福岡女子大学文学部・国際文理学部紀要
「文藝と思想」第 78 号　2014 年 2 月　(19)～(45)頁

How High Can They Jump: An Introduction to Rasch Measurement

Trevor A. HOLSTER
J. LAKE

Abstract

Although classroom tests can be summarized as simple raw scores or percentages, standardized tests require more sophisticated analysis to ensure that sufficient quality of measurement is provided for the intended test use. Rasch analysis, one of the conceptually simplest psychometric measurement models, is introduced though an analogy of measuring jumping ability. This contrasts raw score counts of jumping success with interval level measures of jumping ability. Next the Rasch transformations of raw scores and the units of Rasch measures given in logits are explained. Actual item responses may differ from that predicted by the Rasch model so it is possible to quantify the degree of data-model fit which leads into an explanation of Rasch fit statistics. An extension of the Rasch model to additional facets of measurement is then explained. Two worked examples are presented along with a step-by-step guide to the software and analysis. The first example is for a dichotomously scored test using a simple logistic model. The second example is for polytomous items with raters being an additional facet. This paper is intended as an introduction to test analysis for teachers or researchers developing standardized tests and assumes familiarity with statistical analysis typical of high-school graduates.

Introduction

The conceptual foundation of Rasch measurement differs from the raw score based classical test theory (CTT) that is familiar with from books such as Brown's excellent *Testing in Language Programs* (2005). In their popular introductory text on Rasch measurement, Bond and Fox (2007) use the analogy of jumping. We can think of this as a trait that we could call "jumping ability"; an individual with more jumping ability can jump over higher obstacles than a person with less jumping ability. If we wanted to measure how much jumping ability people have, we would determine how high an obstacle they could jump over and report it in units such as meters that allow us to compare the ability of people with the height of obstacles. This comparison is possible because the same measurement units are used for both the height of an obstacle and the ability of a person. If an individual's ability exceeded the height of an obstacle, they would be expected to jump over that obstacle more than they would fail. The difference in ability between 1.0 m and 2.0 m is the same as the difference between 2.0 m and 3.0 m because the meter is an equal interval measure. Additionally, ability of 2.0 m is twice as much as 1.0 m because the meter scale has a meaningful zero point. This means that meters provide ratio level measurement, so we can express differences in jumping ability or height as ratios, such as 1/2 or 22/7.

Observing and measuring psychological and cognitive abilities is more difficult though because we don't have physical objects to scale them against. For example, if we administer a language test with 100 questions, we don't know how much more ability a score of 90 shows than a score of 80 unless we have calibrated the test items and scores on an interval scale of difficulty. The number of times a person jumps is not much use unless we know the height of the obstacles they jump over, and the number of questions answered correctly is not much use unless we have some measure of how difficult the questions are. This is the type of situation the Rasch measurement model was developed for as it lets us convert test scores into interval measures of difficulty or ability using software such as *Winsteps* (Linacre, 2010b) or *Facets* for many-faceted measurement (Linacre, 2010a).

McNamara (1996) provides an accessible introduction to Rasch analysis explaining the stochastic, or probabilistic, nature of the Rasch model. Crucially, this means that estimating the ability of persons and difficulty of items depends on success or failure not being deterministic. Thus, we might expect the type of results shown by Jumper 1 and Jumper 2 in Table 1, constructed to illustrate how people might perform if they were allowed 15 attempts at jumping objects of various heights. We can see that Jumper 1 always successfully jumps 1.0 m, never successfully jumps 1.8 m, and shows high probability of success up to 1.5 m. We might decide that it's reasonable to estimate this person's ability as 1.5m, because they will probably succeed more than they fail up to this level. However, success or failure is stochastic, with a gradual decrease in the probability of success rather than an abrupt change from 100% success to 100% failure. Similarly, Jumper 2 shows high probability of success up to 2.0m, but low probability after that, so we might estimate ability as 2.0m and claim that Jumper 2 is 0.5m better at jumping than Jumper 1. Of course we need to remember that these are estimates and the real difference will probably be somewhat larger or smaller than our estimate.

If we consider Jumper 3 though, we can see a deterministic pattern of 100% success up to 1.5m and 0% success after that point. Although this intuitively seems to give a more precise measure, what we actually know about Jumper 3 is that she better than 1.5m but not as good as 1.6m. Her ability must be somewhere between these points, but we have no way of estimating it precisely. This data over-fits our model of the trait of jumping, and the over-predictability of the results reduces the precision of the measurement that is possible. Under stochastic models such as the Rasch model, we can't convert purely deterministic data into interval measures. This seems extremely counterintuitive, but the probabilistic Rasch data of Jumper 1 and Jumper 2 can give us a more precise estimate of ability than the deterministic, or Guttman, data of Jumper 3. Jumper 4 and Jumper 5

Table 1 ***Probabilistic versus Deterministic Data***

Height	0.8	0.9	1.0	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0	2.1	2.2	Score
Jumper 1	5	5	5	4	5	5	3	4	2	0	1	0	0	0	0	39
Jumper 2	5	5	5	5	5	5	5	5	4	5	5	5	3	2	0	64
Jumper 3	5	5	5	5	5	5	5	5	0	0	0	0	0	0	0	40
Jumper 4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Jumper 5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	75
Jumper 6	0	0	0	3	3	5	5	5	3	4	2	1	0	0	0	31

illustrate this more dramatically because we cannot estimate their ability at all. We know that Jumper 5 is better than 2.2m and that Jumper 4 is not as good as 0.8m, but we have no way of precisely estimating their ability because the range of our test items is smaller than the range of person abilities.

Jumper 6 illustrates the opposite problem to that of Jumper 3, that of misfit. This person failed on the easy jumps but succeeded on more difficult ones, requiring investigation because this does not conform to the model we have of jumping ability, where low obstacles should be easier to jump over than high obstacles. Perhaps this person misunderstood the task at first but then improved after practice, perhaps the data was recorded incorrectly, or perhaps they cheated when they realized that they could not succeed. Another possibility is that Jumper 6 slept late and missed the first part of the test, only made three attempts at 1.1m, 1.2m, and 1.3m, and then mostly succeeded until 1.7m. If this was so, we would estimate this person to be a better jumper than Jumper 1, even though they have a lower score. In cases like this, where different persons attempt different subsets of test items, the raw score of the number of successful jumps does not allow us to estimate ability unless we know how difficult the jumps were.

Thus, to summarize, Rasch measures are interval estimates of ability of a trait that require success on that trait to be probabilistic, and equal intervals between estimates mean equal intervals in ability. Guttman data may be overly deterministic and can allow less probabilistic ordering of ability, but does not allow for probabilistic measures of ability to the same degree as Rasch models. Classical test theory relies on scores of how many times a candidate succeeded, but these scores may not give useful measures of the underlying trait that we wish to measure. Thus Guttman and CTT scores may only provide ordinal ranking of ability, whereas Rasch estimates generate interval measures of ability.

This transformation of raw scores to interval level measures is achieved through calculation of odds-ratios. For example, a person scoring 80% on a test has an odds-ratio of 4/1, calculated by dividing 80% success by 20% failure, while a person scoring 20% has an odds-ratio of 1/4 and a person scoring 50% has an odds-ratio of 1/1. However, odds-ratios are difficult to manipulate, so these are converted to log odds-ratios, more commonly called “logits”, which can be added

or subtracted easily. The relationship between logit scores and probabilities of success are expressed mathematically as

$$P = \exp(B-D)/(1+\exp(B-D)) \quad (1)$$

where P equals the probability of success, and B, and D respectively equal the logit measures of person ability and item difficulty. When person ability and item difficulty are exactly equal, i.e. the difference between item difficulty and person ability is 0.00 logits, then the probability of success equals 50%, while a person with ability 1.00 logits greater than item difficulty has a probability of success of 73% and a person with ability of -1.00 logits has a probability of success of 27%. Logit scores thus tell us about the relative difference between person ability and item difficulty, not absolute ability. Logits thus represent interval level measures, in a similar manner to the Celsius scale where zero degrees just represents a convenient benchmark, not ratio measures like the meter scale. An important implication of this is that logit scores cannot be calculated when a person fails on all items or succeeds on all items, just as we do not know the jumping ability of a person who succeeds on all jumps or fails on all jumps.

A critical question then is whether the trait or ability we wish to investigate fits the requirements of the model, i.e., is there a stochastic pattern of success increasing with ability? If this is so, then Rasch measures are appropriate. If success is purely random, we could not claim to be measuring anything at all. If success shows strongly deterministic Guttman patterns, then the trait does not meet the requirements of the Rasch model and we would not be able to generate useful measures. Engelhard (2008) proposes that the Rasch measurement is fundamental to the trait of language performance, which is probabilistic, in contrast to a deterministic model of language competence. McNamara (1996) explains Chomsky's distinction between language competence and performance and the crucial developments in describing language performance arising from the work of Hymes (1972), Canale and Swain (1980), and Bachman and Palmer (1996). Taken together with the fundamental role of interlanguage variation in the process of language acquisition outlined by Ellis (1994), we have a model of language proficiency that is fundamentally probabilistic, not a deterministic Guttman model. Thus rather than needing to justify the use of Rasch measurement for language

testing and research, the opposite is the case; the use of non-probabilistic measurement implies investigation of a non-probabilistic trait, but language proficiency is not such a trait.

Rasch Fit Statistics

Despite its conceptual simplicity, the Rasch model provides sophisticated tools to analyze individuals' responses to items, individual items' contribution of information to the whole test, and how the information is organized throughout the test. Rather than relying on a single raw-score summary for each person, the information contributed by each person's response to each item can be considered, providing quality control statistics for persons, items, and the test as a whole (Bond & Fox, 2007; Embretson & Hershberger, 1999; Embretson & Reise, 2000; Engelhard, 2013).

The basis of many of these diagnostic tools is the "score residual", which is the difference between a person's expected score (i.e. probability of a correct response) on an item and their observed score. In a dichotomous test the observed score can only be 0 or 1, but the expected score can have any value between 0 and 1. When person ability equals item difficulty, the probability of success is 50%, so the expected score is 0.50, giving possible score residuals of +/-0.50. As person ability increases relative to item difficulty, the expected score approaches 1.0, so the magnitude of the score residual becomes smaller for a correct response and larger for an incorrect response, while the opposite occurs as person ability decreases relative to item difficulty. Smaller score residuals thus indicate better fit of the data to the Rasch model, while larger score residuals indicate data-model misfit. Score residuals across a dataset are expected to follow a chi-square distribution, and can be calculated for both persons and test items, allowing detailed diagnosis of misfitting persons and items.

Unlike exploratory modeling, which aims to find the model that best fits the observed data, the Rasch model is a confirmatory model which represents a theoretical idealization of measurement, analogous to the meter representing a theoretically perfect definition of length. Physical rulers provide objective measures only to the degree that they can be calibrated against a common scale and, while

they are never perfect, the practical consideration is whether the calibration is sufficiently good for the task at hand. Similarly, Rasch logit measures represent an idealized model of measurement, and fit statistics provide quality control showing how large the distortions in the actual instrument are compared to this.

Although software such as *Winsteps* produces a range of different statistics on the performance of persons and items, the mean-square (MnSq) fit statistic is key to Rasch quality control. The mean-square statistic has a minimum value of 0.00, no upper limit, and an expected mean value of 1.00 that represents perfect fit between the data and the model. Values less than 1.00 indicate over-predictability, or overfit, while values greater than 1.00 indicate unpredictability, or misfit. The range of acceptable fit varies with context and type of measurement instrument, but typically values less than 1.25 are considered well-fitting, values greater than 1.50 require investigation, and values greater than 2.00 indicate a lack of effective measurement. It's also important to remember that overfit is also of concern as it indicates data with less stochastic variation than expected, which reduces the amount of information available to estimate logit measures, resulting in muted measurement. To further confuse matters, fit statistics are reported as both infit and outfit values. Infit values are information weighted, emphasizing responses where person ability is well matched to item difficulty, and thus are a more important indicator of distorted measurement, while outfit values are unweighted and thus provide diagnostic information about unexpected outlying responses.

Many-faceted Rasch Measurement

Rasch's original model was developed for analysis of dichotomous data, but subsequent extensions allowed analysis of Likert type rating scales, where responses can be assigned a range of intermediate values falling between total rejection and total endorsement, and of many-faceted data, where factors such as the severity or lenience of judges must be considered in addition to the facets of person ability and item severity (Linacre, 1994). For judge mediated performance assessments, rater severity must be added to Equation 1, resulting in

$$P = \exp(B-R-D)/(1+\exp(B-R-D)) \qquad (2)$$

where R equals rater severity. Although the Rasch model initially generated criticism

as an over-simplification of complex traits, many-faceted Rasch measurement (MFRM) is now a standard procedure in analyzing judge mediated performance assessments (McNamara, 1996; McNamara & Knoch, 2012). As well as providing logit measures that are adjusted for the severity of different raters, MFRM provides several important advantages over simple raw-score analysis: person ability can be easily criterion referenced because the difficulty of rubric items is reported in the same units as person ability, rater and rubric item performance can be monitored through data-model fit, and diagnosis of individual students in need of remediation is possible through person fit statistics and analysis of unexpected responses.

Worked Example 1: Analyzing a Multiple-choice Test using *Winsteps*

The sample data files needed to replicate this example, a step-by-step guide to importing data into *Winsteps*, and authorized versions of the free *Ministeps* and *Minifacs* software are available for download from:
http://db.tt/CDDvTjWH

The *Ministeps* installation file is in the "Software" directory. This is identical to the full *Winsteps* software package but is limited to 25 items and 75 persons. The example data file contains 25 items and 75 persons, so *Ministeps* can be substituted for *Winsteps* to run the example analysis. The data files for this example are in the "Winsteps Example Files" directory and instructions on importing data into *Winsteps* and setting up a control file are given in Supplement A.

To start the analysis, drag the file "07 Vocab Test Control" onto the *Winsteps* icon on your desktop and *Winsteps* will launch. Press "Enter" to choose a temporary file name, then "Enter" again because we don't need extra specifications. *Winsteps* presents an initial report with some useful data in the table at the bottom. These summary statistics are also available by clicking on the "Output Tables" menu, then "3.1 Summary Statistics." *Winsteps Table 3.1* is reproduced in Table 2. We can see that there were 75 persons on the test and 25 items. However, 3 persons had extreme scores, either 0% or 100%, so the non-extreme person results are most useful here. The mean person measure (i.e. the person ability)

is 56.81 with a mean error of 6.19. The mean item ability is 50.00, an arbitrary value specified when setting up the control file, with an error of 3.24. We can see that the average person measure is better than the average item difficulty, so when a person of mean ability tries a question of mean difficulty, they have a higher probability of success than failure.

If we go to the "Edit" menu and click "Edit control file=…" the control file "07 Vocab Test Control" opens. This is a text file. At the bottom is the data from the test. Above that are the item labels, in this case the correct answer for each question. At the top are the instructions for *Winsteps* about how to read the data. The specification "UIMEAN=50" tells *Winsteps* to assume that the average question has a difficulty of 50. We can set this to any number that is convenient. Researchers usually set it to 0, but for classroom tests, figures in the range of 50 to 100 are easier to understand. We can also see that "USCALE=10". This means that 1 logit is scaled to 10 for this test. That means that the difference between mean person ability and mean item ability is 0.681 logits ((56.81 - 50.00) / 10.00), which is a substantive difference.

There is another way to change the specifications. Close the control file and click "Data Setup". Usually it's easier to create the original control file here and then edit it in *Notepad* later. At the top right we can see the test has been set with "Set item mean" at a value of 50 and "Units per logit" to 10. Close the data setup window and return to *Winsteps*.

We also have some other information in the summary report shown in Table 2. There are reports on "fit", how well the data fitted the model, estimates of person reliability (how reliably the items estimated the ability of the persons) and item reliability (how well the persons allow estimates of the item difficulty). The person reliability of 0.87 is good for such a short test, and we see person separation estimated at 2.53. This means that we can be reasonably confident that this test can separate the persons into 2 groups, but less so that it can separate them into 3 groups. If we used this test as a placement test, we could be confident that the highest students are better than the lowest students, but not that the average students are better or worse than the low or high students.

We should be concerned about the fit statistics though. For persons, the

Table 2 ***Winsteps Table 3.1 Summary Statistics***

```
Table 3.1 Vocab Test Control
INPUT: 75 Person  25 Item  REPORTED: 75 Person  25 Item  2 CATS  WINSTEPS 3.80.0
-------------------------------------------------------------------------------

SUMMARY OF 72 MEASURED (NON-EXTREME) Person
-------------------------------------------------------------------------------
           TOTAL                          MODEL          INFIT        OUTFIT
           SCORE     COUNT     MEASURE    ERROR      MNSQ   ZSTD   MNSQ   ZSTD
 ----------------------------------------------------------------------------
  MEAN      14.7      24.8       56.81     6.19       .94     .0   1.15     .2
  S.D.       6.9        .7       19.18     2.23       .26     .9   1.46    1.2
  MAX.      24.0      25.0       89.09    11.15      1.49    2.9   9.90    4.9
  MIN.       1.0      20.0       12.66     4.41       .45   -2.3    .08   -1.2
 ----------------------------------------------------------------------------
  REAL RMSE   6.80 TRUE SD   17.94  SEPARATION  2.64  Person RELIABILITY  .87
 MODEL RMSE   6.58 TRUE SD   18.02  SEPARATION  2.74  Person RELIABILITY  .88
  S.E. OF Person MEAN = 2.28
-------------------------------------------------------------------------------
  MAXIMUM EXTREME SCORE:      3 Person 4.0%


SUMMARY OF 75 MEASURED (EXTREME AND NON-EXTREME) Person
-------------------------------------------------------------------------------
           TOTAL                          MODEL          INFIT        OUTFIT
           SCORE     COUNT     MEASURE    ERROR      MNSQ   ZSTD   MNSQ   ZSTD
 ----------------------------------------------------------------------------
  MEAN      15.1      24.9       58.65     6.70
  S.D.       7.0        .7       20.86     3.32
  MAX.      25.0      25.0      103.01    18.96
  MIN.       1.0      20.0       12.66     4.41
 ----------------------------------------------------------------------------
  REAL RMSE   7.66 TRUE SD   19.40  SEPARATION  2.53  Person RELIABILITY  .87
 MODEL RMSE   7.48 TRUE SD   19.48  SEPARATION  2.60  Person RELIABILITY  .87
  S.E. OF Person MEAN = 2.43
-------------------------------------------------------------------------------
Person RAW SCORE-TO-MEASURE CORRELATION = .97
CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY = .93


SUMMARY OF 25 MEASURED (NON-EXTREME) Item
-------------------------------------------------------------------------------
           TOTAL                          MODEL          INFIT        OUTFIT
           SCORE     COUNT     MEASURE    ERROR      MNSQ   ZSTD   MNSQ   ZSTD
 ----------------------------------------------------------------------------
  MEAN      45.4      74.6       50.00     3.24      1.00    -.1   1.17     .0
  S.D.      11.2        .6       11.81      .36       .27    1.5   1.46    1.3
  MAX.      68.0      75.0       85.84     4.56      1.86    3.2   7.94    3.9
  MIN.      13.0      73.0       23.10     3.02       .67   -2.6    .43   -1.8
 ----------------------------------------------------------------------------
  REAL RMSE   3.45 TRUE SD   11.30  SEPARATION  3.28  Item   RELIABILITY  .91
 MODEL RMSE   3.26 TRUE SD   11.35  SEPARATION  3.48  Item   RELIABILITY  .92
  S.E. OF Item MEAN = 2.41
-------------------------------------------------------------------------------
Item RAW SCORE-TO-MEASURE CORRELATION = -1.00
1789 DATA POINTS. LOG-LIKELIHOOD CHI-SQUARE: 1501.29 with 1693 d.f. p=.9997
Global Root-Mean-Square Residual (excluding extreme scores): .3676
Capped Binomial Deviance = .1748 for 1864.0 dichotomous observations
UMEAN=50.0000 USCALE=10.0000
```

average of the mean-square infit is 0.94, which is below the expected value of 1.00, but the outfit value is 1.15. The standard deviations are respectively 0.26 and 1.46. From this we can see that some persons did not behave very predictably. We can see a similar pattern in the item fit statistics, with infit mean-square average of 1.00 but outfit of 1.17 and standard deviations of 0.27 and 1.46 respectively. The

standard deviations of the outfit statistics are extremely high, so either some low ability persons succeeded on some difficult items or some high ability persons failed on some easy items.

Distribution of Item Difficulty

Very important information is displayed in the Rasch item map, often called a "Wright Map" after Ben Wright. On the Output Tables menu click "1 Variable Maps", reproduced in Table 3. This shows graphically how the candidates compare to the items; it's a picture of how high they can jump, mapping persons on the left against items on the right. The easiest items and least proficient persons are at the bottom, and the most difficult items and most proficient persons at the top.

Looking at the Wright map in more detail, we can see that Item 11 "dozen" is the most difficult and Item 22 "difficult" the easiest. We can see that Item 17 has a difficulty of about 65, or 1.5 logits, Item 8 has difficulty of 50, or 0 logits, and Item 9 has difficulty of about 35, or -1.5 logits. The interval in difficulty between items 17 and 8 and between items 8 and 9 are about the same. We can also see that there are not enough difficult items. There are many persons of much higher ability than the most difficult item, so we don't really know how much ability these persons have. We also have some big gaps between the easier items, so we can see that this test needs to be modified. We need less items of average difficulty, some much more difficult items, and we need to fill in the gaps between the items. It's very likely that the mismatch between the ability of the persons and the difficulty of the items is responsible for much of the misfit we saw in Table 2 (*Winsteps Table 3.1*).

Point-Measure Correlation

We've seen that this test is mismatched to the sample of persons but we need more information about the quality of the information we have. *Winsteps* has a lot of very sophisticated analyses that are well beyond what we can discuss here, so we'll just cover a few fundamental things. First we need to check to see if we have any really bad items. On the Diagnosis menu, click "A Item Polarity". This brings up *Winsteps Table 26.1*, reproduced in Table 4, which arranges the items in order of

Table 3 ***Winsteps Table 1.0 Person-Item Map***

```
TABLE 1.0 Vocab Test
INPUT: 75 Person  25 Item  REPORTED: 75 Person  25 Item  2 CATS  WINSTEPS 3.80.0
-------------------------------------------------------------------------------
 MEASURE                                  Person - MAP - Item
                                                <more>|<rare>
   90                                1031 1037 1049  +
      1034 1035 1038 1039 1042 1043 1046 1047 1048  |
                                                    |
                                                    |  dozen
                                                    |
                                                    |
                                                    |
                                                    |
   80                                1007 1014 1041  +
                                                    |
                                                    |
                                                   S|
               1004 1017 1020 1032 1036 1045 1050   |
                                                   T|
                                                    |
   70                                      1001  +
                                                    |
                                                    |
                               1010 1012 1040 1044  |
                                                    |  arrange
                                     1013 1027      |  roar
                                                    |  admire
                                           1024    |S
   60                                               +
                               1015 1021 1022 1028  |
                                                    |  debt
                                     1016 1029 M    |
                     1005 1006 1011 1019 1025 1026  |  opportu
                                                    |  skirt
                               1003 1008 1030 1067  |  introdu wage
                                     1018 1033      |  justice lovely  prefer
   50                                              +M clerk
                                     1051 1072      |  elect   tax
                                     1002 1061      |
                                1009 1054 1066      |  melt
                                     1060 1070      |  ancient develop stretch
                                                    |  holy    pride
                                           1057     |  manufac
                                1069 1071 1074      |
   40                                              +
                                1023 1068 1075     |S
                                                   S|
                                1053 1056 1073      |  noise
                                                    |
                                           1065     |
                                                    |  wine
   30                                               +
                                1052 1055 1062      |
                                                    |
                                                   T|
                                           1064     |
                                                    |
                                                    |  difficu
                                                    |
   20                                1058 1059      +
                                                   T|
                                                    |
                                                    |
                                                    |
                                                    |
                                           1063     |
   10                                               +
                                           <less>|<frequent>
```

point-measure correlation. This is the correlation of each item to the overall measure of the test. A value of 0 means the relationship between the item and the overall test is random; some high ability persons got it correct, some low ability persons got it correct, and it doesn't tell us who is high or low ability. Negative values mean that low ability people got it correct more than high ability people, which is a big problem. It may be a badly written item or it might test a different trait than the overall test, maybe an item needing cultural knowledge in a grammar test, for example. The Pt-Measure figures for this test are generally very good, all the values are positive and above 0.4 except for "dozen", which is 0.21. This is fairly encouraging, but we might want to look more closely at Item 11 to see why it has such a low correlation. What is obvious about this item is that it is the most difficult item in the test, it has a "measure" or difficulty of 85. It also has terrible fit statistics, with infit mean-square of 1.86 and outfit of 7.94. It has a raw score of 10, so only 10 candidates got it right. If one or two low candidates got lucky on this item, it would make a big difference to the correlation and to the fit statistics, so we need to be careful about interpreting point-measure correlations for very difficult

Table 4 ***Winsteps Table 26.1 Item Point-measure Correlation***

TABLE 26.1 Vocab Test Control
INPUT: 75 Person 25 Item REPORTED: 75 Person 25 Item 2 CATS WINSTEPS 3.80.0

Person: REAL SEP.: 2.53 REL.: .87 ... Item: REAL SEP.: 3.28 REL.: .91
Item STATISTICS: CORRELATION ORDER

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MODEL MEASURE	S.E.	INFIT MNSQ	ZSTD	OUTFIT MNSQ	ZSTD	PTMEASURE CORR.	EXP	EXACT MATCH OBS%	EXP%	Item
11	13	75	85.84	4.06	1.86	3.2	7.94	3.9	.21	.58	76.4	88.0	dozen
22	68	75	23.10	4.56	.75	-.8	.43	-.5	.46	.38	94.4	91.3	difficult
16	51	75	45.17	3.12	1.30	1.9	1.18	.5	.48	.57	69.4	79.1	develop
17	30	75	64.88	3.17	1.38	2.1	2.65	3.1	.48	.65	76.4	80.7	arrange
7	62	74	32.02	3.79	.99	.0	.57	-.3	.49	.47	84.5	86.5	wine
15	32	74	62.59	3.15	1.43	2.4	1.43	1.2	.51	.65	67.6	80.1	admire
25	45	75	50.81	3.03	1.24	1.6	1.30	.9	.52	.61	73.6	77.4	lovely
19	54	75	42.17	3.21	1.01	.1	.75	-.3	.56	.55	77.8	80.4	manufacture
21	50	75	46.13	3.10	1.09	.7	.81	-.3	.56	.58	70.8	78.6	melt
20	47	75	48.96	3.05	1.11	.8	.92	-.1	.57	.60	73.6	77.7	elect
13	51	75	45.17	3.12	.88	-.8	1.48	1.1	.57	.57	86.1	79.1	stretch
18	44	75	51.73	3.03	1.11	.8	1.09	.4	.58	.61	72.2	77.3	prefer
9	60	75	35.49	3.50	.77	-1.3	.46	-.7	.58	.50	87.5	84.0	noise
23	51	74	44.60	3.17	.91	-.6	.67	-.6	.61	.57	81.7	79.6	ancient
24	52	74	43.81	3.19	.86	-.9	.62	-.7	.62	.56	84.5	79.9	holy
2	37	74	57.75	3.08	.99	.0	1.24	.8	.64	.64	73.2	78.4	debt
6	45	75	50.81	3.03	.91	-.6	.74	-.7	.64	.61	79.2	77.4	justice
3	53	75	43.19	3.17	.74	-1.8	.49	-1.1	.66	.56	86.1	80.0	pride
4	43	74	52.12	3.06	.89	-.7	.77	-.6	.66	.62	77.5	77.5	wage
1	30	74	64.60	3.20	.96	-.2	.89	-.2	.66	.65	81.7	80.8	roar
10	40	74	55.24	3.05	.91	-.6	.72	-.8	.66	.63	77.5	77.2	opportunity
12	47	73	48.25	3.13	.71	-2.1	.53	-1.3	.70	.59	87.1	78.4	tax
8	46	75	49.89	3.04	.75	-1.9	.53	-1.4	.70	.60	80.6	77.6	clerk
14	43	74	52.12	3.06	.73	-2.0	.58	-1.3	.71	.62	83.1	77.5	introduce
5	42	75	53.56	3.02	.67	-2.6	.48	-1.8	.74	.62	86.1	77.1	skirt
MEAN	45.4	74.6	50.00	3.24	1.00	-.1	1.17	.0			79.5	80.1	
S.D.	11.2	.6	11.81	.36	.27	1.5	1.46	1.3			6.5	3.6	

or easy items.

Below *Winsteps Table 26.1* we find *Winsteps Table 26.3* which shows how the distractors for each item functioned, a sample of which is reproduced in Table 5. *Winsteps* provides this information for all items but, for the sake of contrast, Table 5 only shows the results for Item 11 (the worst performing item) and Item 5 (the best performing item). In the "Data Code" column, we can see that each item had five response options, "A" to "E". The correct answer, or "key", for Item 11 was "A" and the key for Item 5 was "D", so these two responses have score values of 1 while the other response options, or "distractors", have score values of 0. The data count columns show how many responses were recorded for each response option as both a raw count and a percentage. The expected pattern in a multiple-choice test is that the distractors attract different numbers of responses depending on how convincing they are to persons at different levels of ability. Item 5 shows 56% of persons choosing the correct answer, and smaller numbers choosing each distractor. Importantly, the average ability of the persons choosing the key was 72.28, a much higher figure than for any of the distractors. This tells us that higher ability persons were not fooled by the distractors very often, indicating that this item is functioning effectively. Item 11 shows nearly equal numbers of persons choosing each response option though, with a much smaller difference between

Table 5 ***Winsteps Table 26.3 Item Distractor Frequencies***

TABLE 26.3 Vocab Test Control
INPUT: 75 Person 25 Item REPORTED: 75 Person 25 Item 2 CATS WINSTEPS 3.80.0

Item CATEGORY/OPTION/DISTRACTOR FREQUENCIES: CORRELATION ORDER

ENTRY NUMBER	DATA CODE	SCORE VALUE	DATA COUNT	%	AVERAGE ABILITY	S.E. MEAN	OUTF MNSQ	PTMA CORR.	Item
11	C	0	13	17	52.20	5.12	1.0	-.14	dozen
	E	0	12	16	55.64	6.64	2.0	-.06	
	D	0	12	16	56.81	6.59	1.8	-.04	
	B	0	13	17	58.33	5.35	1.5	-.01	
	F	0	12	16	60.45	4.63	1.8	.04	
	A	1	13	17	68.26	7.06	9.0	.21	
			23 Items Omitted for Clarity						
5	A	0	7	9	34.19	3.56	.2	-.38	skirt
	E	0	8	11	38.55	5.49	.6	-.33	
	C	0	8	11	41.71	3.80	.4	-.28	
	B	0	7	9	45.69	3.21	.5	-.20	
	F	0	3	4	54.04	2.67	.9	-.05	
	D	1	42	56	72.28	2.46	.5	.74	

Missing % includes all categories. Scored % only of scored categories

the persons choosing the key and those choosing the distractors. This pattern shows an item that is not working effectively so we need to investigate to see if there was a problem with data entry, printing of the answers sheets, etc.

Item Fit

Another important diagnosis available is the item pathway bubble charts, introduced by Bond and Fox (2007). On the Plots menu, click "Bubble Chart (Pathway)". Let's look at the infit mean-square chart for the items as an example. Check the "Items" box, the "Infit" button, and the "Mean-square (chi-square/d.f.)" button, then "OK". A dialogue box will pop up asking what labels you want to use. In this case, let's choose "Label", the name we gave the item will be displayed. *Winsteps* will now create an Excel bubble chart of the item infit, reproduced in Figure 1. The vertical axis shows item difficulty, so "dozen" is the most difficult item, and "difficult" is the easiest. The size of the bubbles shows the standard error of each item. If we hover our mouse pointer over "dozen", we can see that the bubble has a size of 4.06, the item difficulty is 85.84, and the fit is 1.86. The confidence interval of the item difficulty is plus or minus 2 standard errors, so we can be 95% confident that the item difficulty is between 77.72 and 93.96. However the scale on the left is unreadable, so we need to right click on it and select "format axis", then change the major unit interval to 10. The size of the bubbles is too big, we want "dozen" to be about 8 units in diameter, but it's about 20 units. Right click inside the bubble, then click "Format Data Series" and change the bubble size to 40%. This gives a better indication of the standard error in relation to the difficulty. Although "roar" is probably more difficult than "debt", the bubbles overlap slightly, so we can't be 95% confident that it actually is more difficult. We can be quite confident that "roar" is more difficult than "justice", though.

Fit is a much too complex and confusing topic to discuss in-depth here, but a rule-of-thumb for mean-square values is that between 0.8 and 1.3 is good. Values over 2.0 will degrade measurement, so it may be necessary to remove items or persons for test analysis (persons can later be returned before calculating final grades). Mean-square values always average about 1, so if we have items with a high value, they will cause other items to have a low value. High values mean

there is too much unpredictability in the data, maybe it was a bad question and low students were able to guess it or high students were confused, or maybe students cheated. If the misfit is too bad, we might remove the item from the test, then redo the analysis and see how the remaining items function. "Dozen" has a very high misfit value. Look back at the raw data in the Excel file "00 Vocab Test Responses" and see if you can find why. (Hint: the responses to this item have been deliberately changed to cause it to misfit, the pattern should be easy to see.)

The help menu in *Winsteps* gives a lot of advice about diagnosing misfit. A summary of that advice is: (1) First investigate negative point-measure correlations. Fix data entry problems, miskeys, etc. (2) Investigate outfit before infit.(3) Investigate mean-square before t standardized. (4) Investigate high values (too unpredictable) before low values (too predictable). In-depth information about fit is available in *Winsteps Table 26.1* (and in other tables accessible from the Output Tables menu). Four different values are given, two for "Infit" and two for "Outfit". Infit tells us how predictable the responses are on items close to the person's

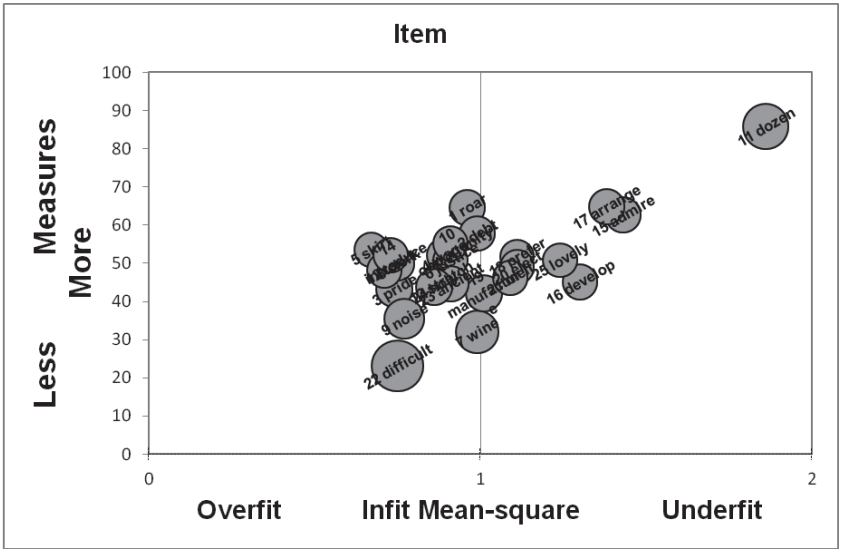

Figure 1 ***Winsteps*** **pathway bubble chart output of item infit. The vertical axis shows item difficulty, the horizontal axis shows item infit mean-square values, and the size of the bubbles shows the 95% confidence interval of item difficulty.**

ability, so too much unpredictability is a serious problem. The MNSQ, mean-square, figures show the size of the misfit and are expected to be around 1.0. Values greater than about 1.3 mean the responses are becoming too unpredictable, the data is misfitting the model, and values less than about 0.7 mean the responses are becoming too predictable and the data overfits the model. Unpredictability is much more serious than being too predictable, so we try to remedy problems with misfit first. The ZSTD values indicate the statistical significance of the fit. Plus or minus 2 means that the misfit is becoming statistically significant. Small misfit MNSQ values that are statistically significant are not usually a problem, but large values will need investigation.

Outfit values show how responses to items much easier or more difficult than the person's ability fit the model, so they are easier to diagnose. On the "Output Tables" menu, click "10 Item Column Fit Order". Scroll down to *Winsteps Table 10.5*, reproduced in Table 6. This shows the most unexpected responses, i.e.

Table 6 ***Winsteps Table 10.5 Most Unexpected Responses***

```
TABLE 10.5 Vocab Test Control
INPUT: 75 Person  25 Item  REPORTED: 75 Person  25 Item  2 CATS  WINSTEPS 3.80.0
-------------------------------------------------------------------------------

MOST UNEXPECTED RESPONSES
Item           MEASURE  |Person
                        |441 543321  44112121216 3156 655776726755665565 56
                        |3147056207414020732559733811264741953836352524983
                     high----------------------------------------------------
22 difficult      23.10 e|.....................0.....................
 7 wine           32.02 L|.....................0....0.0..... ..........
 9 noise          35.49 g|.....................0.....................
19 manufacture    42.17 K|.................0.........................1..
24 holy           43.81 h|.................0..........................
23 ancient        44.60 j|...............0.00..... ....................
13 stretch        45.17 C|..0..0.....................................1
16 develop        45.17 E|....0.....................................1.11...
21 melt           46.13 J|.............0.............................
12 tax            48.25 b|.......................... ..... .....1.....
20 elect          48.96 I|...............00.....................1...1...
 8 clerk          49.89 f|.....................................1........
 6 justice        50.81 l|.............0.....................1...1.....
25 lovely         50.81 F|......0000.0..0.....................
18 prefer         51.73 H|............00.....................1.......1..
 4 wage           52.12 i|.........0.............................1....
14 introduce      52.12 c|.........0.....................
10 opportunity    55.24 k|...........0.................1..1 1..........
 2 debt           57.75 G|.............................1.............1.
15 admire         62.59 D|.0.0.....................1..1..111.1..........
 1 roar           64.60 M|.....................1.1..111.................
17 arrange        64.88 B|0..0.....................1.11....1.1.1........1.
11 dozen          85.84 A|..........1.....1..111....1.........1....1.....
                        |-------------------------------------------------low
                        |441754332141441121212163315626557767267556655 6
                        |314 056207  40207325597 3811 64741953836352524983
```

when low ability persons succeeded on difficult items or high ability persons failed on easy items. The vertical axis shows items, with the easiest items higher on the axis, while the horizontal axis shows persons, with the highest ability persons on the left and the lowest ability persons on the right. Person #43 was thus the highest ability and Person #63 was the lowest. The top left corner therefore matches high ability persons with easy items, so we expect scores of 1 in this area. The bottom right corner matches low ability persons with difficult items, so we expect scores of 0 in this area. Responses where the observed result matched the expected result are marked by a ".", while unexpected failure is marked by a "0" and unexpected success is marked by "1".

We can use this table to go back to the answer sheets and check for problems, maybe the scanner misread an answer, maybe the candidate guessed a difficult item. Often we will recode the responses to "missing data". Doing this with responses that show poor outfit will improve estimates of person ability and the quality of the measures overall, but infit is much more difficult to remedy. Look at the outfit for "Dozen", it is enormous, low students are getting it correct as often as high students. This item needs to be removed from the test. Look at the responses for persons #2, #58, and #75 in the Excel file "00 Vocab Test Responses". #58 has answered A, B, C, D, in a fixed pattern. This person has very bad outfit. #2 and #75 got confused and marked the answer for questions 1 to 11 in the wrong boxes, but then started answering correctly, yet their fit statistics don't ring any alarm bells. Perhaps so many other students answered randomly that these two don't stand out as unusually unpredictable.

Comparing Groups of Students

Often we want to know if different groups of students are significantly different in their ability. A common tool for this is the *t*-test. *Winsteps* can conduct simple *t*-tests if we label each student with a group code. The person codes in this analysis are eight characters long. The first four characters are a student number, followed by a space and then a gender code, then another space and a group code. The students have been placed into class groups using a placement test and the group codes for the three levels are "L", "H", and "K". In *Winsteps*, click on the

"Output Tables" menu, then "28. Person Subtotals". You will be prompted for the information about where the group code is located in the person label by a text box that says "PSUBTOTAL = $S1W1". The first part of this, "$S1", means character 1 in the label and the second part, "W1", means one character wide. Our group labels are one character wide and start at the eighth character in the person label so we need to change that to read "$S8W1" and then click "OK". This will open a text file with the results of the *t*-test, reproduced in Table 7.

The first section of Table 7 shows a summary of the groups and the combined scores, with a mean ability of 59.82 for the "H" group, 83.62 for the "K" group, and 37.29 for the "L" group. The next table shows the difference between each pair of groups and whether the difference is statistically significant. The "H" group was 23.80 lower than the "K" group, with a t = -6.79 and p = .000. This means that it is very, very unlikely that the mean score of these two groups were different by

Table 7 ***Winsteps Table 28.1 t-test of Person Groups***

```
TABLE 28.1 Vocab Test Control
INPUT: 75 Person  25 Item  REPORTED: 75 Person  25 Item  2 CATS  WINSTEPS 3.80.0
-------------------------------------------------------------------------------
EXTREME AND NON-EXTREME Person SCORES
-------------------------------------------------------------------------
Person  MEAN     S.E.  OBSERVED  MEDIAN  MODEL      MODEL
COUNT  MEASURE   MEAN    S.D.          SEPARATION RELIABILITY CODE
-------------------------------------------------------------------------
   75   58.65    2.43   20.86   54.62    2.60        .87     *
   30   59.82    1.82    9.78   57.80    1.52        .70     H
   20   83.62    3.00   13.06   89.09     .59        .26     K
   25   37.29    2.02    9.88   38.93    1.45        .68     L
-------------------------------------------------------------------------
SUBTOTAL RELIABILITY=.98
UMEAN=50 USCALE=10

-------------------------------------------------
Person    MEAN DIFFERENCE        Welch-2sided
CODE CODE MEASURE   S.E.    t    d.f.  Prob.
-------------------------------------------------
H    K     -23.80   3.50  -6.79     32  .000
H    L      22.53   2.71   8.30     50  .000
K    L      46.33   3.61  12.83     34  .000
-------------------------------------------------

-----------------------------------------------------------------
ANOVA - Person
Source  Sum-of-Squares   d.f.  Mean-Squares  F-test  Prob>F
-----------------------------------------------------------------
$S8W1         23919.97    2.00     11959.99   98.73   .0000
Error          8721.56   72.00       121.13
Total         32641.53   74.00       441.10
-----------------------------------------------------------------
Fixed-Effects Chi-Square: 175.3126 with 2 d.f., prob. .0000
-----------------------------------------------------------------
```

chance, so the difference is statistically significant. Although *t*-tests are one of the most common statistical analyses, they need to be interpreted very carefully when comparing multiple groups so you are recommended to refer to references such as Field (2009) for a detailed explanation of how to interpret these results.

Worked Example 2: Analyzing Data with *Facets*

The sample data files needed to replicate this example, a step-by-step guide to constructing a *Facets* specification file, and authorized versions of the free *Ministeps* and *Minifacs* software are available for download from:
http://db.tt/CDDvTjWH

The *Minifacs* installation file is in the "Software" directory. This is identical to the full *Facets* software package except that it is limited to 2000 responses. The example data file contains 1944 responses, so *Minifacs* can be substituted for *Facets* to run the example analysis. The data files for this example are in the "Facets Example Files" directory.

Once you have installed *Facets*, drag the file "01 Presentation Specification with Data" onto the *Facets* icon on your desktop and *Facets* will open the file and ask you for extra specifications. Click "OK" and you will be asked to choose where to save the output file. Save it in the same directory as the specification file. Now *Facets* will analyze the data and produce an output file in plain text format. You can see that this file is called "01 Presentation Specification with Data.out", which is the same name as the specification file, but with ".out" appended. This makes it easy to match the output files to the specification files later on.

This dataset is from practice presentations that were used to introduce students to a rating rubric. There are 19 students, S1 to S19. There were also 19 student raters, R1 to R19, plus three sets of ratings from the classroom teacher, Ta, Tb, and Tc. Rater Ta represents ratings done live in class while Rater Tb and Tc represent ratings of video recordings of the presentations. The rating rubric had nine items, which means that three different things interacted to produce a rating: a rater judged a person against an item to produce a rating from 0 to 3. Therefore we need a 3-faceted analysis. Although the teacher rated all the students, students did not

rate their own performance and some students did not provide ratings for all the other students, so we cannot just add up raw scores because no student had the same set of raters and different raters may have used the rubric differently. Therefore we need to investigate rater performance and adjust for rater leniency.

Facets produces a single plain text output file that contains all the results of the analysis. It's very convenient to have all the results presented like this, but it's also a bit overwhelming at first. A good place to start is *Facets Table 6*, the variable map, reproduced in Table 8. This looks very similar to the Wright maps produced by *Winsteps*, except that it contains a column for raters in addition to students and items. Looking at the items column, we can see that "Eyes" (i.e. eye contact) was the most difficult item and "Intonation" was the easiest. Overall, the students were rated very highly, only eye contact was difficult for most of them. However, the range of rater severity was huge, more than three logits. We definitely cannot use raw scores to measure student ability.

Facets Table 7.2.1, reproduced in Table 9, shows the rater measurement report arranged in order of rater severity. Rater R8 was the most severe rater. This rater's average rating was 1.67 compared with an average of 2.13 for all raters. This corresponds to a logit measure of 1.18 logits compared with 0.00 logits for all raters. The most lenient rater was R2, with an average rating of 2.71 and a logit measure of −2.07. These are enormous differences that make it clear that the raw ratings from different raters are not interchangeable. Another interesting point is that the teacher's ratings, Ta, Tb, and Tc, changed in severity between the different sessions, showing that we cannot assume that rater's performances are stable between different rating sessions.

Facets Table 7.2.2, reproduced in Table 10, also details the performance of raters but is arranged in order of misfit. Two raters, R14 and R16, show misfit large enough to be of concern, with mean-square values exceeding 1.50. Also, there are several raters who over-fit the model, with values much less than the expected 1.00. A likely cause of this is holistic rating, where raters can identify overall good or weak performances, but assign similar ratings on all rubric items instead of considering each item independently. The resulting lack of variance in the data means that there is less information available about the performances, so highly

Table 8 ***Facets Table 6.0 All Facet Vertical Rulers***

```
Rated Performance Test
Table 6.0  All Facet Vertical "Rulers".
Vertical = (1A,2*,3A,S) Yardstick (columns lines low high extreme)= 0,8,-3,3,End
+---------------------------------------------------------------------+
|Measr|+Students      |-Raters|-Items                          |Scale|
|-----+--------------+-------+--------------------------------+-----|
|   3 +              +       +                                + (3) |
|     |              |       |                                |     |
|     |              |       |                                |     |
|     |              |       |                                |     |
|     |              |       |                                | --- |
|     | S19  S8      |       |                                |     |
|     | S11          |       |                                |     |
|     |              |       |                                |     |
|   2 + S16          +       +                                +     |
|     | S10          |       |                                |     |
|     | S13  S9      |       |                                |     |
|     | S2   S7      |       | Eyes                           |     |
|     | S1   S14  S3 |       |                                |     |
|     | S12  S15     |       |                                |     |
|     | S17  S5      |       |                                |     |
|     |              | *     |                                |  2  |
|   1 + S4           + *     +                                +     |
|     | S6           | **    |                                |     |
|     | S18          | *     | Hands                          |     |
|     |              |       | Pausing                        |     |
|     |              | *     |                                |     |
|     |              | ****  | Volume                         |     |
|     |              | *     |                                |     |
|     |              | **    |                                |     |
| *   0 *            * **    *                                * --- *
|     |              | *     |                                |     |
|     |              |       | Confidence     Speed           |     |
|     |              |       |                                |     |
|     |              | *     |                                |     |
|     |              | *     |                                |     |
|     |              |       | Body                           |     |
|     |              |       | Notes/reading                  |     |
|  -1 +              + *     +                                +     |
|     |              | *     |                                |  1  |
|     |              |       |                                |     |
|     |              |       | Intonation                     |     |
|     |              |       |                                |     |
|     |              |       |                                |     |
|     |              |       |                                |     |
|     |              |       |                                |     |
|  -2 +              + *     +                                +     |
|     |              | *     |                                |     |
|     |              |       |                                |     |
|     |              |       |                                | --- |
|     |              |       |                                |     |
|     |              |       |                                |     |
|     |              |       |                                |     |
|     |              |       |                                |     |
|  -3 +              +       +                                + (0) |
|-----+--------------+-------+--------------------------------+-----|
|Measr|+Students     | * = 1 |-Items                          |Scale|
+---------------------------------------------------------------------+
```

over-fitting raters are undesirable. Overall, *Facets Table 7.2.1* and *Facets Table 7.2.2* have shown us that rater performance is of concern here; many of these students do not seem to be using the rubric in the same way as the teacher.

However, we need to know that the teacher was consistent in his use of the rubric before we can diagnose problems with student raters. The specification file “03

Table 9 ***Facets Table 7.2.1 Raters Measurement Report Arranged by Severity***

Rated Performance Test
Table 7.2.1 Raters Measurement Report (arranged by mN).

Total Score	Total Count	Obs Ave	Fair(M) Ave	Logit	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim Discr	Corr PtMea	Corr Exp	Exact Agree Obs %	Exact Agree Exp %	Raters
135	81	1.67	1.70	1.18	.17	1.08	.5	1.08	.5	.90	.19	.53	40.0	36.8	11 R8
125	72	1.74	1.75	1.06	.18	1.13	.8	1.10	.6	.89	.64	.53	39.3	38.4	18 R15
308	171	1.80	1.82	.90	.12	.92	-.8	.93	-.6	1.09	.67	.54	41.3	39.4	2 Tb
153	81	1.89	1.84	.86	.17	1.69	3.8	1.67	3.8	.27	.63	.53	41.6	41.0	17 R14
120	63	1.90	1.87	.79	.20	1.05	.3	1.04	.2	.90	.22	.54	38.9	41.2	10 R7
160	81	1.98	2.01	.46	.18	.84	-1.0	.84	-1.0	1.16	.38	.51	43.5	41.5	14 R11
170	81	2.10	2.02	.42	.18	.46	-4.5	.48	-4.3	1.61	.34	.52	48.1	43.7	7 R4
125	63	1.98	2.04	.36	.20	1.11	.7	1.13	.7	.83	.52	.53	38.3	42.2	16 R13
328	162	2.02	2.05	.35	.13	1.15	1.4	1.16	1.4	.84	.64	.53	42.5	42.7	3 Tc
165	81	2.04	2.05	.35	.18	.76	-1.6	.78	-1.5	1.24	.40	.53	43.9	42.6	12 R9
165	81	2.04	2.07	.29	.18	.84	-1.1	.86	-.9	1.12	.19	.53	42.9	42.7	9 R6
152	72	2.11	2.11	.19	.19	.75	-1.6	.80	-1.3	1.22	.30	.52	46.0	43.6	20 R17
172	81	2.12	2.14	.12	.18	.92	-.5	.96	-.2	1.05	.32	.51	42.5	42.9	13 R10
366	171	2.14	2.17	.05	.13	.87	-1.2	.86	-1.3	1.19	.65	.52	45.3	43.4	1 Ta
177	81	2.19	2.18	.01	.18	1.43	2.5	1.35	2.1	.56	.57	.53	41.9	43.9	21 R18
155	72	2.15	2.22	-.09	.19	.95	-.2	.96	-.2	1.04	.49	.53	43.9	43.3	8 R5
162	72	2.25	2.36	-.47	.20	.86	-.8	.90	-.6	1.08	.34	.51	45.5	42.2	22 R19
191	81	2.36	2.42	-.63	.19	1.62	3.4	1.38	2.2	.51	.67	.51	38.6	42.9	19 R16
179	72	2.49	2.52	-.95	.22	1.18	1.0	1.13	.7	.83	.43	.48	42.6	42.0	6 R3
180	72	2.50	2.59	-1.18	.22	.73	-1.7	.68	-1.9	1.33	.59	.48	46.6	40.9	4 R1
221	81	2.73	2.78	-1.99	.25	1.22	1.1	1.01	.1	.97	.54	.39	34.4	36.7	15 R12
195	72	2.71	2.79	-2.07	.26	.60	-2.3	.50	-2.3	1.36	.66	.41	38.9	36.1	5 R2
186.5	88.4	2.13	2.16	.00	.19	1.01	-.1	.98	-.2		.47		Mean (Count: 22)		
63.7	32.2	.28	.30	.87	.03	.30	1.9	.27	1.7		.16		S.D. (Population)		
65.1	32.9	.28	.31	.89	.04	.30	2.0	.27	1.8		.17		S.D. (Sample)		

Model, Populn: RMSE .19 Adj (True) S.D. .85 Separation 4.50 Strata 6.33 Reliability.95
Model, Sample: RMSE .19 Adj (True) S.D. .87 Separation 4.61 Strata 6.48 Reliability) .96
Model, Fixed (all same) chi-square: 372.6 d.f.: 21 significance (probability): .00
Model, Random (normal) chi-square: 19.8 d.f.: 20 significance (probability): .47
Inter-Rater agreement opportunities: 10341 Exact agreements: 4366 = 42.2% Expected: 4283.6 = 41.4%

Table 10 ***Facets Table 7.2.2 Raters Measurement Report Arranged by Fit***

Rated Performance Test
Table 7.2.2 Raters Measurement Report (arranged by fN).

Total Score	Total Count	Obs Ave	Fair Ave	Model Logits	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	Exact Agree. Obs %	Exact Agree. Exp %	Raters
153	81	1.89	1.84	.86	.17	1.69	3.8	1.67	3.8	.27	.63	.53	41.6	41.0	17 R14
191	81	2.36	2.42	-.63	.19	1.62	3.4	1.38	2.2	.51	.67	.51	38.6	42.9	19 R16
177	81	2.19	2.18	.01	.18	1.43	2.5	1.35	2.1	.56	.57	.53	41.9	43.9	21 R18
221	81	2.73	2.78	-1.99	.25	1.22	1.1	1.01	.1	.97	.54	.39	34.4	36.7	15 R12
179	72	2.49	2.52	-.95	.22	1.18	1.0	1.13	.7	.83	.43	.48	42.6	42.0	6 R3
328	162	2.02	2.05	.35	.13	1.15	1.4	1.16	1.4	.84	.64	.53	42.5	42.7	3 Tc
125	72	1.74	1.75	1.06	.18	1.13	.8	1.10	.6	.89	.64	.53	39.3	38.4	18 R15
125	63	1.98	2.04	.36	.20	1.11	.7	1.13	.7	.83	.52	.53	38.3	42.2	16 R13
135	81	1.67	1.70	1.18	.17	1.08	.5	1.08	.5	.90	.19	.53	40.0	36.8	11 R8
120	63	1.90	1.87	.79	.20	1.05	.3	1.04	.2	.90	.22	.54	38.9	41.2	10 R7
155	72	2.15	2.22	-.09	.19	.95	-.2	.96	-.2	1.04	.49	.53	43.9	43.3	8 R5
172	81	2.12	2.14	.12	.18	.92	-.5	.96	-.2	1.05	.32	.51	42.5	42.9	13 R10
308	171	1.80	1.82	.90	.12	.92	-.8	.93	-.6	1.09	.67	.54	41.3	39.4	2 Tb
366	171	2.14	2.17	.05	.13	.87	-1.2	.86	-1.3	1.19	.65	.52	45.3	43.4	1 Ta
162	72	2.25	2.36	-.47	.20	.86	-.8	.90	-.6	1.08	.34	.51	45.5	42.2	22 R19
160	81	1.98	2.01	.46	.18	.84	-1.0	.84	-1.0	1.16	.38	.51	43.5	41.5	14 R11
165	81	2.04	2.07	.29	.18	.84	-1.1	.86	-.9	1.12	.19	.53	42.9	42.7	9 R6
165	81	2.04	2.05	.35	.18	.76	-1.6	.78	-1.5	1.24	.40	.53	43.9	42.6	12 R9
152	72	2.11	2.11	.19	.19	.75	-1.6	.80	-1.3	1.22	.30	.52	46.0	43.6	20 R17
180	72	2.50	2.59	-1.18	.22	.73	-1.7	.68	-1.9	1.33	.59	.48	46.6	40.9	4 R1
195	72	2.71	2.79	-2.07	.26	.60	-2.3	.50	-2.3	1.36	.66	.41	38.9	36.1	5 R2
170	81	2.10	2.02	.42	.18	.46	-4.5	.48	-4.3	1.61	.34	.52	48.1	43.7	7 R4
186.5	88.4	2.13	2.16	.00	.19	1.01	-.1	.98	-.2		.47		Mean (Count: 22)		
63.7	32.2	.28	.30	.87	.03	.30	1.9	.27	1.7		.16		S.D. (Population)		
65.1	32.9	.28	.31	.89	.04	.30	2.0	.27	1.8		.17		S.D. (Sample)		

Model, Populn: RMSE .19 Adj (True) S.D. .85 Separation 4.50 Strata 6.33 Reliability) .95
Model, Sample: RMSE .19 Adj (True) S.D. .87 Separation 4.61 Strata 6.48 Reliability.96
Model, Fixed (all same) chi-square: 372.6 d.f.: 21 significance (probability): .00
Model, Random (normal) chi-square: 19.8 d.f.: 20 significance (probability): .47
Inter-Rater agreement opportunities: 10341 Exact agreements: 4366 = 42.2% Expected: 4283.6 = 41.4%

Table 11 *Facets Table 7.2.2 Rater Measurement Report for Teacher Ratings*

```
Rated Performance Test
Table 7.2.2  Raters Measurement Report  (arranged by fN).
-------------------------------------------------------------------------------------------------------------
-------
Total  Total   Obs    Fair         Model  Infit      Outfit    Estim. Correlation  Exact Agree.
Score  Count   Ave    Ave  Logits  S.E. MnSq ZStd  MnSq ZStd  Discrm PtMea PtExp  Obs %  Exp % Raters
-------------------------------------------------------------------------------------------------------------
328    162    2.02   1.85   -.09   .14  1.07   .7  1.07   .6   .96    .67   .67   60.5   54.6  3 Tc
308    171    1.80   1.56    .63   .15  1.01   .1  1.01   .1   .96    .72   .72   58.6   53.0  2 Tb
366    171    2.14   2.04   -.54   .14   .88 -1.1   .88 -1.0  1.10    .66   .64   55.6   53.0  1 Ta
-------------------------------------------------------------------------------------------------------------
334.0  168.0  1.99   1.82    .00   .14   .99  -.1   .99  -.1           .68           Mean (Count: 3)
 24.1    4.2   .14    .20    .48   .00   .08   .8   .08   .7           .03           S.D. (Population)
 29.5    5.2   .17    .24    .59   .00   .10  1.0   .10   .9           .03           S.D. (Sample)
-------------------------------------------------------------------------------------------------------------
Model, Populn: RMSE .14  Adj (True) S.D. .46  Separation 3.20  Strata 4.60  Reliability.91
Model, Sample: RMSE .14  Adj (True) S.D. .57  Separation 3.98  Strata 5.64  Reliability.94
Model, Fixed (all same) chi-square:  33.7  d.f.: 2  significance (probability): .00
Model,  Random (normal) chi-square:  1.9  d.f.: 1  significance (probability): .17
Inter-Rater agreement opportunities: 495  Exact agreements: 288 =  58.2%  Expected:  265.1 =  53.5%
-------------------------------------------------------------------------------------------------------------
```

Presentation Teacher Ratings Specification.txt" includes only the teacher's ratings. When we open this in *Facets*, *Facets Table 7.2.2*, reproduced in Table 11, shows a pattern of Rater Ta (live ratings) being lenient and slightly over-fitting, Rater Tb (the first video rating) being strict and well fitting, and Rater Tc (the second video rating) being of average severity and slightly misfitting. Rater behavior has changed between sessions so raw ratings would not be suitable for high-stakes purposes, but the level of misfit is quite small so logit measures provide effective measurement and can be used to diagnose students' use of the rubric.

Table 12 *Facets Table 7.1.2 Student Measurement Report Anchored Against Teacher Ratings*

```
Rated Performance Test
Table 7.1.2  Students Measurement Report  (arranged by fN).
-------------------------------------------------------------------------------------------------------------
Total  Total   Obs    Fair          Model  Infit      Outfit     Estim. Correlation
Score  Count   Ave    Ave  Measure  S.E.  MnSq ZStd  MnSq ZStd  Discrm PtMea  Exp   Nu Students
-------------------------------------------------------------------------------------------------------------
209      90   2.32   2.43    2.32   .19  1.43  2.6  1.29  1.7   .61    .56   .61   14 S11
233     117   1.99   2.06    1.21   .16  1.40  2.8  1.43  2.9   .51    .36   .66   20 S17
244     126   1.94   1.94     .89   .15  1.42  3.1  1.42  3.0   .51    .49   .66    7 S4
190      81   2.35   2.50    2.55   .21  1.38  2.2  1.28  1.6   .61    .46   .59   11 S8
243     117   2.08   2.12    1.39   .16  1.35  2.5  1.31  2.1   .59    .51   .62   15 S12
216     108   2.00   2.20    1.61   .16  1.34  2.4  1.30  2.2   .64    .50   .62    5 S2
217     108   2.01   2.08    1.26   .16  1.32  2.3  1.30  2.2   .63    .50   .62    8 S5
132      72   1.83   1.92     .82   .20  1.31  1.7  1.27  1.5   .66    .55   .68    9 S6
217      99   2.19   2.25    1.76   .18  1.29  1.9  1.19  1.2   .66    .53   .64   16 S13
294     135   2.18   2.26    1.80   .15  1.23  1.8  1.15  1.0   .77    .57   .64   12 S9
203      99   2.05   2.12    1.40   .17  1.18  1.2  1.18  1.2   .79    .51   .65   18 S15
194      90   2.16   2.18    1.56   .19  1.12   .8  1.16   .9   .86    .64   .66    6 S3
156      81   1.93   1.84     .59   .19  1.16  1.0  1.14   .9   .79    .59   .67   21 S18
221      99   2.23   2.19    1.59   .18  1.14  1.0  1.06   .3   .86    .61   .67   10 S7
235     108   2.18   2.33    2.02   .17  1.14  1.0  1.08   .5   .85    .55   .61   19 S16
212      90   2.36   2.48    2.48   .20  1.14   .9  1.12   .8   .83    .41   .58   22 S19
230     108   2.13   2.18    1.57   .17  1.13  1.0  1.10   .7   .85    .62   .65   17 S14
238     117   2.03   2.14    1.45   .16  1.03   .3  1.10   .7   .94    .61   .64    4 S1
220      99   2.22   2.28    1.87   .18   .99   .0  1.07   .4  1.01    .64   .64   13 S10
-------------------------------------------------------------------------------------------------------------
216.0  102.3   2.11   2.18    1.59   .18  1.24  1.6  1.21  1.4           .54    Mean (Count: 19)
 33.3   15.8    .14    .17     .51   .02   .13   .9   .11   .8           .07    S.D. (Population
 34.2   16.2    .15    .18     .53   .02   .13   .9   .11   .8           .08    S.D. (Sample)
-------------------------------------------------------------------------------------------------------------
Model, Populn: RMSE .18  Adj (True) S.D. .48  Separation 2.73  Strata 3.98  Reliability .88
Model, Sample: RMSE .18  Adj (True) S.D. .50  Separation 2.82  Strata 4.09  Reliability .89
Model, Fixed (all same) chi-square:  145.2  d.f.: 18  significance (probability): .00
Model,  Random (normal) chi-square:  16.0  d.f.: 17  significance (probability): .52
-------------------------------------------------------------------------------------------------------------
```

The specification file “04 Presentation Anchored Specification.txt” includes both the student raters and the teacher but the rubric has been “anchored” against the teacher ratings by specifying the difficulty of each rubric item when rated only by the teacher. The ratings made by students will be compared against this and the fit statistics will show which students tended to follow the same rating patterns as the teacher. When we look at the results from this analysis, it is clear that student raters are not behaving comparably to the teacher. *Facets Table 7.1.2*, reproduced in Table 12, shows student ability organized by fit to the model, and all students show some level of misfit. This shows that the ratings by students do not follow the same pattern as the ratings by the teacher.

Facets Table 7.2.2, reproduced in Table 13, shows the rater measurement report arranged by fit-to-the-model and this allows us to identify which student raters are most distorted compared to the teacher. We can see there are four badly misfitting raters with mean-square fit statistics greater than 1.50; R14, R16, R18, and R12. These four raters seem to be causing most of the misfit, so we need to investigate what happened.

Table 13 ***Facets Table 7.2.2 Raters Measurement Report Anchored Against Teacher Ratings***

Rated Performance Test
Table 7.2.2 Raters Measurement Report (arranged by fN).

Total Score	Total Count	Obs Ave	Fair(M) Ave	Logits	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Est. Disrm	Corr PtMea	Corr Exp	Exact Agree Obs%	Exact Agree Exp%	Raters
153	81	1.89	1.85	.97	.18	2.16	5.9	2.09	5.7	-.24	.52	.61	41.6	42.9	17 R14
191	81	2.36	2.42	-.71	.21	2.15	5.6	1.85	4.2	-.03	.54	.59	38.6	44.9	19 R16
177	81	2.19	2.19	.01	.20	2.00	5.2	1.87	4.7	-.11	.42	.60	41.9	46.2	21 R18
221	81	2.73	2.80	-2.25	.27	1.89	3.8	1.61	1.9	.40	.30	.47	34.4	37.9	15 R12
135	81	1.67	1.71	1.34	.18	1.49	2.8	1.48	2.8	.46	.09	.61	40.0	38.3	11 R8
125	72	1.74	1.76	1.20	.19	1.46	2.5	1.40	2.2	.55	.55	.60	39.3	39.9	18 R15
179	72	2.49	2.54	-1.08	.23	1.43	2.2	1.36	1.7	.59	.40	.56	42.6	43.9	6 R3
120	63	1.90	1.87	.90	.21	1.30	1.6	1.29	1.6	.60	.20	.61	38.9	43.0	10 R7
125	63	1.98	2.05	.41	.21	1.20	1.1	1.22	1.3	.72	.56	.60	38.3	44.2	16 R13
162	72	2.25	2.36	-.53	.21	1.19	1.1	1.20	1.2	.70	.25	.59	45.5	44.1	22 R19
172	81	2.12	2.14	.13	.19	1.19	1.2	1.20	1.3	.72	.27	.59	42.5	45.0	13 R10
165	81	2.04	2.08	.32	.19	1.18	1.1	1.20	1.3	.72	.09	.60	42.9	44.8	9 R6
328	162	2.02	2.05	.40	.13	1.18	1.6	1.20	1.8	.78	.67	.61	42.5	44.8	3 Tc
155	72	2.15	2.23	-.11	.21	1.19	1.1	1.18	1.1	.74	.44	.60	43.9	45.4	8 R5
165	81	2.04	2.05	.40	.19	1.10	.6	1.08	.5	.85	.29	.60	43.9	44.6	12 R9
308	171	1.80	1.83	1.02	.13	1.02	.2	1.05	.5	.95	.66	.61	41.3	41.2	2 Tb
160	81	1.98	2.01	.51	.19	.97	-.1	.98	.0	1.01	.40	.59	43.5	43.4	14 R11
366	171	2.14	2.17	.06	.13	.91	-.8	.91	-.8	1.12	.68	.60	45.3	45.5	1 Ta
152	72	2.11	2.12	.21	.20	.87	-.7	.93	-.4	1.09	.34	.60	46.0	45.7	20 R17
180	72	2.50	2.61	-1.34	.24	.81	-1.0	.77	-1.2	1.21	.61	.56	46.6	42.6	4 R1
170	81	2.10	2.02	.48	.19	.61	-2.9	.64	-2.8	1.43	.32	.60	48.1	45.9	7 R4
195	72	2.71	2.82	-2.34	.28	.69	-1.7	.55	-1.7	1.29	.64	.49	38.9	37.2	5 R2
186.5	88.4	2.13	2.17	.00	.20	1.27	1.4	1.23	1.2		.42		Mean (Count: 22)		
63.7	32.2	.28	.31	.99	.04	.43	2.3	.38	2.0		.18		S.D. (Population)		
65.1	32.9	.28	.31	1.01	.04	.44	2.3	.39	2.0		.18		S.D. (Sample)		

Model, Populn: RMSE .20 Adj (True) S.D. .97 Separation 4.81 Strata 6.75 Reliability.96
Model, Sample: RMSE .20 Adj (True) S.D. .99 Separation 4.93 Strata 6.90 Reliability.96
Model, Fixed (all same) chi-square: 423.9 d.f.: 21 significance (probability): .00
Model, Random (normal) chi-square: 19.9 d.f.: 20 significance (probability): .46
Inter-Rater agreement opportunities: 10341 Exact agreements: 4366 = 42.2% Expected: 4479.6 = 43.3%

At the very bottom of the output file, we find *Facets Table 4.1*, reproduced in Table 14, which shows the most unexpected responses. This table is generated by comparing the observed score with the statistically expected score. The difference between these is called the score residual. The score residuals are then standardized and *Facets* reports standardized residuals with absolute values greater than 3.0, which roughly corresponds to statistical significance of $p < .01$. Looking at Table 14, it is obvious that raters R12, R14, R16, and R18 are causing most of the misfit problems and that Item 3 "Eyes" is the most problematic item for these raters. This very fine grained diagnostic information about how students, items, and raters interact shows us where problems are arising and lets us identify individual students, raters, and items for remedial attention.

Summary and Conclusions

This paper introduced Rasch analysis through a practical process of working through two examples. The first example was based on a type of test typically given in educational settings. This was a dichotomously scored, that is, correct or incorrect answers to multiple choice items. The second example was a judged

Table 14 ***Facets Table 4.1 Most Misfitting Responses***

Rated Performance Test
Table 4.1 Unexpected Responses (19 residuals sorted by u).

Cat	Score	Exp.	Resd	StRes	Nu	Stu	Nu	Rat	N	Items
2	2	2.9	-.9	-4.2	13	S10	15	R12	2	Notes/reading
0	0	2.3	-2.3	-4.0	14	S11	19	R16	3	Eyes
1	1	2.7	-1.7	-3.8	20	S17	6	R3	2	Notes/reading
2	2	2.9	-.9	-3.6	6	S3	15	R12	2	Notes/reading
0	0	2.1	-2.1	-3.6	11	S8	21	R18	3	Eyes
0	0	2.1	-2.1	-3.6	12	S9	19	R16	3	Eyes
0	0	2.1	-2.1	-3.6	18	S15	17	R14	5	Body
0	0	2.1	-2.1	-3.5	5	S2	19	R16	3	Eyes
0	0	2.1	-2.1	-3.5	14	S11	21	R18	3	Eyes
2	2	2.9	-.9	-3.4	4	S1	15	R12	2	Notes/reading
1	1	2.7	-1.7	-3.4	17	S14	6	R3	1	Confidence
0	0	2.0	-2.0	-3.3	15	S12	19	R16	3	Eyes
0	0	1.9	-1.9	-3.2	8	S5	19	R16	3	Eyes
0	0	1.9	-1.9	-3.2	9	S6	2	Tb	5	Body
1	1	2.6	-1.6	-3.2	12	S9	15	R12	3	Eyes
1	1	2.6	-1.6	-3.2	13	S10	15	R12	3	Eyes
1	1	2.6	-1.6	-3.2	22	S19	17	R14	9	Intonation
3	3	1.0	2.0	3.1	8	S5	17	R14	8	Pausing
0	0	1.9	-1.9	-3.1	18	S15	11	R8	6	Speed
Cat	Score	Exp.	Resd	StRes	Nu	Stu	Nu	Rat	N	Items

performance test typically given when raters use a scoring rubric that is also common in educational settings. The examples were worked through with two different software programs *Winsteps* and *Facets*. Practical guidelines and rules-of-thumb were provided so that researchers doing their own analysis with their own data would be able to follow the steps and be able to evaluate their own measures.

References

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model* (2nd ed.). London: Lawrence Erlbaum.

Brown, J. D. (2005). *Testing in language programs: A comprehensive guide to English language assessment (New Ed.)*. New York: McGraw-Hill.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1*(1), 1-47.

Ellis, R. (1994). *The study of second language acquisition*. Oxford: Oxford University Press.

Embretson, S. E., & Hershberger, S. L. (Eds.). (1999). *The new rules of measurement: What every psychologist and educator should know*. Mahwah, NJ: Lawrence Erlbaum.

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

Engelhard, G. (2008). *Historical view of the influences of theories of measurement and language proficiciency on English language tests*. Paper presented at the PROMS 2008: Pacific Rim Objective Measurement Symposium, Ochanomizu University, Tokyo, Japan.

Engelhard, G. (2013). *Invariant measurement*. New York: Routledge.

Field, A. P. (2009). *Discovering statistics with SPSS* (3rd ed.). London: Sage.

Hymes, D. (1972). On communicative competence. In J. B. Pride & J. Holmes (Eds.), *Sociolinguistics: Selected readings* (pp. 269-293). Harmondsworth: Penguin.

Linacre, J. M. (1994). *Many-facet Rasch measurement* (2nd ed.). Chicago: MESA Press.

Linacre, J. M. (2010a). *Facets* (Version 3.67.0). Retrieved from http://www.winsteps.com/facets.htm

Linacre, J. M. (2010b). Winsteps (Version 3.70.02). Retrieved from http://www.winsteps.com

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

McNamara, T. F., & Knoch, U. (2012). The Rasch wars: The emergence of Rasch measurement in language testing. *Language Testing, 29*(4), 555-576. doi: 10.1177/0265532211430367