a rating operation progresses. In this paper, we have looked at the detection and measurement of static rater effects. But some researchers have been asking whether rater effects might rather be dynamic effects that change over time. A number of researchers have examined the stability of rater leniency/severity measures over time (Braun, 1988; Hoskens and Wilson, 2001; Lumley and McNamara, 1995; Lunz, Stahl, and Wright, 1996; Myford, Marr, and Linacre, 1996; O'Neill and Lunz, 2000; Wilson and Case, 2000). While most researchers studying rater drift have focused on detecting changes in leniency/severity, few have looked at the stability of measures of other rater effects over time (Wolfe, Moulder, and Myford, 2001; Wolfe, in press). This is a critical area for future research, since it is important to know just how variable the rating behavior of individual raters is. If rater effects are, indeed, dynamic rather than static, then that can have serious implications for how we might handle the ratings they assign at different points in time.

The *Facets* computer program adjusts ratee performance measures for differences in rater leniency/severity, generally basing those adjustments on a single overall measure of severity for each rater. If rater behavior does fluctuate as a rating operation progresses, then some might question the appropriateness of this method for adjusting ratings for rater severity differences: Does this adjustment process produce "fairer" ratings for all ratees, or only for some? As we have shown in this paper, a rater severity effect can present itself in several distinct ways, some more subtle than others:

- Some severe raters may underestimate the level of ratee performance across the entire performance continuum. These raters tend to consistently assign lower ratings than other raters to all ratees.
- Other raters may exhibit a tendency to cluster their ratings around a particular category on a rating scale (i.e., show restriction of range in their ratings). That category may be at the high end of the scale, the low end of the scale, or in the middle of the scale. If a rater's ratings tend to cluster at the lower end of the scale, then that may signal severity. Unlike the previous example, the rater

does not underestimate ratee performance across the entire performance continuum—only along a portion of that continuum. The net effect is still detectable as rater severity, though the pattern of ratings for a rater showing restriction of range may differ somewhat from the pattern of ratings for a rater who consistently assigns lower ratings than other raters to all ratees.

- A rater may selectively exhibit a severity effect. That is, a rater may be differentially severe, showing a tendency to assign ratings that are lower than expected to ratees in certain subgroups, given the ratings that other raters assign these same ratees. However, the raters may not show this same tendency when rating other ratee subgroups.

The key question that arises, then, is this: Is it appropriate to use a "one-size-fits-all" approach to adjusting ratings for rater severity differences if, indeed, rater severity differences can present in these categorically different ways? Perhaps what is needed is a more mathematically sophisticated approach to adjustment that would take into account the potentially localized nature of a rater severity effect. Such an approach would not make the assumption that a severe rater exercises a constant level of severity no matter what ratee he or she is rating, no matter what day the rating occurs, no matter whether the ratee is rated in the morning or in the afternoon, no matter whether the ratee is the first to be rated or the last, no matter what subgroup the ratee belongs to, etc. Rather, this alternative approach would take into consideration these contextualized (and potentially powerful) facets of the rating operation and would use information about differences in rater performance related to these facets in adjusting ratings. Accordingly, additional research is needed to determine the best way to interpret these types of interaction effects and their impact on adjustments for rater severity.

A second area in which additional research is needed concerns the development and implementation of methods for detecting and adjusting for a greater variety of rater effects in commercially available software. Adjusting ratings for differences in the level of leniency/severity that raters exercise represents an important methodological step toward increasing the objectivity of measurement. However, adjusting ratings for differences in rater leniency/severity does not remove all subjectivity from ratings, as we have seen. Currently available computer software programs do not attend to other important rater effects. Additional research and development is needed to refine existing computer programs (or develop new ones) that embody a more sophisticated approach to the detection of multiple rater effects and that will enable the adjustment of ratings for these multiple effects, not just for rater leniency/severity effects.[6]

A third area of needed future research concerns how best to use model-based rater indices to help in determining whether certain raters need additional training, and in deciding what the nature of that training might be. As we have pointed out in this paper, one of the major advantages of using a many-facet Rasch measurement approach to analyze rating data is that the output from such analyses provides valuable information that can help determine how well the various aspects of a rating operation are functioning. Using selected pieces of output from *Facets* analyses, one can pinpoint aspects of a complex rating operation that were functioning as intended, as well as potentially problematic aspects. The analyses provide specific information about how each "element" of each facet (e.g., each rater, ratee, trait, rating scale) within the rating operation performed—detailed information that those in charge of monitoring quality control for a rating operation could use to initiate meaningful changes to improve it. For example, by reviewing the detailed information provided in quality control charts and/or tables for misfitting raters, supervisors could gain an understanding of the specific nature of each rater's misfit. The supervisor would then be in a much stronger position to determine how best to work with each rater, providing individually targeted feedback and retraining activities to help them learn to use the rating scales in a more consistent fashion. Further, if *Facets* analyses could be conducted in "real time" (i.e., while a rating operation is taking

place), then supervisors could use the rater fit statistics to identify early on those who need additional training before they are allowed to score operationally.

Studies are needed that look at the effectiveness of providing feedback to raters who exhibit rater effects. Are they able to use that feedback to change their behavior so that they no longer exhibit that effect (or does providing such feedback lead to their exhibiting other unwanted rater effects, creating perhaps yet another set of challenges to be addressed in rater training)? What types of feedback are most effective? How often should it be given? How can we determine whether raters are able to use the feedback they receive to change their rating behavior in appropriate ways? What statistical indicators should we be using to measure the change? Hoskens and Wilson (2001) have carried out pioneering research in this area within the context of the scoring of essays included in a statewide test, but more studies are needed in different rating contexts.

More research is also needed to determine whether supervisors in charge of monitoring rating operations can make use of the results from the *Facets* analyses to help them devise targeted retraining activities for raters showing evidence of bias in their ratings. For example, the output from *Facets* bias analyses could help supervisors identify individual raters who showed a differential severity effect related to a specific ratee background characteristic (e.g., gender, race/ethnicity). An example of this would be the real-time evaluation of rater behavior in the Olympic Figure Skating competition, which would have detected the aberrant judge behavior at the 2002 Winter Olympics before it became a scandal. By reviewing bias analyses quality control charts and/or tables for each rater, supervisors could pinpoint the particular ratees most affected. Additionally, the supervisors could use *Facets* output to identify specific patterns of differential severity and determine whether there are groups of raters who exhibit similar patterns. With this information in hand, supervisors should be in a better position to devise training activities that could be tailored to meet those raters' specific needs, helping them to become aware of the biases they exhibit as they explore positive steps they could take to deal with those biases. But can supervisors develop these types of targeted retraining activities? What are the characteristics of effective retraining activities that lead to the elimination of rater biases? How effective are such activities in helping raters overcome their biases? How would we measure the effectiveness of such activities?

Fourth, additional research is needed to gain an understanding of how various constraints and operational practices impact the detection and measurement of rater effects. For example, because of the cost associated with assigning ratings, most rating operations collect ratings from only one or two raters for each example of ratee performance. As a result, the data matrices from such rating designs are sparse, containing only a small percentage of ratings, with a concomitant large percentage of missing data. Unless care is taken when setting up the rating design to ensure that all raters rate a subset of ratees in common, disconnected subsets of ratee performances may exist. When these disconnected subsets occur, it is not possible to place all ratee and rater measures onto a common linear continuum that would then allow one to make direct comparisons among all raters and ratees. The influence of disconnected subsets of raters and ratees and the impact that extensive missing data have on the validity of indicators of rater effects has not been examined, and this is another important area for potential future research.

A fifth area of needed future research would be to determine how various procedures and practices for examining and resolving rater disagreements impact the validity of ratee performance measures. Raters sometimes differ in the manner in which they use the various categories on a rating scale. Consequently, they do not always agree in the ratings they assign. Assessment programs have adopted a variety of different procedures for resolving discrepancies between raters when they occur (Johnson, Penny, and Johnson, 2000). Unfortunately, most research involving rater effects has focused on the raw ratings that

raters assign to a particular ratee, not the "resolved" ratings. Researchers have directed very little attention toward understanding the impact of different procedures for resolving discrepant ratings on the detection and measurement of rater effects. Myford and Wolfe (2002) determined that common criteria for defining a pair of ratings as being "discrepant," and therefore requiring "resolution," do not necessarily identify the same cases as requiring attention as do the rater effect indices presented in this paper. For example, their preliminary research revealed that about 7% of the ratings in a particular administration of the *Test of Spoken English* were identified as being potentially problematic by criteria focusing on either rater resolution criteria or rater effect indices—only 1% of the ratings were simultaneously identified as being potentially problematic by *both* of these sets of criteria.

A sixth potentially fruitful area for investigation involves the development of a comprehensive "rater reliability" index which balances agreement (Cohen's kappa), trends (simple correlations), and variances (intra-class correlations). We need group-wise, not merely pair-wise, summary statistics. At present, there is no comprehensive index that summarizes the overall quality of a rating operation. Are the raters doing a better job this year than last? What difference did the extra day of rater training make? As a whole, the field seems to have advanced little since the publication of Saal, Downey, and Lahey, (1980) in its attempts to devise informative group-wise summary statistics.

Clearly, we have come a long way over the last three quarters of a century in our quest to better understand rating behavior. We have developed (and are continuing to develop) a number of useful tools for detecting and measuring rater effects, and we have learned a great deal about how raters can differ in their uses of rating scales. We are gaining an understanding of the cognitive processes that raters employ and the biases they use to filter information to arrive at their ratings. But we have miles to go yet on our journey toward understanding, with many promising adventures awaiting us along the way.

Acknowledgments

We would like to acknowledge the helpful comments and suggestions from Everett Smith, Richard Smith, Gwyneth Boodoo, Yong-Won Lee, Daniel Eignor, Michael Linacre, and Terrence Jackson in the preparation of this paper. The material contained herein is based on work supported by the Educational Testing Service, Michigan State University, and the University of Illinois at Chicago. Any opinions, findings, conclusions, and recommendations expressed are those of the authors and do not necessarily reflect the views of the Educational Testing Service, the University of Illinois at Chicago, or Michigan State University.

Footnotes

[1] The differences between the means for the "normal" and "effect" males (and the differences between the means for the "normal" males and females) are due to randomness in the Rasch model. The model is modeling a probabilistic event.

[2] Researchers studying restriction of range using a classical test theory approach often examine the standard deviation of the ratings across all ratees for a given trait (i.e., a group-level statistical indicator of restriction of range). The smaller the standard deviation, the greater the restriction-of-range effect in the ratings, they reason. However, comparing the standard deviation for an individual rater to the standard deviation of all raters is not a useful approach for deciding whether a particular rater is showing a leniency/severity effect or a restriction-of-range effect. The researcher will not be able to determine whether that rater is exhibiting a leniency/severity effect or is showing restriction of range. To make this distinction, the researcher would need to compare conditional standard deviations, because the standard deviations for severe or lenient raters will tend to be smaller than those for other more "normal" raters. In addition, because severe/lenient raters tend not to use the scale categories with the same frequency as more "normal" raters, it is difficult to know how much smaller these conditional standard deviations

would need to be before a researcher could say with much confidence that a rater showed a restriction-of-range effect as opposed to a leniency/severity effect.

[3] For example, note that even the small differences in rater severities displayed in Table 3 result in a statistically significant fixed chi-square statistic.

[4] Note that the "single rater—rest of the raters" correlation for Rater 10 is somewhat smaller than the correlations for the remaining raters. Introducing central tendency into the ratings of Rater 10 resulted in range restriction, and that range restriction impacted the correlation coefficient.

[5] It is important to emphasize that threshold reversals (or disordered average measures) are frequently signals that one or more rating scale categories are not working as intended. The root problem may not be an individual rater's use of the scale. Rather, the problem may be with the rating scale itself and the way in which it was constructed. For example, when there are threshold reversals (or disordered average measures), sometimes one or more rating scale category labels are confusingly worded, making it difficult for raters to distinguish among categories. In this situation, raters may assign ratees to those categories somewhat haphazardly, since the meaning of one or more categories may not be clear. Threshold reversals (or disordered average measures) may also signal that two or more rating scale categories overlap in meaning. When categories are not defined such that they are mutually exclusive, then the boundaries between those categories are blurred. It becomes difficult to tell where one category ends, and the next one begins. Reliably assigning ratees to categories under such circumstances becomes highly problematic. So how does the researcher determine whether the problem is with an individual rater and his/her use of the scale, or with the rating scale itself and the way in which it was constructed? When there are threshold reversals or disordered average measures, look at the Rater Measurement Report included as Table 7 of output from the MFRM analysis to see how many misfitting raters there are. If there are a number of misfitting raters, then chances are there is a problem with the rating scale rather than with an individual rater's use of the scale. Alternatively, fit Hybrid Model #2 or #3 to the data, and examine the manner in which individual raters employ the rating scale.

[6] As we noted earlier, Longford's (1994a, 1994b, 1995, 1996) model adjusts ratee performance measures for variance due to differences in rater consistency as well as differences in rater leniency/severity, but, like *Facets*, his adjustment approach does not take into consideration all the various rater effects that we have described in this paper.

References

Adams, R. J., Wilson, M. R., and Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*, 1-23.

Adams, R. J., Wilson, M., and Wu, M. (1997). Multilevel item response modeling: An approach to errors in variables regression. *Journal of Educational and Behavioral Statistics, 22*, 47-76.

Andrich, D. (1998). Thresholds, steps and rating scale conceptualization. *Rasch Measurement: Transactions of the Rasch Measurement SIG, 12*, 648.

Baker, E. L., Abedi, J., Linn, R. L., and Niemi, D. (1996). Dimensionality and generalizability of domain-independent performance assessments. *Journal of Educational Research, 89*, 197-205.

Beale, E. M. L., and Little, R. J. A. (1975). Missing data in multivariate analysis. *Journal of the Royal Statistical Society (B)*, 129-145.

Bernardin, H. J., and Pence, E. C. (1980). Effects of rater training: Creating new response sets and decreasing accuracy. *Journal of Applied Psychology, 65*, 60-66.

Borman, W. C. (1977). Consistency of rating accuracy and rating errors in the judgment of human performance. *Organizational Behavior and Human Performance, 20*, 238-252.

Braun, H. I. (1988). Understanding scoring reliability: Experiments in calibrating essay readers. *Journal of Educational Statistics, 13*, 1-18.

Brennan, R. L. (1983). *Elements of generalizability theory*. Iowa City, IA: The American College Testing Program.

Brennan, R. L. (2000). (Mis) conceptions about generalizability theory. *Educational Measurement: Issues and Practice, 19*, 5-10.

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

Cason, G. J., and Cason, C. L. (1984). A deterministic theory of clinical performance rating. *Evaluation and the Health Professions, 7*, 221-247.

Clauser, B. E., Clyman, S. G., and Swanson, D. B. (1999). Components of rater error in a complex performance assessment. *Journal of Educational Measurement, 36*, 29-45.

Crocker, L., and Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart, and Winston.

Cronbach, L. J., Gleser, G. C., Nanda, H., and Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability scores and profiles*. New York: Wiley.

DeGruijter, D. N. M. (1984). Two simple models for rater effects. *Applied Psychological Measurement, 8*, 213-218.

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

Guilford, J. P. (1954). *Psychometric methods* (2nd ed.). New York: McGraw Hill.

Hedge, J. W., and Kavanagh, M. J. (1988). Improving the accuracy of performance evaluations: Comparison of three methods of performance appraiser training. *Journal of Applied Psychology, 73*, 68-73.

Hill, C. E., O'Grady, K. E., and Price, P. (1988). A method for investigating sources of rater bias. *Journal of Counseling Psychology, 35*, 346-350.

Hoskens, M., and Wilson, M. (2001). Real-time feedback on rater drift in constructed response items: An example from the Golden State Examination. *Journal of Educational Measurement, 38*, 121-146.

Houston, W. M., Raymond, M. R., and Svec, J. C. (1991). Adjustments for rater effects in performance assessment. *Applied Psychological Measurement, 15*, 409-421.

Hoyt, W. T. (2000). Rater bias in psychological research: When is it a problem and what can we do about it? *Psychological Methods, 5*, 64-86.

Hoyt, W. T., and Kerns, M. D. (1999). Magnitude and moderators of bias in observer ratings: A meta-analysis. *Psychological Methods, 4*, 403-424.

Johnson, R. L., Penny, J., and Johnson, C. (2000). The relationship between score resolution methods and interrater reliability: An empirical study of an analytic scoring rubric. *Applied Measurement in Education, 13*, 121-138.

Keaveny, T. J., and McGann, A. F. (1975). A comparison of behavioral expectation scales and graphic rating scales. *Journal of Applied Psychology, 60*, 695-703.

Kenny, D. A., and Kashy, D. A. (1992). Analysis of the multitrait-multimethod matrix by confirmatory factor analysis. *Psychological Bulletin, 112*, 165-172.

Lance, C. E., LaPointe, J. A., and Stewart, A. M. (1994). A test of the context dependency of three causal models of halo rater error. *Journal of Applied Psychology, 79*, 332-340.

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*, 71-92.

Latham, G. P., Wexley, K. N., and Pursell, E. D. (1975). Training managers to minimize rating errors in the observation of behavior. *Journal of Applied Psychology, 60*, 550-555.

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

Linacre, J. M. (1996). Generalizability and many-facet Rasch measurement. In G. Engelhard, Jr., and M. Wilson (Eds.), *Objective measurement: Theory into practice: Vol. 3* (pp. 85-98). Norwood, NJ: Ablex.

Linacre, J. M. (1999). Investigating rating scale category utility. *Journal of Outcome Measurement, 3*, 103-122.

Linacre, J. M. (2001a). *FACETS* [Computer program, version 3.36.2]. Chicago: MESA Press.

Linacre, J. M. (2001b). *A user's guide to Facets: Rasch measurement computer program* [Computer program manual]. Chicago: MESA Press.

Linacre, J. M. (2002). Number of person or item strata. *Rasch Measurement: Transactions of the Rasch Measurement SIG, 16*, 888.

Linn, R. L., Burton, E., DeStefano, L., and Hanson, M. (1996). Generalizability of New Standards Project 1993 pilot study tasks in mathematics. *Applied Measurement in Education, 9*, 201-214.

Little, R. J. A., and Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: John Wiley.

Longford, N. T. (1994a). Reliability of essay rating and score adjustment. *Journal of Educational and Behavioral Statistics, 19*, 171-201.

Longford, N. T. (1994b). *A case for adjusting subjectively rated scores in the Advanced Placement tests* (ETS Technical Report 94-5). Princeton, NJ: Educational Testing Service.

Longford, N. T. (1995). *Measurement of uncertainty in educational testing*. New York: Springer-Verlag.

Longford, N. T. (1996). *Adjustment for reader rating behavior in the Test of Written English* (TOEFL Research Report No. 55). Princeton, NJ: Educational Testing Service.

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

Lunz, M. E., Stahl, J. A., and Wright, B. D. (1996). The invariance of rater severity calibrations. In G. Engelhard, Jr., and M. Wilson (Eds.), *Objective measurement: Theory into practice: Vol. 3* (pp. 99-112). Norwood, NJ: Ablex.

Lynch, B. K., and McNamara, T. F. (1998). Using G-theory and many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing, 15*, 158-180.

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

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.

Marcoulides, G. A., and Drezner, Z. (1997). A method for analyzing performance assessments. In M. Wilson, G. Engelhard, Jr., and K. Draney (Eds.), *Objective measurement: Theory into practice: Vol. 4* (pp. 261-277). Greenwich, CT: Ablex.

Marsh, H. W. (1989). Confirmatory factor analysis of multitrait-multimethod data: Many problems and a few solutions. *Applied Psychological Measurement, 13*, 335-361.

Marsh, H. W., and Bailey, M. (1991). Confirmatory factor analyses of multitrait-multimethod data: A comparison of alternative models. *Applied Psychological Measurement, 15*, 47-70.

McNamara, T. F., and Adams, R. J. (2000). The implications of halo effects and item dependencies for objective measurement. In M. Wilson, and G. Engelhard, Jr. (Eds.), *Objective measurement: Theory into practice: Vol. 5* (pp. 243-257). Stamford, CT: Ablex.

Muraki, E. (1999, April). *The introduction of essay questions to the GRE: Toward a syn-*

thesis of item response theory and generalizability theory. Paper presented at the annual meeting of the American Educational Research Association, Montréal, Canada.

Muraki, E., and Bock, R. D. (2003). *PARSCALE* [Computer program, version 4]. St. Paul MN: Assessment Systems Corporation.

Murphy, K. R., and Anhalt, R. L. (1992). Is halo error a property of the rater, ratees, or the specific behavior observed? *Journal of Applied Psychology, 77*, 494-500.

Murphy, K. R., and DeShon, R. (2000a). Interrater correlations do not estimate the reliability of job performance ratings. *Personnel Psychology, 53*, 873-900.

Murphy, K. R., and DeShon, R. (2000b). Progress in psychometrics: Can industrial and organizational psychology catch up? *Personnel Psychology, 53*, 913-924.

Myford, C. M., Marr, D. B., and Linacre, J. M. (1996). *Reader calibration and its potential role in equating for the TWE* (TOEFL Research Report No. 95-40). Princeton, NJ: Educational Testing Service.

Myford, C. M., and Mislevy, R. J. (1995). *Monitoring and improving a portfolio assessment system* (MS #94-05). Princeton, NJ: Educational Testing Service, Center for Performance Assessment.

Myford, C. M., and Wolfe, E. W. (2002). When raters disagree, then what: Examining a third-rating discrepancy resolution procedure and its utility for identifying unusual patterns of ratings. *Journal of Applied Measurement, 3*, 300-324.

O'Grady, K. E., and Medoff, D. R. (1991). Rater reliability—a maximum-likelihood confirmatory factor-analytic approach. *Multivariate Behavioral Research, 26*, 363-387.

O'Neill, T. R., and Lunz, M. E. (2000). A method to study rater severity across several administrations. In M. Wilson, and G. Engelhard, Jr. (Eds.), *Objective measurement: Theory into practice: Vol. 5* (pp. 135-146). Stamford, CT: Ablex.

Patz, R. J., Junker, B. W., Johnson, M. S., and Mariano, L. T. (2002). The hierarchical rater model for rated test items and its application to large-scale educational assessment data. *Journal of Educational and Behavioral Statistics, 27*, 341-384.

Patz, R. J., Wilson, M. J., and Hoskens, M. (1997). *Optimal rating procedures and methodology for NAEP open-ended items* (Working Paper No. 97-37). Washington, DC: U. S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics. Retrieved Sept. 5, 2001 from the World Wide Web: http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=9737

Pulakos, E. D., Schmitt, N., and Ostroff, C. (1986). A warning about the use of a standard deviation across dimensions within ratees to measure halo. *Journal of Applied Psychology, 71*, 29-32.

Raymond, M. R. (1986). Missing data in evaluation research. *Evaluation and the Health Professions, 9*, 395-420.

Raymond, M. R., and Houston, W. H. (1990). *Detecting and correcting for rater effects in performance assessment* (ACT Research Report Series 90-14). Iowa City, IA: The American College Testing Program.

Raymond, M. R., and Viswesvaran, C. (1993). Least-squares models to correct for rater effects in performance assessment. *Journal of Educational Measurement, 30*, 253-268.

Raymond, M. R., Webb, L. C., and Houston, W. M. (1991). Correcting performance-rating errors in oral examinations. *Evaluation and the Health Professions, 14*, 100-122.

Rost, J. (1988). Rating scale analysis with latent class models. *Psychometrika, 53*, 327-348.

Saal, F. E., Downey, R. G., and Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin, 88*, 413-428.

Scullen, S. E. (1999). Using confirmatory factor analysis of correlated uniquenesses to estimate method variance in multitrait-multimethod

matrices. *Organizational Research Methods, 2*, 275-292.

Scullen, S. E., Mount, M. K., and Goff, M. (2000). Understanding the latent structure of job performance ratings. *Journal of Applied Psychology, 85*, 956-970.

Shavelson, R. J., Webb, N. M., and Rowley, G. L. (1989). Generalizability theory. *American Psychologist, 44*, 922-932.

Wang, W. (1997). *Estimating rater severity with multilevel and multidimensional item response modeling*. Taipei, Taiwan: Taiwan National Science Council. (ERIC Document Reproduction Service No. ED 408 340).

Wang, W., Wilson, M. R., and Adams, R. J. (1997). Rasch models for multidimensionality between and within items. In M. Wilson, G. Engelhard, Jr., and K. Draney (Eds.), *Objective measurement: Theory into practice: Vol. 4* (pp. 139-156). Greenwich, CT: Ablex.

Widaman, K. F. (1985). Hierarchically nested covariance structure models for multitrait-multimethod data. *Applied Psychological Measurement, 9*, 1-26.

Wilson, H. G. (1988). Parameter estimation for peer grading under incomplete design. *Educational and Psychological Measurement, 48*, 69-81.

Wilson, M. R., and Adams, R. J. (1995). Rasch models for item bundles. *Psychometrika, 60*, 181-198.

Wilson, M. R., and Case, H. (2000). An examination of variation in rater severity over time: A study of rater drift. In M. Wilson, and G. Engelhard, Jr. (Eds.), *Objective measurement: Theory into practice: Vol. 5* (pp. 113-134). Stamford, CT: Ablex.

Wilson, M. R., and Hoskens, M. (1999, April). *The rater bundle model for constructed-response items: An example in the context of real-time feedback on rater effects*. Paper presented at the annual meeting of the American Educational Research Association, Montréal, Canada.

Wolfe, E. W. (1998, April). *A two-parameter logistic rater model (2PLRM): Detecting rater harshness and centrality*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Wolfe, E. W. (in press). Identifying rater effects using latent trait models. *Psychology Science*.

Wolfe, E. W., Chiu, C. W. T., and Myford, C. M. (1999). *The manifestation of common rater effects in multi-faceted Rasch analyses* (MS #97-02). Princeton, NJ: Educational Testing Service, Center for Performance Assessment.

Wolfe, E. W., Moulder, B. M., and Myford, C. M. (2001). Detecting differential rater functioning over time (DRIFT) using a Rasch multi-faceted rating scale model. *Journal of Applied Measurement, 2*, 256-280.

Wu, M., Adams, R., and Wilson, M. (1997). *ConQuest* [Computer program]. Melbourne, Australia: Australian Council for Educational Research.