International Journal of Assessment and Evaluation in Education
Vol 8, Dec 2018 (23-35) ISSN 2232-1926 / eISSN 2600-870X

Application of Many Faceted Rasch Measurement with FACETS

Priyalatha Govindasamy[1], Antonio Olmos[2], Kathy Green[3] & Maria del Carmen Salazar[4]
[1]Faculty of Education and Human Development, Sultan Idris Education University
Tanjong Malim, 35900, Perak, Malaysia
[2]Aurora Mental Health Center, Aurora, Colorado, 80014, USA.
[3]Research Methods and Statistics, Morgridge College of Education
University of Denver, Denver Colorado, 80208, USA
[4]Morgridge College of Education, University of Denver, Denver, Colorado, 80208, USA
gpriyalatha@fpm.upsi.edu.my[1], antonioolmos@aumhc.org[2], Kathy.Green@du.edu[3], msalazar@du.edu[4]

Received: 20 Mac 2018; Accepted: 5 June 2018; Published: 21 December 2018

Abstract

Many facet Rasch measurement (MFRM) is a type of measurement application that aims to perform analysis of multiple variables that potentially influence results of a test or outcome measure. A facet is a component with a systematic contribution to the variability of the measurement error. Linacre (1989) created the technique as an extension of the Rasch model to model the consistency of judges/raters in rating performances. The purpose is to provide a step- by step guide for practitioners on how to conduct an MFRM analysis in FACETS software using real data.

Keywords: FACET, Many facet Rasch Measurement, Rasch, measurement, Rasch application

INTRODUCTION

Measurement is a process of assigning numbers or symbols to attributes, according to a specific set of rules (Stevens, 1951). The assigned number represents the amount of an attribute that an individual possesses that assists further understanding of the individual. Measuring physical objects is relatively straightforward compared to measuring ability or proficiency. In the social sciences, measurement can be difficult. For example, in the case of ability, there are many factors contributing to one's ability that need to be accounted for to determine the true ability.

When using individuals (judges) as measurement instruments, variation is an expected result, because raters may not agree completely on a judgment. Variation due to disagreements in judgment is usually considered as error, and measurement models can be used to account for it. However, there are some sources of disagreement that may result from bias, which is systematic, and should be eliminated. Systematic errors such as leniency/severity or the so-called "halo effect" (Popham, 1990) can lead to a biased outcome.

In social sciences, Inter-rater reliability (IRR) (Crocker & Algina, 1986) is an extension of the classical test theory framework to assess discrepancy among raters. The IRR coefficient evaluates the degree of discrepancy among raters using a score similar to a reliability coefficient (0 to 1, where numbers closer to 1 indicate higher inter-rater reliability). However the IRR is not intended to analyze rater bias. Usually, low inter-rater reliability is addressed through training (Harding-DeKam et. al., 2015).

Many Facet Rasch measurement (MFRM) is a measurement application that aims to address multiple variables/factors that can potentially influence the outcomes of a test or assessment. In that sense, a **facet** is a factor/component/variable with a systematic (non-random) contribution to the variability of the measurement (measurement variance; Eckes, 2015). The MFRM model is an extension of the basic Rasch model (Bond & Fox, 2007) intended to account for variability introduced by facets, such as raters, occasions, or conditions, in addition to the difficulty of items and examinee ability. Linacre (1989) created the model to study the consistency of judges in rating performances. In this case, the judge's rating represented one facet associated with the performance of the individuals. In this demonstration, a four facet Rasch model is used. The facets are: (1) student teaching apprentice, (2) item, (3) supervisor, and (4) rating time. According to this four-facet example, the probability of an apprentice (n) with competence (B) obtaining a rating of x (where x = 1, 2, 3, 4) on item D from supervisor C with item difficulty F at time T (where T = 1, 2,..6) is expressed as follows:

$$\text{Log}\,(P_{nijkl}/P_{nij(k-1)l}) = B_n - D_i - C_j - F_k - T_l \quad (1)$$

The MFRM model can also be used to explore the interactions among specific facets. The analysis of these interactions allows the evaluation of systematic bias terms in ratings of apprentice performance. Downing (2005) highlights MFRM as capable of not only monitoring the rater effect but also of adjusting for systematic error.

The MFRM has three specific requirements that need to be met (Farrokhi, Esfandiari, & Dalili, 2011):

1. A positive relationship between scores and ability. That is, students with higher ability are more likely to receive higher scores on more items than less skilled students.
2. Local independence. That is, students' performance or response should not be dependent on their response to previous items.
3. Unidimensionality. That is, only one latent or underlying ability is measured at a time. Related to unidimensionality is the idea that all the items in an instrument should measure a single variable/construct.

The purpose of this paper is to provide practitioners with a step-by step guide in performing Many Facet Rasch Measurement analysis using the FACETS software (Linacre, 2013). Our aim is to present a model for the practitioner that includes data preparation, analysis, and interpretation.

What is FACETS?

The software FACETS was developed by John M. Linacre (Linacre, 2013) as an extension of the Rasch Model (Rasch, 1960/1980) which is used to model the relationship between ability and difficulty for items in a measurement instrument. FACETS allow users to account for the effect of factors such as the effect of raters (judges) in a measurement model (Linacre, 1989). It is important to notice that facets can be **fixed** (i.e., we are only interested in understanding the specific instances included in the category), or **random** (i.e., the instances are part of a population and they can be exchanged). FACETS version 3.71.4 has the capacity of testing up to 2 billion observations (Linacre, 2015). The full version of the FACETS is available through purchase at the Winsteps website[1]. In addition, there is a reduced version of FACETS called MINIFAC that is accessible from the same website. MINIFAC has similar functions as FACETS but limits the analysis to 2000 observations.

An illustrative example

Information about the data for this demonstration

Data for this demonstration was from a study conducted by faculty at the University of Denver on pre-service teachers' performance over one year of coursework on the Framework for Equitable and Effective Teaching (FEET) (Salazar, Green, Govindasamy & Lerner, 2016). The FEET is a rater-

[1] http://www.winsteps.com/index.htm

completed measure of teaching performance. Each rater/teacher supervisor is assigned a group of 8-9 pre-service teachers who are evaluated/rated twice during each of three academic quarters, for a total of six ratings. The FEET assessment consists of 13 items on a 4-point rating scale. The four facets in this study were (1) apprentices (pre-service teachers), (2) rater (supervisor), (3) FEET items (item), and (4) time (two ratings in each academic quarter). The probability of a pre-service teacher receiving a specific rating is dependent on all four facets. This means that all the facets introduce variability into the data and need to be partitioned to understand the true ability of the pre-service student teacher. The purpose of the study was to examine the effects of rater lenience/severity on pre-service teacher performance scores and to provide directions for rater training as well as to revise the FEET items. In total, there were 9 raters assessing 68 apprentices on 13 items over 6 times.

STEPS IN CONDUCTING MANY FACETED RASCH MODEL (MFRM)

This section describes the steps needed to conduct a MFRM model using the setting described in the previous section.

Step 1: Setting up the data

First, we discuss the file organization needed for the MFRM software. FACETS has the ability to retireve data from a text file, Excel, R, SPSS, SAS, and Stata. For this demonstration, the data were stored in SPSS (v.22). Figure 1 is an example of data set up in SPSS for a four-facet model.

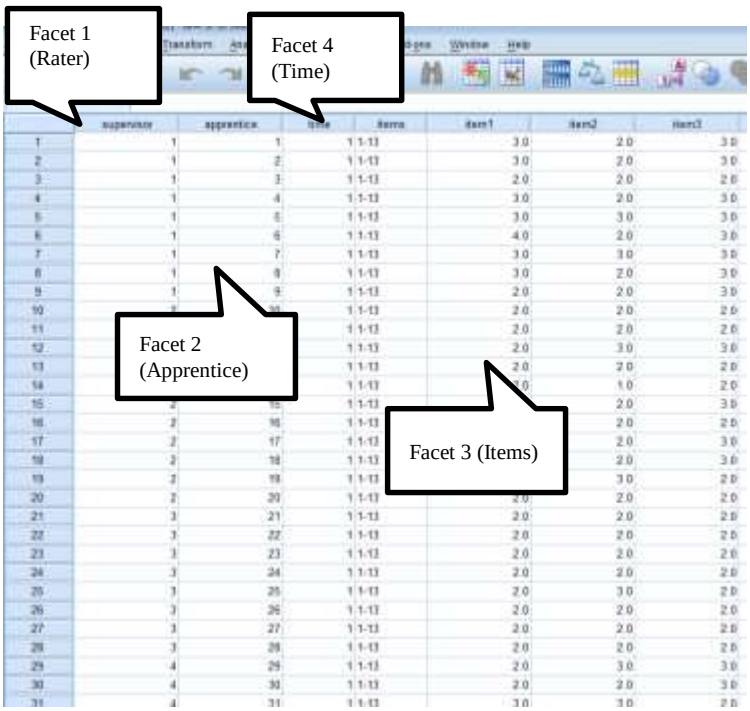
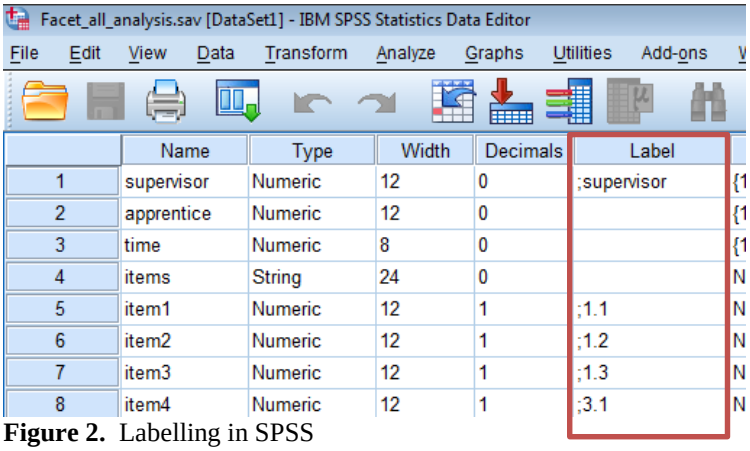


Figure 1. Display of the data used for the example (SPSS file)

The first column in Figure 1 is the supervisors (rater facet) represented by their identification number. Next is the apprentice (facet 2). Facet 4 in the third column is time. The fourth column is a string variable created indicating the total number of items to be used in the analysis. The fifth column onwards represents the responses to items that were administered to the apprentices. In SPSS (v.22), variables can be labelled with the inclusion of a semicolon (;) before the text. For example, the label for item 1 is; 1.1 (Figure 2; red box), indicating time 1-item 1.



	Name	Type	Width	Decimals	Label	
1	supervisor	Numeric	12	0	;supervisor	{1
2	apprentice	Numeric	12	0		{1
3	time	Numeric	8	0		{1
4	items	String	24	0		N
5	item1	Numeric	12	1	;1.1	N
6	item2	Numeric	12	1	;1.2	N
7	item3	Numeric	12	1	;1.3	N
8	item4	Numeric	12	1	;3.1	N

Figure 2. Labelling in SPSS

Step 2: Writing syntax for FACETS

Unlike programs such as SPSS, FACETS is a syntax-based program. The syntax is saved as a text (.txt) file. Figure 3 is the syntax used to run the FACETS analysis illustrated in this paper. Commands are in **boldface** in this example for clarity's sake, but *do not* use boldface when running examples in FACETS. Comments are included after a semi-colon on the right hand side of the file, and are not interpreted by the software

Title = Ratings of FEET study	; title of the output
anchorfile= anchor for overall.txt	; name of the anchor file
Facets = 4; supervisor, apprentice, time, items	; number of facets
Positive =2	; All except for apprentice has a negative direction
Non-centered = 2	; All except for apprentice are centered
Model = ?B,?B,?,?, R4	; observations are ratings in range 1-4 ; examine for interaction/bias between supervisor and apprentice
Rating Scale = item,R4	; Specifying facet that used rating scale & the number categories in the rating scale
0 = Missing	; Specifying the "0" as missing values
*	
Unexpected = 2	; Standardized residual ±2 is flagged
Usort = 4,2,1,3	; Sorting unexpected ratings starting from facet 4 to 3
Vertical = 1A,2A,2*,3A,4A	; Arranging the variable map, where labels are presented for the facets and the asterisk requests the frequency distribution of the apprentice facet
Arrange = M	; Arrange the output based on the measure
Inter-rater = 1	; Facet 1 is the supervisor intended to examine for inter-rater agreement
*	
Labels =	; Specify the facets and label them.
1, supervisor	
1 = Rater1	
2 = Rater2	
3 = Rater3	
*	
2, apprentice	
1-68; 68 apprentices (anonymous)	
*	
3,time	

```
1= PreFall
2= PostFall
3= PreWinter
*
4,items
1=Item1
2=Item2
3=Item3
4=Item4
5=Item5
*
Data=Facet_all_analysis.sav                    ; Specifying the name of the datafile for the
                                                 analysis
```

Figure 3. An example of the script used to run MFRM analysis in the FACETS software.

The following section describes the functions Illustrated in this example (in bold-face) and their associated meaning. The function description is followed by an example.

Title
Assign a name to the analysis and the output.

Anchorfile
This command creates and saves a file containing anchoring values for all the available facets. The anchor file can be created from the "output file" drop down menu on FACETS. **Anchors** in FACETS are intended to fix the scale for some of the facets in the study, Anchors are used for **fixed** facets (for example, items, because we may be only interested in understanding the effect of the specific items included in the example), while other facets are free to vary (i.e., **random**; for example, apprentices, because apprentices may be part of a pool of apprentices). This command generates anchor values/estimates for all the facets including for step/categories in the model. The output is presented in a temporary text file format. It is important to note that an anchor file may not be required in many analyses but in this example it is used to anchor raters, items, and time, because they are considered **fixed**, while letting apprentices be considered as **random**. It should be noted that in many facets analyses, raters would also be considered a random facet if raters represent a random sample from a population of raters.

Facets
Specify the number of facets in the model. In this example there are four facets: 1) Supervisor, 2) apprentice, 3) session (time) and 4) items.

Positive
The function specifies which facets have a positive direction. If a facet has a positive direction, then a higher number (positive logit positions) means more of that variable. In our example, only apprentices have a positive direction, thus in this case higher positions mean highly proficient apprentices. If the direction of the facet is negative (negative logit position), then a lower number means more of that variable. In our example, rater, session, and items are negative; thus a higher position on the graph represents: 1) more severe raters, 3) harder items, and 4) later sessions. The direction of the facets, whether positive (+) or negative (-), is indicated at the top of the ruler on the variable map [see Wright map below].

Non-Centered
Centering creates a frame of reference for a facet(s) based on the other facets (in this example, the frame of reference is provided by items, sessions, and raters). Centering creates a "coordinate system" on which the facet being framed (apprentice in this example) is positioned. Centering allows for an easier interpretation of how the non-centered facets relate to the centered one. In our example, one could focus on how severe were raters with the same apprentice, or item difficulty with regard to a specific apprentice, etc.

Model
This statement identifies the specifics of a model that can be tested. That means the possibility of including some interaction among the facets. Facets in this command are indicated by the symbol "?" and each "?" represents one of the facets in the model. For example, there will be four "?" in a four facet model. Bias interaction terms between the facets can be generated by including the letter "B" after the appropriate facet (i.e., "?") . In our example, the statement "?B,?,?,?B, R4" indicates an analysis that will include only the interaction between supervisor (facet 1) and item (facet 4). It is important to note that higher order interactions can also be defined.

Rating scale
Indicates the facet on which responses are rated as well as the number of categories in the rating scale. This facet is typically the item facet. In our example, the facet is item, and the number of categories is 4.

Unexpected
This command flags large residuals from model-predicted estimates. Residuals quantify the degree to which observations do not fit the model. Usually, the standardized residuals for an observation are flagged as unexpected if they exceed ±2. In this example, this criterion (±2) is used to flag residuals.

Usort
The "Usort" command sorts unexpected ratings or responses using facets in some specific sequence. For example, the command: Usort = 4, 3, 2, 1 sorts the unexpected residuals from the fourth facet followed by facets 3, 2 and 1. In this example, sorting by items first (facet 4), can help with the identification of unexpected responses based on the items before displaying residuals on the other facets.

Vertical
This statement is used to organize the facets in the variable map/Wright map. Facets are organized in numerical order from left to right. Numbers, letters and symbols are used to classify and illustrate the facets in the variable map. For example, the command: "1A,2A,2*,3A,4A" instructs FACETS to add labels for all four facets in the variable map. It also instructs FACETS to include a frequency distribution for the apprentice (facet 2) in the variable map [see Wright Map below]. Table 1 describes the letters and symbols used to enhance the presentation of facets (Linacre, 2013).

Table 1. Description of the Letters and Symbols used to generate the Variable Map/Wright map

Letter/symbol	Description
N	Show element numbers for a specific facet
*	Show frequency distribution of the elements for a specific facet
C	Show count of the elements in this position for a specific facet
A	(or any character other than N nor *, e.g., L) Show element labels (alphabetically)
S	Show the rating scale category numbers (default)
SL	Show the rating scale category numbers and labels
#S	Do not show scoring (rating scale or partial credit) values

Arrange
This function specifies the arrangement of the output. Table 2 presents the organizing options allowed for the output (Linacre, 2013).

Table 2. Description of the letters for generating potential outcome options

Letters	Description
A a	Element labels (alphabetically)
M m	Element Measure
F f	Fit - more extreme of INFIT and OUTFIT (also Z, T)
N n	Element numbers

P p	Point-biserial correlation Arrangements for bias/interaction: use facet 0
M m	Bias Measure
T t	t-statistic of bias measure (also F, Z)
N n	Bias term serial numbers

Inter-rater
This command reports the facet to be used to assess agreement among raters. The statement requires specification of the facet number for raters.

Step 3: Analysis procedure

This section presents a detailed description of the indicators used to evaluate each facet in the model. Each facet introduces variability into the measurement model. Each facet is evaluated to ensure that the facet and all its levels are functioning as intended. First, a variable map/Wright map can be examined to gain an overall understanding of the measure. Subsequently, the detailed measurement results for each facet are examined separately. In most studies, the evaluation of facets begins with the rater/supervisor facet. In our example, this initial assessment is followed by evaluation of the apprentice, session(time), and item facets. Following is the description of the indicators used to evaluate each facet.

Overall indicators

Variable/Wright map
A variable map also known as a Wright map (see Figure 4) is a very informative output. This figure presents the calibrations of multiple facets (in our example, supervisor, apprentice, time, and items) in a single layout. All facets in the analysis are displayed on a single frame of reference that helps to make comparisons within and between various facets. A detailed interpretation of the Wright map is presented under Step 4: Results from a simple analysis.

Individual indicators

Fixed/Random chi-square
The fixed/Random chi-square statistic (Fixed/Random effects) describes the heterogeneity/difference within the facets. A significant chi-square result rejects the null hypothesis that all facet levels are equal. The chi-square statistic is an omnibus test. Thus a significant chi-square indicates the presence of heterogeneity, but does not identify where the measure differs within the facets.

Separation ratio
This statistic gives the spread of the elements of the facet's measures relative to the precision of those measures. Values closer to zero indicate less variability within the facets. This index helps to determine if the facets' differences are truly larger than measurement error. The separation ratio can be estimated as the ratio of the "true" standard deviation of the facet measure after adjusting for measurement error ($SD_{t(J)}$) over the average facet's standard error, ($RMSE_J$). The ratio can be expressed as follows:

$$G_j = SD_{t(J)} \,/\, RMSE_J \qquad (2)$$

Where G_j is an indicator presenting the spread of the facets in measurement error units. A higher rater separation ratio (G_j) shows greater spread of the facet's measures.

Separation index
The separation index can be used to further understand the facets patterns. This index informs us about the potential groups for that facet, due to their similar response patterns. The separation index can be expressed as follows:

$$H_J = (4G_J + 1)/3 \qquad (3)$$

The value, (H_J) represents the potential strata or groups that raters can be classified into. Higher separation index values indicate the presence of more variation among the elements of a facet.

Reliability of separation

In the present context, reliability is the variance of the facet's measure over measurement error. Higher separation values mark the presence of heterogeneity within the facets' measures. The computed separation ratio (G_j) is used to compute the reliability. The equation used to estimate separation reliability is:

$$R_{J=}\ G_J^2\ /\ (1+\ G_J^2\ ) \qquad (4)$$

Indicators to further explain rater behavior

Bias interaction

The bias/interaction terms can be examined to understand the interaction between two or more facets (for example, rater and items). Chi-square tests are used to test the statistical significance of bias interactions, followed by an in-depth analysis of the statistically significant interactions.

Step 4: Results from a sample analysis

This section presents the results from the analysis. Output tables from FACETS are presented along with interpretation of the findings.

Variable map

The variable map indicates the positions of supervisor, apprentice, time (sessions) and item (facets) on the same scale (Figure 4 below). The first column in the map provides the scale in logit units. Supervisors are displayed in the next column. The location of the supervisors indicates the severity/leniency when rating the apprentice. Supervisors appearing higher on the column are harder raters while supervisors displayed lower in the column are more lenient when the facet takes a positive orientation. The third and fourth columns represent the apprentices' proficiency estimates on the rated FEET items. The numbers in the 3rd column are the apprentice's ID number. The 4th column repeats the information from the 3rd column, but allows greater understanding of the distribution of apprentices, with every star representing an apprentice. Apprentices with higher proficiency are near the top of the figure while less proficient apprentices are near the bottom of the column. The fifth column represents time (sessions) of the evaluations. In this example, apprentices were rated up to six times (twice per quarter). The session distribution has a specific meaning, with values near the top indicating sessions in which apprentices received lower ratings, and values near the bottom representing sessions where apprentices received higher ratings. Finally, item difficulties are included in column six. Items at the top are more difficult, whereas items at the bottom are easier.

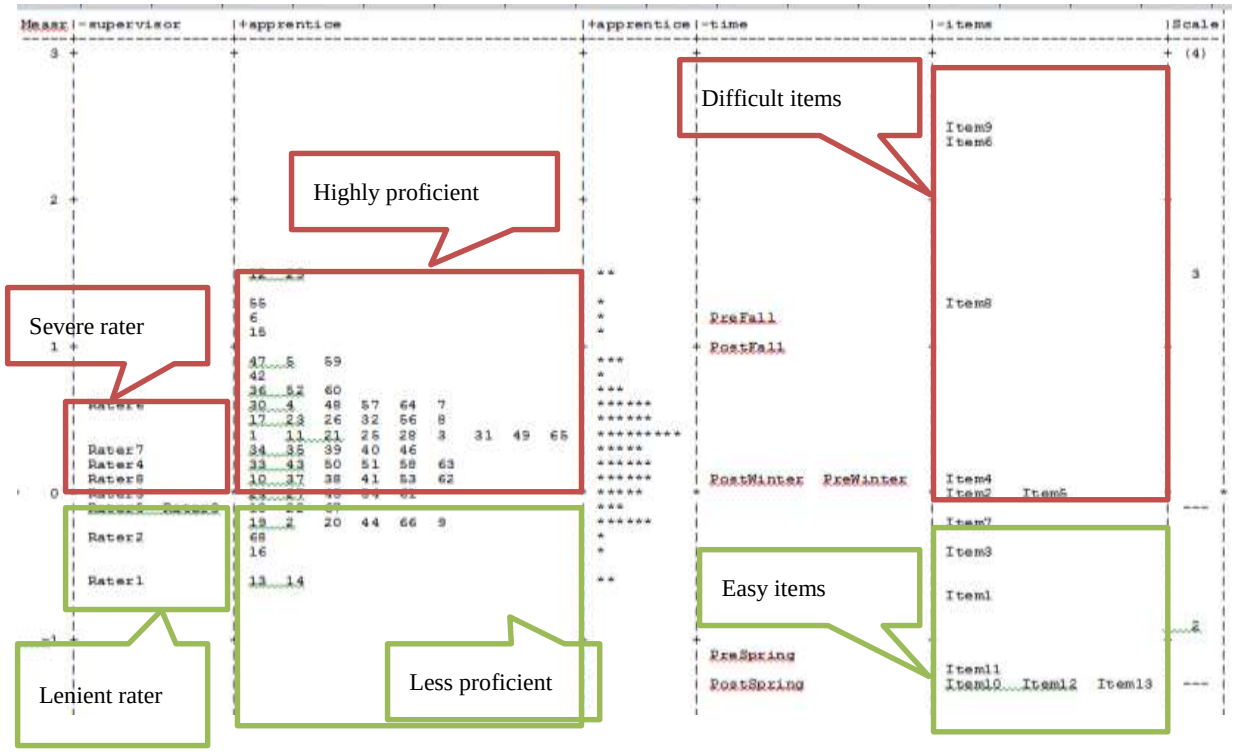


Figure 4. Wright Map

Figure 4 shows that a majority of the items are located above 0.0 logit position (column 6, red box) with items 10, 12, and 13 being the easiest items. Items 8, 6, and 9 were the hardest items. This result must be interpreted with caution as these three items (Items 6, 8, and 9) were not rated across all six time points (sessions). Based on the figure, supervisors can be categorized into three distinctive groups at the cut-off point of .00 logits. The first group is a high severity group (supervisors 6, 7, 4, and 8; second column, red box) in which the supervisors rated the apprentices more severely. Supervisor 3, positioned at the origin, showed intermediate rating severity. The third group of supervisors (supervisors 1, 2, 5, and 9; second column, green box) with logit positions below 0.00 is a lenient group. Overall, supervisor one (-0.59 logits) was identified as the most lenient in rating while supervisor six (+0.61 logits) was the most severe supervisor. The examinees were distributed from -0.63 logits to 1.48 logits. The items were located between -1.30 logits and 2.50 logits. Highly proficient examinees are those above the 0.00 logit (third column, red box), and less proficient are those below the 0.0 logit (third column, green box). There were differences in the time (session) facet. There was a gradual increase in the apprentice proficiency levels over the time points.

The FACETS software also generates tables where the effect of each facet can be analyzed in detail. Tables 3 through 6 (below) provide outputs for rater, apprentice, session, and items respectively.

Rater effect

Table 3 presents the output associated with the rater effect, which was set as fixed (see anchorfile section above)

Table 3. Rater Effect Output

Total Score	Total Count	Obsvd Average	Fair(M) Average	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	Exact Agree. Obs %	Exact Agree. Exp %	N	supervisor
2037	754	2.70	2.84	-.59	.06	1.10	1.5	1.11	1.6	.94	.79	.73	49.0	50.3	1	Rater1
1969	793	2.48	2.77	-.35	.05	.91	-1.4	1.01	.1	.91	.81	.76	60.8	51.3	2	Rater2
1691	702	2.41	2.71	-.12	.06	.63	-6.0	.69	-4.4	1.21	.75	.76	64.3	52.4	5	Rater5
1538	624	2.46	2.69	-.09	.06	.92	-1.1	1.02	.3	.96	.78	.76	63.3	52.7	9	Rater9
1452	611	2.38	2.67	.00	.06	.91	-1.1	1.11	1.2	.84	.78	.76	46.9	52.2	3	Rater3
1662	702	2.37	2.62	.12	.06	.66	-5.5	.66	-4.7	1.25	.75	.76	69.7	52.8	8	Rater8
2020	845	2.39	2.61	.16	.05	.94	-1.0	1.17	2.2	1.05	.71	.76	59.0	52.8	4	Rater4
1363	637	2.14	2.57	.26	.06	1.44	5.7	1.50	3.9	.78	.70	.77	58.5	53.3	7	Rater7
1394	624	2.23	2.41	.61	.06	.84	-2.3	.83	-1.6	1.02	.76	.77	53.2	50.9	6	Rater6

```
|------------------------------+-------------+---------------------+------+------------+-------------+-----------------
-- |
|  1680.7   699.1    2.40   2.66 |    .00   .06 |  .93 -1.3  1.01  -.1 |      |   .76      |             | Mean (Count: 9)
| |   254.2    78.9     .15    .12 |    .33   .00 |  .23  3.3   .25  2.8 |      |   .03      |             | S.D. (Population)
| |   269.6    83.7     .16    .12 |    .35   .00 |  .24  3.5   .26  3.0 |      |   .04      |             | S.D. (Sample)
| +-------------------------------------------------------------------------------------------------------------------
-----|
Model, Populn: RMSE .06  Adj (True) S.D. .32  Separation 5.72  Strata 7.95  Reliability (not inter-rater) .97
Model, Sample: RMSE .06  Adj (True) S.D. .34  Separation 6.07  Strata 8.43  Reliability (not inter-rater) .97
Model, Fixed (all same) chi-square:  300.1  d.f.: 8  significance (probability): .00
Model,  Random (normal) chi-square:  7.8  d.f.: 7  significance (probability): .35
Inter-Rater agreement opportunities: 1235  Exact agreements: 714 =  57.8%  Expected:  640.2 =  51.8%
```

The third line from the bottom shows the Chi-square statistic for the rater (supervisors) effect using a fixed model[2]. A significant Chi square statistic (X^2 (df=8) = 300.1, $p < .001$), as is the case in our example, suggests that the supervisors' severity ratings were statistically different from each other. That is, the supervisors were not "equally calibrated" when evaluating the apprentices. The fourth line from the bottom shows the separation and reliability scores for the sample. The high separation (6.07) and high reliability (.97); suggests different severity ratings by the supervisors. A strata of 8.43 indicates eight different levels of supervisor severity among the supervisors, which matches the Chi-square result described earlier in this paragraph.

A significant Chi-square in addition to separation with high reliability is an indication of unwanted variation between supervisors in their levels of severity. The logit measure of supervisor severity (fifth column from the left) ranged from a low of -0.59 (lenient; Rater 1) to a high of 0.61 (severe; Rater 6). The fit statistics (Infit Mnsq and Outfit Mnsq, columns 7 and 9 from the left) identified all nine supervisors' mean square values as fitting within the accepted range of 0.5 to 1.5 logit of "productive of measurement" (Linacre, 2013). A final word about this section: Please note that the reliability coefficient included in this output is not an inter-rater reliability score. That is, a high reliability score is not an indication that the raters are reliable with each other. In fact as illustrated by this example, in this context, a high "reliability coefficient" indicates the opposite.

Apprentice effect

Table 4 presents the output associated with the ability measure for a small subset (9 of 68) of apprentices. Overall, the apprentice facet reflects adequate quality in measurement.

Table 4. Apprentice Measure Output

```
+-------------------------------------------------------------------------------------------------------------------+
|  Total   Total   Obsvd  Fair(M)|        Model | Infit      Outfit    |Estim.| Correlation |                     |
|  Score   Count  Average Average|Measure  S.E. | MnSq ZStd  MnSq ZStd |Discrm| PtMea PtExp | Nu apprentice       |
|-------------------------------+--------------+---------------------+------+-------------+---------------------|
|   237     117      2.03   2.21 |   -.63   .13 | 1.66  3.6  2.31  2.8 |  .19 |   .54   .79 | 14 14               |
|   179      91      1.97   2.21 |   -.62   .14 |  .78 -1.3   .72  -.6 | 1.17 |   .80   .78 | 13 13               |
|   204      91      2.24   2.35 |   -.40   .15 |  .54 -2.9   .56 -2.0 | 1.23 |   .80   .77 | 16 16               |
|   184      78      2.36   2.58 |    .10   .17 |  .60 -2.1   .70 -1.3 | 1.13 |   .76   .76 | 41 41               |
|   177      78      2.27   2.60 |    .15   .16 |  .49 -3.0   .55 -1.9 | 1.32 |   .78   .77 | 53 53               |
|   238     104      2.29   2.61 |    .16   .14 |  .59 -2.5   .73 -1.3 | 1.12 |   .79   .77 | 58 58               |
|   105      52      2.02   2.61 |    .17   .20 |  .44 -2.6   .39 -2.2 | 1.30 |   .89   .79 | 63 63               |
|   199      91      2.19   2.61 |    .18   .15 |  .96  -.1  1.40  1.4 |  .92 |   .65   .77 | 33 33               |
|   285     117      2.44   2.63 |    .21   .14 |  .63 -2.4   .63 -2.2 | 1.27 |   .72   .75 | 43 43               |
|   208      91      2.29   2.63 |    .22   .15 |  .53 -2.9   .52 -2.4 | 1.19 |   .79   .77 | 51 51               |

|-------------------------------+--------------+---------------------+------+-------------+---------------------|
|   222.4    92.5    2.39   2.65 |    .32   .16 |  .92  -.6   .99  -.2 |      |   .76       | Mean (Count: 68)    |
|    54.4    19.8     .25    .16 |    .44   .03 |  .41  2.2   .49  1.9 |      |   .07       | S.D. (Population)   |
|    54.8    19.9     .25    .16 |    .44   .03 |  .41  2.2   .49  2.0 |      |   .07       | S.D. (Sample)       |
+-------------------------------------------------------------------------------------------------------------------+
Model, Populn: RMSE .16  Adj (True) S.D. .41  Separation 2.50  Strata 3.66  Reliability .86
Model, Sample: RMSE .16  Adj (True) S.D. .41  Separation 2.52  Strata 3.69  Reliability .86
Model, Fixed (all same) chi-square:  540.0  d.f.: 67  significance (probability): .00
Model,  Random (normal) chi-square:  59.1  d.f.: 66  significance (probability): .71
-------------------------------------------------------------------------------------------------------------------
```

The first line from the bottom shows a non-significant Chi-square statistic for apprentice ($X^2(66)$= 59.1, n.s.) under a random effects model (i.e., apprentices are part of a population of similar individuals). This result indicates that statistically speaking, the apprentices did not have different proficiency levels. The third line from the bottom shows a separation ratio of 2.52, with approximately four strata and a reliability index of 0.86. Based on the Wright map (Figure 4), apprentices can be categorized as low, immediate, and high ability using a cut-off value of 0.0 logit. Based on the cut-off value, the majority of the apprentices (about 50 out of 68) demonstrated a high teaching skills

[2] The second line from the bottom shows the chi-square assuming random effects. This result would have been used, had the study assumed that the raters were random.

proficiency level. Finally, from the total sample, fifteen apprentices did not have a good fit (they either underffit or overfit).

Item effect

Table 5 presents the logit item difficulty, standard error of the measure, infit, and outfit indices.

Table 5. Item measure output

Total Score	Total Count	Obsvd Average	Fair(M) Average	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	Nu	items
1462	484	3.02	3.02	-1.30	.09	1.26	2.9	1.18	2.2	1.03	.45	.48	10	Item10
1462	484	3.02	3.02	-1.30	.09	1.27	3.0	1.20	2.4	1.04	.47	.48	12	Item12
1459	484	3.01	3.02	-1.28	.09	1.66	6.7	1.59	6.4	.80	.45	.48	13	Item13
1449	484	2.99	3.00	-1.20	.08	1.47	4.9	1.41	4.6	.88	.48	.48	11	Item11
1365	484	2.82	2.86	-.65	.08	.90	-1.1	1.01	.1	1.03	.51	.49	1	Item1
1327	484	2.74	2.80	-.44	.07	.85	-1.6	.99	-.1	.96	.44	.50	3	Item3
1286	484	2.66	2.74	-.23	.07	.71	-3.5	.85	-1.7	1.14	.51	.51	7	Item7
1229	484	2.54	2.66	.03	.06	.72	-3.8	.93	-.8	.97	.47	.54	2	Item2
1227	484	2.54	2.65	.04	.06	.80	-2.6	1.04	.5	.93	.44	.54	5	Item5
1211	484	2.50	2.63	.10	.06	.65	-5.0	.86	-1.6	.97	.51	.54	4	Item4
814	484	1.68	1.88	1.33	.05	.92	-1.2	.91	-1.1	1.05	.73	.69	8	Item8
431	484	.89	.61	2.41	.06	.91	-1.2	.65	-2.3	1.08	.77	.69	6	Item6
404	484	.83	.53	2.50	.06	.80	-2.8	.55	-3.0	1.11	.78	.69	9	Item9
1163.5	484.0	2.40	2.42	.00	.07	.99	-.4	1.01	.4		.54		Mean (Count: 13)	
359.0	.0	.74	.84	1.27	.01	.30	3.5	.27	2.7		.12		S.D. (Population)	
373.7	.0	.77	.87	1.32	.01	.32	3.7	.28	2.8		.13		S.D. (Sample)	

Model, Populn: RMSE .07 Adj (True) S.D. 1.27 Separation 17.70 Strata 23.94 Reliability 1.00
Model, Sample: RMSE .07 Adj (True) S.D. 1.32 Separation 18.43 Strata 24.90 Reliability 1.00
Model, Fixed (all same) chi-square: 4966.8 d.f.: 12 significance (probability): .00
Model, Random (normal) chi-square: 12.0 d.f.: 11 significance (probability): .37

The second line from the bottom shows the Chi-square statistic for the item's effect assuming a fixed effect model[3]. The significant chi-square statistic, ($X^2(12) = 4966.8$, $p < .001$), indicates statistically significant differences in the item difficulties. Item 13 (signalled by a red arrow) had an infit Mean square larger than 1.5, which suggests a misfitting item (Item 13 asked supervisors to "Analyze practice for continuous improvement"). In general, the items functioned well in capturing the apprentice's teaching skills proficiency. There was sufficient variability among the items used, as indicated by a separation ratio of 18.43 and the presence of 25 potential strata, both indicators of the items' uniqueness.

Time effect

Table 6 presents the time (session's) difficulty measure, standard error, and fit indices. The negative logit measure for spring means it was easier to obtain higher ratings in spring than in fall.

Table 6. Time Measure Output

Total Score	Total Count	Obsvd Average	Fair(M) Average	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	N	time
3211	1092	2.94	3.02	-1.28	.05	1.13	2.4	1.18	3.5	.76	.47	.64	6	PostSpring
2829	988	2.86	2.96	-1.07	.05	1.23	4.1	1.33	5.8	.77	.47	.65	5	PreSpring
2421	1014	2.39	2.64	.07	.05	.86	-2.5	.89	-1.8	1.12	.83	.76	3	PreWinter
2726	1144	2.38	2.63	.11	.04	.92	-1.5	.98	-.3	1.07	.82	.76	4	PostWinter
1870	936	2.00	2.17	1.00	.05	.70	-5.4	.73	-3.4	1.17	.86	.78	2	PostFall
2069	1118	1.85	2.03	1.17	.04	.85	-2.9	.94	-.6	1.06	.80	.77	1	PreFall
2521.0	1048.7	2.40	2.58	.00	.05	.95	-1.0	1.01	.6		.71		Mean (Count: 6)	
456.6	74.6	.40	.37	.93	.00	.18	3.3	.19	3.2		.17		S.D. (Population)	
500.2	81.7	.44	.40	1.02	.00	.20	3.6	.21	3.5		.19		S.D. (Sample)	

Model, Populn: RMSE .05 Adj (True) S.D. .93 Separation 19.89 Strata 26.85 Reliability 1.00
Model, Sample: RMSE .05 Adj (True) S.D. 1.02 Separation 21.79 Strata 29.39 Reliability 1.00
Model, Fixed (all same) chi-square: 2389.6 d.f.: 5 significance (probability): .00
Model, Random (normal) chi-square: 5.0 d.f.: 4 significance (probability): .29

[3] The first line from the bottom shows the chi-square assuming random effects. This result would have been used, had the study assumed that the items were random.

In our example, time was the fourth facet modeled. In Table 6, the second line from the bottom shows the Chi-square statistic for the time (session) facet under a fixed effects model[4]. The significant chi-square statistic ($X^2(5) = 2389.6$, $p < .001$) indicates a statistically significant difference by time of evaluation. The third line from the bottom shows a separation ratio of 21.79 with approximately 30 strata and reliability >0.99. All these values further indicate the difference in time (session) when evaluating the apprentices. Further confirmation of this fact can be found in the Wright map (Figure 4). The map shows a gradual increase in the apprentice's teaching skills from Fall (top of the map) to Spring (bottom of the map) quarter. By the end of the Spring quarter, the apprentices had improved their teaching skills significantly.

Bias interaction

In our example, the objective of the bias-interaction analysis was to determine if some supervisors had specific biases related to some of the items. As explained earlier, the core of the analysis depends on a statistically significant chi-square. Table 9 below include an excerpt of the bias interaction report, including the chi-square table (first row from the bottom), as well as sections of the interaction.

Table 9. Bias Interaction Summary Output

Observd Score	Expctd Score	Observd Count	Obs-Exp Average	Bias Size	Model S.E.	t	d.f.	Prob.	Infit MnSq	Outfit MnSq	supervisor Sq	N	supervisor	measr	items Nu	items	measr
116	139.58	49	-.48	-1.10	.19	-5.92	48	.0000	1.9	1.8	115	7	Rater7	.26	13	Item13	-1.28
124	141.04	47	-.36	-1.08	.22	-4.87	46	.0000	.7	1.0	102	3	Rater3	.00	12	Item12	-1.30
…																	
57	76.75	58	-.34	-.49	.16	-3.02	57	.0038	.7	.5	73	1	Rater1	-.59	9	Item9	2.50
34	23.31	49	.22	.39	.18	2.13	48	.0387	.5	.3	79	7	Rater7	.26	9	Item9	2.50
39	25.49	48	.28	.42	.17	2.49	47	.0162	.6	.4	78	6	Rater6	.61	9	Item9	2.50
…																	
…																	
217	186.26	58	.53	2.35	.32	7.38	57	.0000	1.3	1.6	82	1	Rater1	-.59	10	Item10	-1.30
129.3	129.28	53.8	.00	.03	.21	.00			.8	.9	Mean (Count: 117)						
45.6	44.60	6.1	.20	.53	.04	2.23			.4	.4	S.D. (Population)						
45.8	44.79	6.1	.20	.53	.04	2.24			.4	.4	S.D. (Sample)						

Fixed (all = 0) chi-square: 583.2 d.f.: 117 significance (probability): .00

The bias for the supervisor-items interaction was statistically significant, ($X^2(117) = 583.2$, $p < 0.01$; first row from the bottom), with bias size values between -1.10 and 2.35 (Table 9; fifth column from the left). Given its statistical significance, there is an indication of potential biases in the supervisors' ratings. To illustrate this idea of bias, a few sections of the table were selected (enclosed in a red box). In this section, the responses of three raters (two considered severe; raters 6 and 7 and one considered lenient; rater 1) with regard to the same item (item 9; considered the most difficult item) are presented. In this section of the table it can be observed that the two severe raters have a very different bias size (0.39, 0.42 for raters 7 and 6 respectively), compared to the lenient rater (-0.49). The difference among raters serves as diagnostic to further investigate their behavior towards each item.

CONCLUSION

This paper provides a primer on the analyses used to check the reliability and validity of data coming from different raters. Unfortunately, proper data analyses from raters is an area that does not draw a lot of attention from either practitioners or researchers. And yet, it is as important as any other area associated with data collection. As illustrated in this example, judges can be unreliable and provide biased ratings. The ability of researchers and practitioners to make decisions that could potentially affect programs or interventions depends on the validity of their results. Thus it is important to address any form of validity threat, whether it comes from research designs or measurement error.

Researchers and practitioners can (and often will) conduct inter-rater reliability tests to check how reliable different raters are. Inter-rater reliability is a good option, just like Cronbach's alpha can

[4] The first line from the bottom shows the chi-square assuming random effects. This result would have been used, had the study assumed that the time (sessions) were random.

be to check the reliability of an instrument using classical test theory (Nunnally & Bernstein, 1994). However, inter-rater reliability falls short when it comes to a deeper understanding of rater bias. Many Facet Rasch Measurement models can be used to determine the degree of judge bias. It also can help to assure that the judges are "calibrated."

However, the reader should be aware that the output from FACETS can be long. Therefore, researchers need to be selective in the analysis of their output. Generally results are presented in a text file. However, there are options from a drop-down menu such as "score and measure files" with an option in order to generate tables for publication and for further statistical analysis.

REFERENCES

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). United States of America: Lawrence Erlbaum Associates, Inc.

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

Downing, S. M. (2005). Threats to the validity of clinical teaching assessments: What about rater error? *Medical Education*, 39,350-355.

Eckes, T. (2015). Introduction to Many-Facet Rasch measurement: Analyzing and evaluation rater-mediated assessments. Frankfurt: Peter Lang Edition

Farrokhi, F., Esfandiari, R., & Dalili, M. V. (2011). Applying the Many Facet Rasch Model to detect centrality in self-assessment, peer assessment and teacher assessment. *World Applied Sciences Journal*, *15*, 70-77.

Harding-DeKam, J.L., Reinsvold, L., Olmos, A., Song, Y., Franklin, B., Higgins, T., & Enriquez, M. (2014). Mathematics and science teaching for English learners (MAST-EL) partnership: A relationship among elementary school teachers, pre-service teachers, principals, coaches, and college faculty. *Teacher Education & Practice* (Special theme issue STEM Teacher Preparation and Practice: Prepare and Inspire Students), 27, 2-3, 267-281.

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

Linacre, J. M. (2013). A user's guide to Facets: Rasch measurement computer program [Computer program manual]. Chicago: MESA Press.

Linacre, J. M. (2015). FACETS [Computer program, version 3.71.4]. Chicago: MESA Press.

Nunally & Bernstein. (1994). *Psychometric theory*. New York: McGraw Hill

Popham, W. J. (1990). *Modem educational measurement: a practitioner's perspective* (2nd ed.). Englewood Cliffs NJ: Prentice-Hall.

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests*. (Copenhagen, Danish Institute for Educational Research), expanded edition (1980) with foreword and afterword by B.D. Wright. Chicago: The University of Chicago Press.

Salazar, M.D., Green, K., Govindasamy, P., Lerner, J. (2016, 04, 12). *The Power of the Margins: The Design and Implementation of an Evaluation Model for Equitable and Effective Teaching*. Paper presented at American Education, Research Association, Washington D.C.

Stevens, S. S. (1951). Mathematics, measurement and psychophysics. In Stevens, S. S, *Handbook of Experimental Psychology* (pp.22). New York: John Wiley and Sons Inc.