

**TWO STRATEGIES FOR FITTING REAL DATA
TO RASCH POLYTOMOUS MODELS**

Antonio J. Rojas Tejada*

Andrés González Gómez**

José L. Padilla García**

Cristino Pérez Meléndez**

UNIVERSITY OF ALMERÍA*

UNIVERSITY OF GRANADA**

For correspondence, please, contact with:

Prof. Dr. ANTONIO J. ROJAS TEJADA

Área de Metodología de las Ciencias del Comportamiento

University of Almería

04120 ALMERÍA

SPAIN

Email: arojas@ual.es

TWO STRATEGIES FOR FITTING REAL DATA TO RASCH POLYTOMOUS MODELS

ABSTRACT

A comparative study of the results provided by two strategies for fitting data to Latent Trait Theory Models has been performed. The first, called Total-Persons-Items (TPI), is structured in three phases: 1) assessment of item fit, 2) assessment of person fit; and finally, 3) overall fit of data to the models (items and persons). The second strategy, the Total-Items-Persons (TIP), changes the order of the phases: 1) assessment of person fit, 2) assessment of item fit and, 3) overall fit of data to the models. To verify the results of these two strategies, a set of 30 items, designed to measure religious attitude, was administered to a sample of 821 persons. The Latent Trait Theory Models used were the Partial Credit Model and the Rating Scale Model. The results underline an important difference between the two procedures: the TPI maximizes the number of persons with good fit and the TIP maximizes the number of items with good fit. Moreover, a procedure for controlling the sensitivity of fit to sample size is proposed.

INDEX TERMS: strategies for fitting data to model, sensitivity of fit to sample size, Rasch-fit statistics, Partial Credit Model, Rating Scale Model

TWO STRATEGIES FOR FITTING REAL DATA TO RASCH POLYTOMOUS MODELS

Study of data-to-model fit is obligatory in the application of Latent Trait Models. If the data do not fit the model, then they cannot be used either to calibrate the items or to measure the persons (Wright, 1980). All the advantages and opportunities of the Latent Trait Theory for the construction of tests are obtained only when the data fit the model satisfactorily (i.e. Hambleton, 1990; Hambleton and Swaminathan, 1985; Wright and Stone, 1979).

All Rasch-fit statistics are based on the similarity between empirical and expected data (Gustafsson, 1980). Good fit allows ill-functioning items, suspicious persons, surprising item-person combinations, as well as responses that fit too well to be identified.

The study of fit can be considered from a triple perspective: 1) item fit (which helps identify poor items); 2) person fit (which helps identify a suspicious pattern of responses); and, 3) overall fit (which enables fit of data to the model to be evaluated as a whole). Overall data-to-model fit is the last of the three ways of analyzing the adjustment. However, there is no reason why either one of the two remaining perspectives must be carried out first. This represents a problem in psychometric work based on Latent Trait Models. The professional that wishes to employ these models has no sequence of clearly-defined steps available to him.

In this study, two possibilities for analyzing fit are proposed:

1.- Total-Person-Item Strategy (TPI). Consisting of: 1) evaluating item fit, 2) eliminating badly-fitting items, 3) then recalibrating, and only when items with good fit are obtained, 4) evaluating person fit, and finally, 5) evaluating overall fit of data to the model.

2.- Total-Item-Person Strategy (TIP). Consisting of: 1) evaluating person fit, 2) eliminating the badly-fitting persons, 3) then recalibrating, and only when persons which fit well are obtained, 4) evaluating item fit, and finally, 5) evaluating the overall fit of data to the model.

It is generally believed that one of the reasons for bad fit of persons is that several items used to estimate their ability were 'bad-quality' items. If these items are eliminated, it is then

reasonable to expect better estimation of person ability and, therefore, that the TPI strategy will maximize the number of persons with adequate fit. On the contrary, one of the reasons for detecting items with poor fit is that the persons show an inappropriate response pattern (Fed Li and Olejnik, 1997). Thus, if the TIP strategy is used and persons with poor fit are eliminated from the analysis, it may be expected that the number of items with good fit will increase.

The strategy used to study data-model fit can therefore lead to different results depending on the order in which the same actions are performed.

Moreover, the study of fit has a shortcoming: sensitivity of fit statistics to the sample size. This problem is well described by Hambleton (1989), Hambleton and Murray (1983) and López Pina and Hidalgo (1996). Hambleton (1989) and Hambleton and Murray (1983) show how, for a three-parameter logistic model with simulated data (and, therefore, with perfect theoretical adjustment), by varying the person sample size to determine how good item fit is, differences in the percentage of items with bad fit are obtained. Thus with large person samples ($N=2400$), for a total of 50 items, percentages of items with bad fit vary between 76% and 84%, while with small samples ($N=150$) these percentages oscillate between 10% and 40%. Similar results were obtained by López Pina and Hidalgo (1996) for the two-parameter logistic model (also with simulated data) and 40 items. With small samples ($N=50$), 15% of the items fit badly, while in large samples ($N=2000$) this percentage increased to 67.5%. In fact, with large sample sizes it is very easy for small model-data discrepancies to become statistically significant, leading to the conclusion that the items do not fit the model; however, with small sample sizes the risk of erroneous estimates is very high. To partially solve this problem, a procedure, applied by Rojas, González, Padilla y Pérez (2000), which is consistent when different sample sizes are used for different analyses of fit, was followed.

Two models belonging to the family of Rasch Polytomous Models (Rasch, 1960) were used to study how this strategy of fit works, specifically, the Rating Scale Model -RSM- (Andrich, 1978a, 1978c; Masters, 1980; Wright and Masters, 1982) and the Partial Credit Model -PCM- (Masters, 1982, 1988a, 1988b; Masters and Wright, 1984, 1997; Wright and Masters, 1982). Several tests of fit have been developed for these models. Masters and Wright (1997) classified them into the three types mentioned above: 1) tests for item fit, where items that have 'problems' with the model are identified, 2) tests of person fit, where persons that do

not follow the general standard answers proposed by the model are indicated; and, 3) tests of total fit, that indicate the degree of general adjustment of the entire set of data to the model (items and persons).

The object of this work is an empirical study that corroborates the predicted functioning of the different strategies of fit. A complementary objective is to show the utility of the procedure used to control the problem of sensitivity to sample size.

Method

Persons

The study sample consisted of 821 students. Of these, 22.61% were men and 77.4% were women; the mean age was 21.30 (median=21 and moda=20), with a range of 28 (max.=46 and min=18), and standard deviation of 3.39; all of them were students at the Universities of Almería (80.4%) or Sevilla (19.6%); mostly in Psychology (59.7%), followed by Psychopedagogy (8.8%), Teaching (7.3%), Environmental Sciences (7.1%), Pedagogy (6.8%) and others (10.3%).

Materials

A set of 30 items designed to measure religious attitudes as employed by Morales (1988) in his 'Scale of Religious Attitudes (R-1)' to measure 'religious beliefs or nearness to the faith' (Morales, 1988, p. 488). The definition of the trait is very generic, and does not intend to allude to any specific religion or religious practice, such as conceived by Hood (1970) or Pargament, et al. (1988). The items are expressed on a 5-point Likert scale. This particular test was selected for two fundamental reasons: 1) it allows RSM and PCM control of polytomous items with a classification scale format that collects the religious attitude in only one dimension and, 2) it is open-ended. Furthermore, as it was designed for didactic purposes, Morales (1988) provides three scales, from a first version with 30 items to a third or final test consisting of 18 items, that describe the process of classical item analysis.

Analysis

Two statistical methods have been developed that provide the degree of fit to the RSM and PCM (for items as well as persons) models. These are called: The 'Outfit' or Standardized Outlier-Sensitive Mean Square Fit Statistic, and 'Infit' or Standardized Information-Weighted Mean Square Fit Statistic (i.e. Wright, 1999; Wright and Linacre, 1998a).

The criterion to be considered in determining whether or not an item or a person is a misfit is whether 'infit' or 'outfit' values obtained are equal to or greater than 2 (or a negative item-total or person-total correlation). To prove that data (items or persons) fit to the model, the distribution of 'infit' and 'outfit' values is examined in two ways: a) by calculating mean values for 'infit' and 'outfit' and standard deviations for items and for persons, keeping in mind that when fit to the model is good, the means should be nearly 0 and standard deviations near 1; and, b) observing whether the distribution of the 'infit' and 'outfit' statistic values is approximately normal.

The BIGSTEPS program, ver. 2.82, developed by Wright and Linacre (1998b) was used to apply the RSM and PCM. The PROX and UCON estimation procedures as described by Wright and Masters (1982) were used.

Procedure

The questionnaire was collectively administered by a researcher in the classroom where the students attended their classes. All of them had to answer all the questions. Data was collected at the facilities of the Universities of Almería and Sevilla.

Results

Strategy TPI

Stage I: Fit and selection of items with different sample sizes.

This stage is designed to determine how many items in a set of 30 adjust to the RSM and PCM models. Items with good fit are selected as 'good' items and used to measure the variable proposed. But evaluation of goodness of fit to the model presents an important problem: sensitivity of the statistics to sample size. This is especially important for item fit, since it is in this phase when it is decided which items fit and are therefore selected to measure the variable, and which items are to be discarded as misfits. Because of this, an attempt has been made to control sensitivity to sample size at this stage. The total sample was divided into four samples of different sizes. The first and largest, is made up of the total sample ($n=821$). The other three are random samples containing 75%, 50% and 25% of the persons.

Results with the RSM indicate that 18 items are misfits with the 100%-size sample according to the 'infit' and 'outfit' values; the 75%-size sample has only 8 misfits; the 50% sample has 16 misfits and only 10 items are considered misfits with the 25%-size sample. It is thus clear how sample size is influencing the 'infit' and 'outfit' statistics. All the items show positive Item-Total Correlations (ITC), except Item 4 in the 25% sample, which has a correlation of -0.04. Items 1, 4, 11, 13, 16, 20, 21, 24, 28 and 30 are considered misfits in all four sample sizes. Items 2, 3, 5, 10, 14, 19 and 27 show poor fit in the 50%, 75% and 100% samples. Item 22 only appears as a misfit in the 75% and 100% samples.

Results with the PCM indicate that 16 items are misfits with the 100% sample size; with the 75% sample, 15 items are misfits; 14 items are misfits with the 50% sample and only 7 items are misfits with a 25% sample. All the items had a positive ITC, except Item 4 in the 25% sample, which has a correlation of -0.04. Items 1, 3, 4, 11, 21, 24 and 28 are considered misfits in all four samples sizes. Items 2, 12, 23, 25, 27 and 30 show poor fit for the 50%, 75% and 100% samples sizes. Items 10 and 26 only appear as misfits in the 75% and 100% sample sizes, while items 19 and 20 are only misfits with the 50% and 100% sample sizes, respectively.

This must be the determining factor in a decision as to which items should be rejected or accepted based on their fit to the model. In this case, items are considered misfits when they have shown poor fit in three or more samples sizes. According to this criterion, the 17 following items are misfits with the RSM: 1, 2, 3, 4, 5, 10, 11, 13, 14, 16, 19, 20, 21, 24, 27, 28 and 30. For the PCM, the following 13 items are considered misfits: 1, 2, 3, 4, 11, 12, 21, 23, 24, 25, 27, 28 and 30.

Stage II: Fit and selection of persons.

Once items with good fit, that is, the 13 for the RSM and 17 for the PCM, have been selected, person fit to both models may be analyzed with these items.

For the RSM, of 821 persons, 193 show poor fit (either because of inappropriate 'infit' and 'outfit' values, or negative Person-Total Correlation -STC-). This is 23.5% of the total sample.

For the PCM, 188 persons have poor fit (22.9% of the total sample).

These persons have response patterns which are very different from those expected from the model and will be eliminated from the sample (as was done with the items).

Stage III: Total fit of selected items with the selected persons.

Total fit of items

Results with the RSM show evidence that item fit is acceptable (Table 1 and Figure 1). Only one item (8) obtained an excessive 'outfit' value (2.07). On the other hand, the means of the 'infit' and 'outfit' statistics are relatively near to 0 (0.16 and 0.22, respectively), with standard deviations of 1.21 and 1.28, respectively.

.....
Table 1
.....
.....

.....
Figure 1
.....

For the PCM, the results also indicate evidence leading to the conclusion that item fit is acceptable (Table 2 and Figure 2). Only Item 19 showed excessive 'infit' and 'outfit' values (2.31 and 2.72, respectively). On the other hand, the statistical means of 'infit' and 'outfit' are

nearly 0 (0.10 and 0.11, respectively) with standard deviations of 1.20 and 1.29, respectively.

.....
Table 2
.....

.....
Figure 2
.....

Total fit of persons

For the RSM, results show evidence leading to the conclusion that person fit is adequate (Figure 3). The means of the 'infit' and 'outfit' statistics are nearly 0 (0.02 and 0.03, respectively) and standard deviations are nearly 1 (1.01 and 0.96, respectively).

.....
Figure 3
.....

In the case of the PCM, the results also show evidence of adequate fit (Figure 4). The means of the 'infit' and 'outfit' statistics are nearly 0 (0.02 and 0.01, respectively) and standard deviations are nearly 1 (1.00 and 0.95, respectively).

.....
Figure 4
.....

The results obtained with the persons, together with the previous results for the items, support the interpretation of an adequate total fit of persons and items.

Strategy TIP

Stage I: Adjustment and selection of the persons.

In this first stage, the process is carried out for both models (30 items with a sample of 821 persons). Person fit is examined in order to discard those persons with response patterns that differ from those expected from the model.

For the RSM, of 821 persons, 249 are observed to show incoherent response patterns (from high 'infit' and 'outfit' values or negative STC). These results mean 30.33% of the sample.

For the PCM, 243 persons show response patterns different from those patterns expected from the model. That is 28.5% of the sample.

These badly fitting persons will be eliminated from the sample for the second stage.

Stage II: Fit and selection of the items with different samples sizes.

In this stage, just as in the TPI strategy, control of the sensitivity of fit has been attempted using 572 persons from Stage I. The new 100% sample has 572 persons, and the three random 75%, 50% and 25% samples have 429, 286 and 143 persons, respectively.

In this case, for the RSM 19 badly fitting items are observed with the 100% sample size; with the 75% sample 18 items fit badly; with the 50% sample, there are 16 badly fitting items and 12 items are considered misfits with a sample size of 25%. All the items have a positive ITC, except item 21 in the 25% sample size, which has a value of -0.02. Items 1, 3, 4, 11, 13, 14, 16, 19, 20, 21, 27 and 28 are considered misfits with all four samples. Items 10, 22, 24 and 30 are misfits with the 50%, 75% and 100% samples. Only items 5 and 25 appear to be misfits with the 75% and 100% samples.

For the PCM, 16 items are considered misfits for the 100% sample size; with the 75% sample, there are 12 misfitting items; with the 50% sample there are 10, and 10 with the 25% sample size also. All the items have a positive ITC, except for Item 4 in the 25% sample (value of -0.04) and the Item 3 in the 50% sample (value of -0.01). Items 1, 3, 4, 11, 21, 24, 25, 27 and 28 do not fit in any of the four samples. Item 2 does not fit with the 50%, 75% and 100% samples sizes. Items 10, 20 and 30 appear to be misfits with two sample sizes. Items 12, 23 and 29 are only misfits with 100% sample size.

To offset the sensitivity of fit to sample size, as with the TPI strategy, items that did not fit with three or more samples sizes are considered bad fits. Thus, for the RSM, the following 16 items are considered misfits: 1, 3, 4, 10, 11, 13, 14, 16, 19, 20, 21, 22, 24, 27, 28 and 30. For the PCM, the following 10 items are considered misfits: 1, 2, 3, 4, 11, 21, 24, 25, 27 and 28.

Stage III: Total fit of selected items with the elected persons

Total fit of items

For the RSM, the overall fit of the items was good in all cases, except for 2 and 25 (for values of 'infit' and 'outfit'). All the ITC are positive and oscillate between the minimum value of 0.34 and maximum of 0.59 (Table 3 and Figure 5).

.....

Table 3

.....

.....

Figure 5

.....

When the distribution of the statistics of fit (mean and standard deviation) is studied, the following is observed: the means of the 'infit' and 'outfit' statistics are near 0 (0.11 and 0.08, respectively) with standard desviations of 1.38 and 1.30, respectively

For the PCM, fit was appropriate in all cases, except for 2 (for 'infit') and 20 (for 'infit' and 'outfit' values). All the ITC are positive and oscillate between the minimum value of 0.31 and the maximum of 0.50 (Table 4 and Figure 6).

.....

Table 4

.....

.....

Figure 6

.....

From study of the distribution of the fit statistics, it is observed that the means of the 'infit' and 'outfit' statistics are nearly 0 (0.09 and 0.07, respectively), with standard deviations of 1.22.

Total fit of persons

For the RSM, the mean and standard deviation were -0.01 and 1.09 for 'infit', and 0.01 and 1.07 for 'outfit' (Figure 7).

.....

Figure 7

.....

For the PCM, the mean and standard deviation were 0.00 and 0.01 for 'infit' and 1.09 and 1.05 for 'outfit' (Figure 8).

.....

Figure 8

.....

These results also support the interpretation of good total fit with this strategy.

Conclusions

In this paper, two strategies for fitting data to the RSM and PCM are described. When the two strategies are compared, differences are observed with regard to number of items and persons that are selected for their good fit, although both approaches obtain adequate fit of data to the models. As the two strategies employ equivalent operations and only differ in the order of the first two stages, it may be said that depending on the strategy followed, either the number

of items that fit well or the number of persons that fit well is maximized (Table 5). It is thus demonstrated that with the TPI strategy, the number of persons selected is maximized by showing response patterns coherent with the model (76.5% for the RSM and 87.1% for the PCM) compared to the TIP strategy (69.7% for the RSM and 71.5% for the PCM). On the contrary, and as was noted at the beginning, the TIP strategy maximizes the number of items selected by showing adequate fit to the model (46% for the RSM and 66.7% for the PCM) as compared to the TPI strategy (43.3% for the RSM and 56.7% for the PCM).

.....
 Table 5

These results are very important, since their implications depend on the objectives for which item calibration and person measurement are going to be used. When a set of items is designed to measure a psychological variable, and the major interest is in obtaining the greatest number of items that measure said variable, for example, to create an item bank, and not in measuring the persons, the TIP strategy would be used. If the purpose of the items is not so much inclusion in an item bank as measuring the persons with regard to the psychological variable they are intended to measure, the strategy that would obtain the most person measurements coherent with the model would be the TPI.

It should be pointed out that bad fit of either items or persons should cause reflection on their possible causes, as already indicated by O'Brien (1992) for items: problems with the psychological theory on which they are based or the process by which the theory is been made operative through the items, or perhaps both (Bohlig, Fisher, Masters and Bond, 1998). An atypical response pattern may indicate (Fred Li and Olejnik, 1997): 1) examinee misconceptions (Tatsuoka, 1984, 1985); 2) exceptional creativity (Levine and Drasgow, 1982); 3) cultural differences (Van Der Flier, 1982); 4) cheating (Mandsen, 1987); 5) instructional differences (i.e. Padilla, Pérez and González, 1998; Harnish, 1983); 6) test bias (Frary, 1982); 7) social desirability (Schmitt, Cortina and Whitney, 1993); or, 8) traitedness/response inconsistency (Reise and Waller, 1993).

The new development presented in this study is a way of controlling sensitivity of the fit statistic to sample size by employing the two strategies followed for data fit. This effect has been controlled by dividing the total sample into four random samples, in which fit to the items is calculated separately, the criterion for item selection being adequate fit in at least two samples.

It should also be noted that the results obtained with the RSM and PCM may be extended to application with any Latent Trait Theory model, and therefore, should be investigated using simulated data.

References

- Andrich, D. (1978a). A Rating Formulation for Ordered Response Categories. *Psychometrika*, 43. 561-573.
- Andrich, D. (1978b). Application of a Psychometric Rating Model to Ordered Categories which are Scored with Successive Integers. *Applied Psychological Measurement*, 2(4). 581-594.
- Bohlig, M., Fisher, W.P., Masters G.N., and Bond, T. (1998). Content Validity and Misfitting Items. *Rasch Measurement Transactions* 12:1, 607.
- Frary, R.B. (1982). A Comparison of Person Fit Measures. Paper presented at the *Annual Meeting of the American Educational Research Association*. New York.
- Fred Li, M. and Olejnik, S. (1997). The Power of Rasch Person-fit Statistics in Detecting Unusual Response Patterns. *Applied Psychological Measurement*, 21(3). 215-231.
- Gorsuch, R.L. (1988). Psychology Of Religion. *Annual Review of Psychology*, 39. 201-221.
- Gustafsson, J. (1980). Testing and Obtaining Fit of Data to the Rasch Model. *British Journal of Mathematical and Statistical Psychology*, 33. 205-233.
- Hambleton, R.K. (1989). Principles and Selected Applications of Item Response Theory. In R.L. Linn (Ed.). *Educational measurement*. (pp. 147-200). New York: McMillan.
- Hambleton, R.K. (1990). Item Response Theory: Introduction and Bibliography. *Psicothema*, 2(1). 97-107.

- Hambleton, R.K. and Murray, L. (1983). Some Goodness of Fit Investigations for Item Response Models. In R.K. Hambleton (Ed.). *Applications of item response theory*. (pp.71-94). Vancouver, B.C.: Educational Research Institute of British Columbia.
- Hambleton, R.K. and Swaminathan, H. (1985). *Item Response Theory: Principles and Applications*. Boston: Kluwer Academic Publishers.
- Harnish, D.L. (1983). Item Response Patterns: Applications for Educational Practice. *Journal of Educational Measurement*, 20. 191-206.
- Hojtink, H., Molenaar, I. and Post, W. (1994). *PARELLA. User's Manual*. IEC ProGAMMA. Groningen- The Netherlands.
- Hood, R.W. Jr. (1970). Religious Orientation and the Reported Religious Experience. *Journal for the Scientific Study of Religion*, 9(4). 285-291.
- Levine, M.V. and Drasgow, F. (1982) Appropriateness Measurement: Review, Critique and Validating Studies. *British Journal of Mathematical and Statistical Psychology*, 35. 42-56.
- López Pina, J.A. e Hidalgo, M.D. (1996). Bondad de ajuste y teoría de respuesta a los ítems. In J.Muñiz (Coor.). *Psicometría*. (pp. 643-703). Madrid: Universitas.
- Madsen, H.S. (1987). *Utilizing Rasch Analysis to Detect Cheating on Language Examinations*. ERIC Document Reproduction Service N° ED 287 284.
- Masters, G.N. (1980). A Rasch Model for Rating Scales. *Dissertation Abstracts International*, 41, 215A-216A.
- Masters, G. N. (1982). A Rasch Model for Partial Credit Scoring. *Psychometrika*, 47(2). 149-174.
- Masters, G.N. (1988a). The Analysis of Partial Credit Scoring. *Applied Measurement in Education*, 1(4). 279-297.
- Masters, G.N. (1988b). Partial Credit Model. In J.P.Keeves, (Ed.). *Educational Research, Methodology and Measurement: an International Handbook*. (pp. 292-297). Elmsford, NY.: Pergamon Press.
- Masters, G.N. and Wright, B.D. (1984). The Essential Process in a Family of Measurement Models. *Psychometrika*, 49(4). 529-544.

- Masters, G.N. and Wright, B.D. (1997). The Partial Credit Model. In W. J. van der Linden and R.K. Hambleton (Eds.). *Handbook of Modern Item Response Theory*. (pp. 101-121). New York: Springer-Verlag.
- Morales, P. (1988). *Medición de actitudes en psicología y educación*. San Sebastián: Ttattalo.
- O'Brien, M.L. (1992). Using Rasch Procedures to Understand Psychometric Structure in Measures of Personality. In M. Wilson (Ed.). *Objective Measurement: Theory into Practice*. (pp. 61-76). Norwood, NJ: Ablex Publishing Corporation.
- Padilla, J.L.; Pérez, C. and Gómez. A. (1998). La explicación del sesgo en los items de rendimiento. *Psicothema*, 10(2). 481-490.
- Pargament, K.I., Kennell, J., Hathaway, W., Grevengoed, N., Newman, J. and Jones, W. (1988). Religion and the Problem-solving Process: Three Styles of Coping. *Journal for the Scientific Study of Religion*, 27(1), 90-104.
- Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen: Danish Institute for Educational Research. (Reprinted by the Chicago University Press, 1980).
- Reise, S.P. and Waller, N.G. (1993). Traitness and the Assessment of Response Pattern Scalability. *Journal of Personality and Social Psychology*, 65. 143-151.
- Rojas, A.J., González, A., Padilla, J.L. and Pérez (2000). Comparison of Strategies of the Data Fit to the Partial Credit Model. *Psicothema*, 12, n° 2, 296-302. [In spanish].
- Schmitt, N. Cortina, J.M. and Whitney, D.J. (1993). Appropriateness Fit and Criterion-related Validity. *Applied Psychological Measurement*, 17. 143-150.
- Tatsuoka, K.K. (1984). Caution Indices Based on Item Response Theory. *Psychometrika*, 49. 95-110.
- Tatsuoka, K.K. (1985). A Probabilistic Model for Diagnosing Misconceptions by the Pattern Classification Approach. *Journal of Educational Statistics*, 10. 55-73.
- Van Der Flier, H. (1982). Deviant Response Patterns and Comparability of Tests Scores. *Journal of Cross-Cultural Psychology*, 13. 267-298.
- Wright, B.D. (1980). Afterword. In Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. (pp. ix-xxiii). Chicago: The University of Chicago

Press.

Wright, B.D. (1999). Fundamental Measurement for Psychology. In S. E. Embetson and S.L. Hershberger (Ed.). *The New Rules of Measurement*. (pp. 65-104). Mahwah, NJ: Lawrence Erlbaum Associates, Pub.

Wright, B.D. and Linacre J.M. (1998a). *A User's Guide to BIGSTEPS*. Chicago. MESA Press.

Wright, B.D. and Linacre, J.M. (1998b). *BIGSTEPS ver. 2.2. Computer Program*. Chicago: MESA Press.

Wright, B.D. and Masters, G.N. (1982). *Rating Scale Analysis*. Chicago. MESA Press.

Wright, B.D. and Stone, M.H. (1979). *Best Test Design: Rasch Measurement*. Chicago. MESA Press.

Nº Item	Infit	Outfit	ITC
6	1.09	1.29	0.37
7	0.67	0.76	0.34
8	1.80	2.07	0.39
9	0.84	0.75	0.40
12	-1.94	-1.53	0.51
15	1.17	1.77	0.40
17	-1.10	-1.54	0.36
18	1.18	1.26	0.42
22	0.85	0.45	0.41
23	-0.41	-0.44	0.50
25	-1.84	-1.80	0.57
26	0.17	-0.26	0.45
29	-0.41	0.03	0.43
Mean	0.16	0.22	0.43
S. Desv.	1.21	1.28	0.07

Table 1. RSM. Total fit: TPI (13 items and 627 persons).

Nº Item	Infit	Outfit	ITC
5	0.38	0.44	0.39
6	1.57	1.94	0.35
7	1.04	0.93	0.36
8	0.64	0.64	0.39
9	0.46	0.63	0.38
10	-1.41	-1.90	0.47
13	-1.38	-1.23	0.43
14	-1.04	-1.13	0.42
15	-0.90	-0.73	0.44
16	1.26	1.41	0.29
17	1.93	1.45	0.29
18	-0.40	-0.39	0.42
19	2.31	2.72	0.35
20	-0.93	-0.84	0.43
22	-0.13	-0.58	0.40
26	-1.01	-1.22	0.42
29	-0.71	-0.24	0.42
Mean	0.10	0.11	0.39
S. Desv.	1.20	1.29	0.05

Table 2. PCM. Total fit: TPI (17 items and 633 persons).

Nº Item	Infit	Outfit	ITC
2	-2.14	-2.02	0.52
5	-0.20	0.00	0.39
6	1.82	1.72	0.35
7	0.53	0.49	0.34
8	0.86	0.73	0.41
9	1.33	1.05	0.39
12	-1.68	-1.30	0.52
15	1.35	1.42	0.38
17	-0.11	0.00	0.34
18	1.58	1.46	0.40
23	-0.49	-0.72	0.52
25	-2.11	-2.16	0.59
26	1.38	1.22	0.41
29	-0.54	-0.72	0.45
Mean	0.11	0.08	0.43
S. Desv.	1.38	1.30	0.08

Table 3. RSM. Total fit: TIP (14 items and 572 persons).

Nº Item	Infit	Outfit	ITC
5	-0.04	-0.02	0.42
6	2.02	1.85	0.35
7	1.47	1.16	0.36
8	0.89	0.89	0.40
9	1.10	1.09	0.38
10	-0.81	-1.27	0.47
12	-1.37	-1.24	0.48
13	-0.41	-0.27	0.39
14	-0.63	-0.68	0.42
15	0.00	0.25	0.43
16	0.93	1.05	0.31
17	1.35	1.39	0.32
18	0.31	0.09	0.42
19	1.98	1.99	0.39
20	-1.30	-1.27	0.46
22	1.14	1.32	0.37
23	-1.38	-1.21	0.49
26	-0.38	-0.55	0.43
29	-0.93	-1.08	0.46
30	-2.07	-2.11	0.50
Mean	0.09	0.07	0.41
S. Desv.	1.22	1.22	0.05

Table 4. PCM. Total fit: TIP (20 items and 578 persons).

	TPI strategy		TIP strategy	
	RSM	PCM	RSM	PCM
Persons	76.5% (628)	87.1% (633)	69.7% (572)	71.5% (578)
Items	43.3% (13)	56.7% (17)	46.0% (14)	66.7% (20)

Table 5. Fitted items and persons by model and strategy.

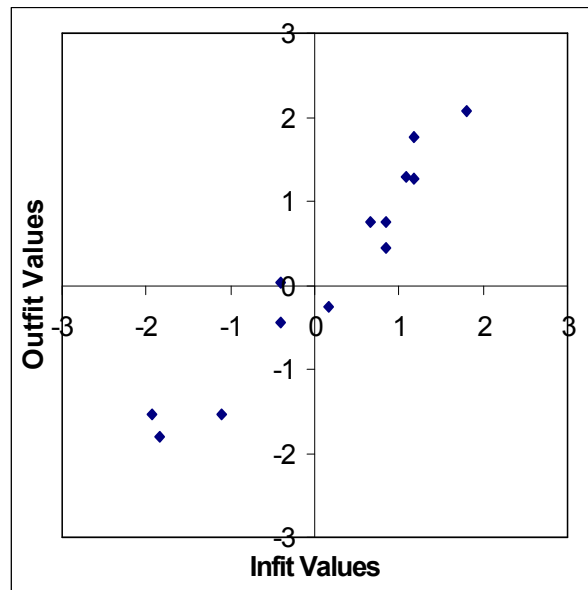


Figure 1. Scatterplot RSM. Total fit of items: TPI (13 items and 627 persons).

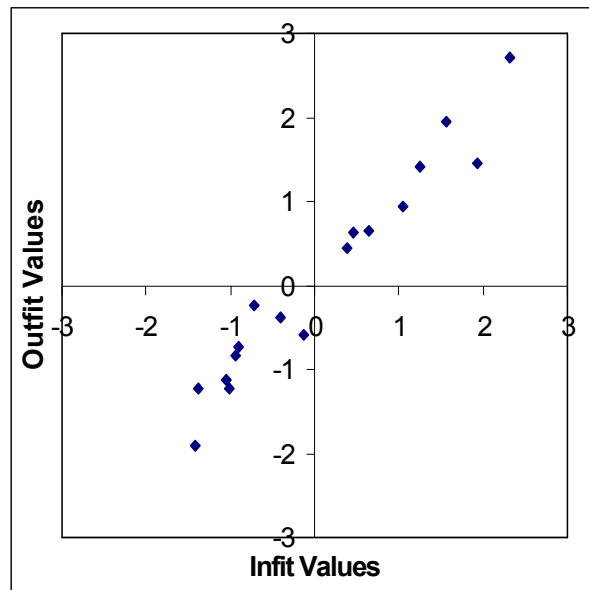


Figure 2. Scatterplot PCM. Total fit of items: TPI (17 items and 633 persons).

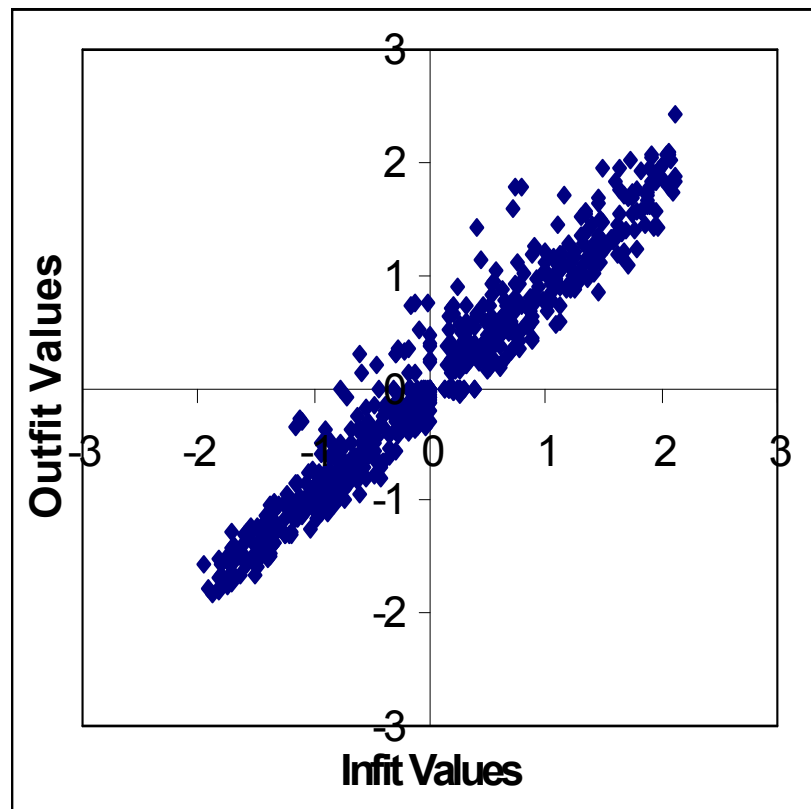


Figure 3. Scatterplot RSM. Total fit of persons: TPI (13 items and 628 persons).

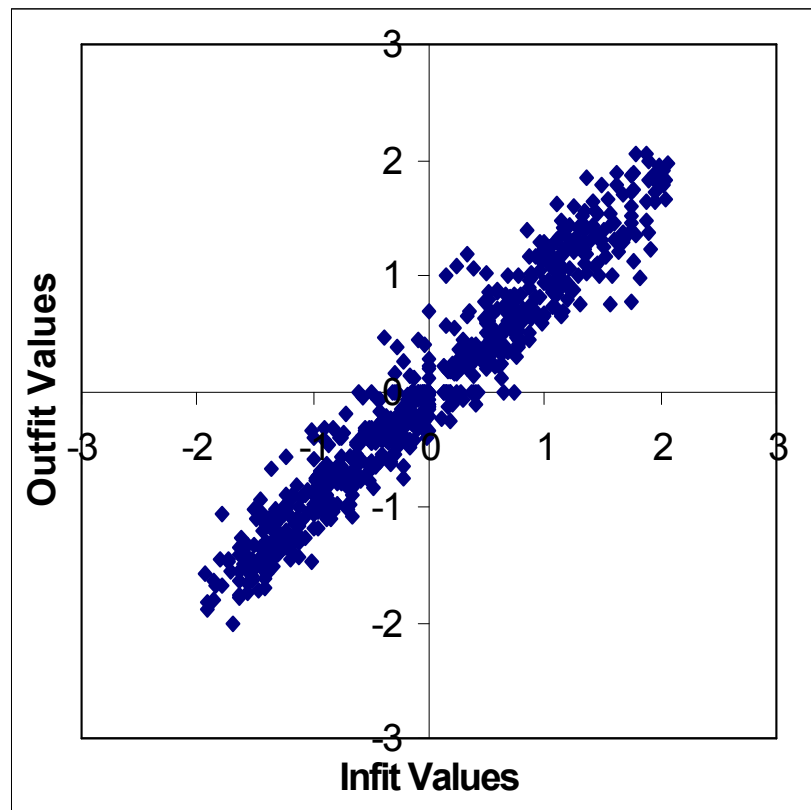


Figure 4. Scatterplot PCM. Total fit of persons: TPI.17 items and 633 persons.

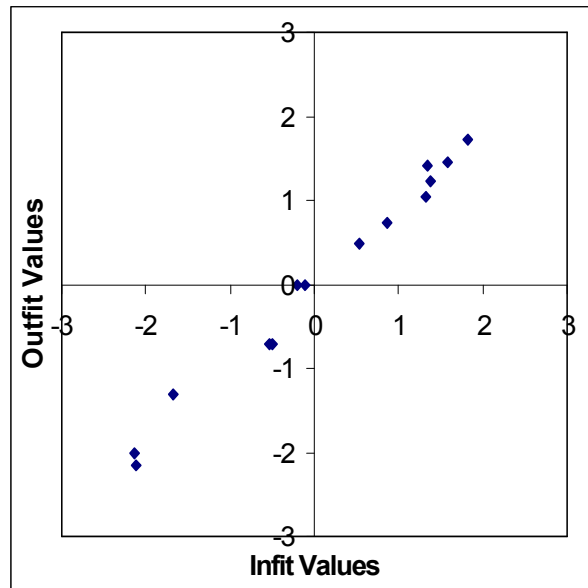


Figure 5. Scatterplot RSM. Total fit of items: TIP (14 items and 572 persons).

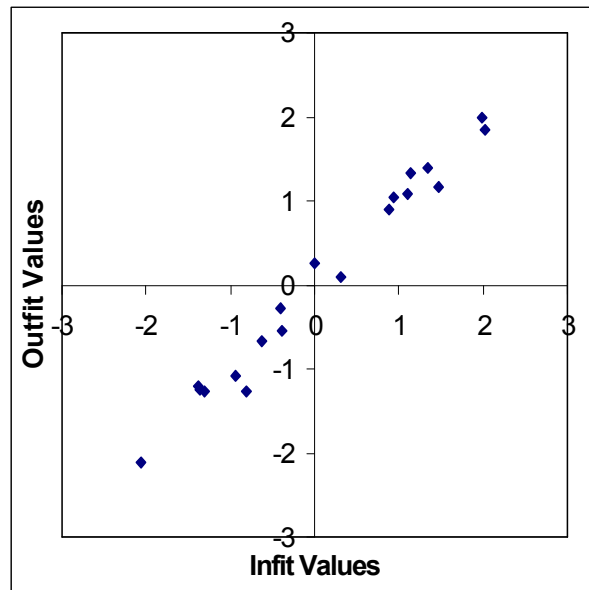


Figure 6. Scatterplot PCM. Total fit of items: TIP (20 items and 578 persons).

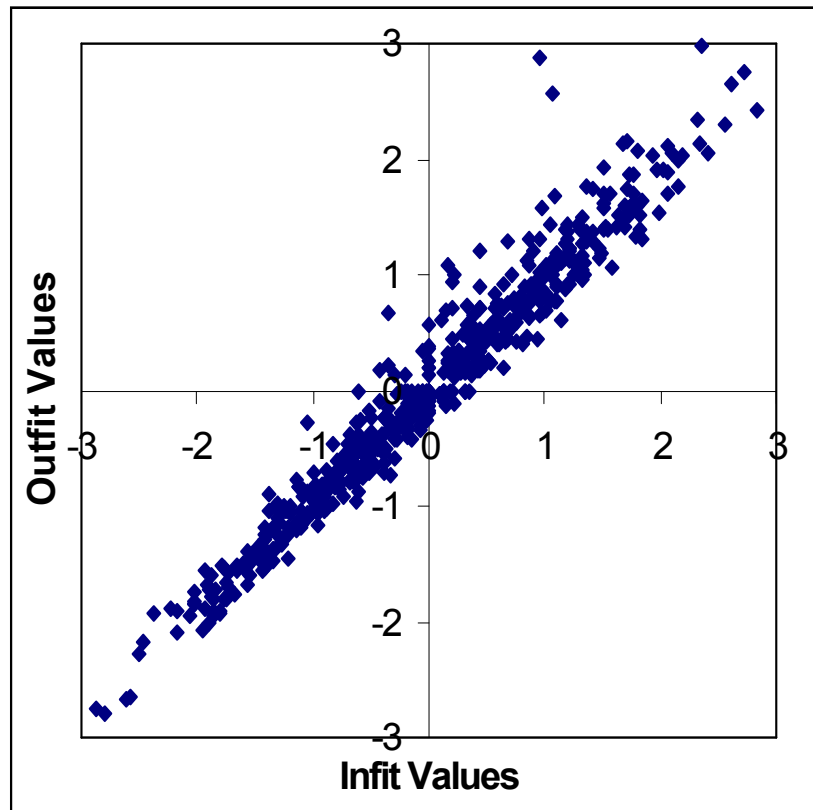


Figure 7. Scatterplot RSM. Total fit of persons: TIP (14 items and 572 persons).

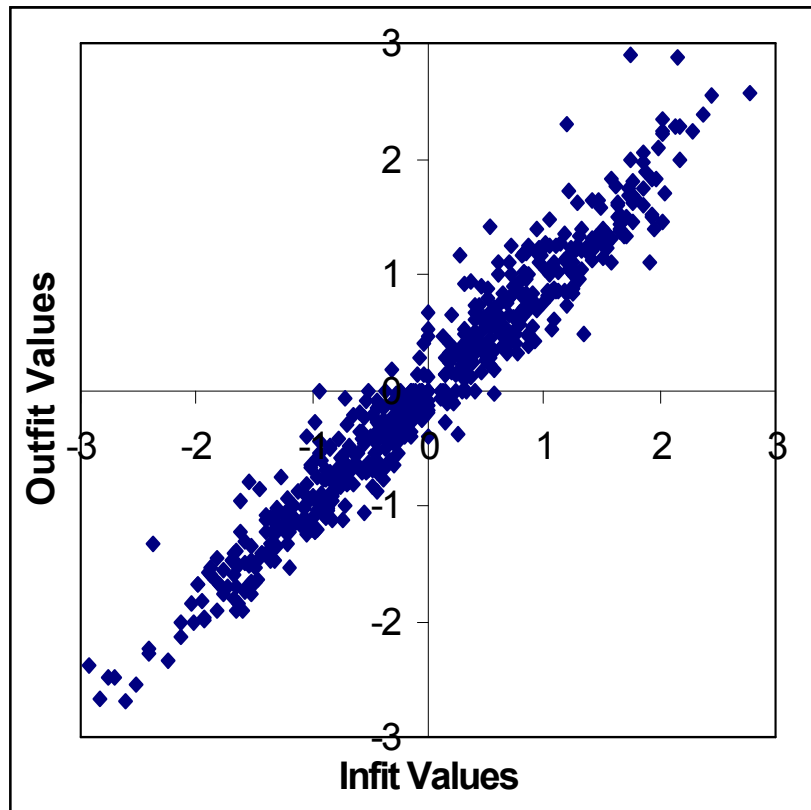


Figure 8. Scatterplot PCM. Total fit of persons: TIP (20 items and 578 persons).