Extensions and Applications of the Rasch Model

Afshin Khosravi

Extensions and Applications of the Rasch Model

Afshin Khosravi

Eurometric Institute

Afshinkhosravi1000@gmail.com

Abstract

In this monograph different extensions of the Rasch model are briefly introduced. The aim is to provide simple and accessible descriptions of the most commonly used Rasch models. Applications of each model in research are also listed for more information.

Rating Scale Model

Rating scale model (Andrich, 1978) model expresses that the probability of person *n* scoring *x* on the *m*-step item *i* is a function of person's location θ *n* and the difficulties of the *m* steps. In this model the thresholds are combined with the item parameter estimates. Andrich advanced RSM for ordered item categories data. This model assumes equal category thresholds across the items. Step difficulties in ordered item categories are deemed to be governed by a predefined set of response categories which are repeated for all the questions. Since the same responses alternatives such as 'strongly disagree', 'disagree', 'undecided', 'agree' and 'strongly agree' are given for all the items it is assumed that step difficulties do not vary across the items. In other words, the distance between 'strongly agree' and 'agree' in all the items throughout the test is the same (Masters & Wright, 1984). That is, the increment in the level of the construct as a result of endorsing 'strongly agree' rather than 'agree' is equal for all the items. However, the model does not require that the distances between 'strongly disagree', 'disagree', 'undecided', 'agree' and 'strongly agree' be equal within a single question. The level of increment in the construct can be different when a respondent endorses 'strongly agree' rather than 'agree' compared to when he endorses 'undecided' rather than 'disagree'.

The model has been applied in validating questionnaires and in accounting for local dependence in testlet-based assessments (Baghaei, 2007; Baghaei, 2008a; Baghaei, 2008b; Baghaei, 2009; Baghaei, Monshi Toussi, & Boori, 2009; Baghaei, 2010a; Baghaei, 2011a; Baghaei, 2011b;

Pishghadam, Baghaei, Shams, & Shamsaee, 2011; Pishghadam, Baghaei, Shahriari Ahmadi, 2011; Baghaei, 2014; Baghaei & Cassady, 2014; Pishghadam, Baghaei, Bazri, & Ghaviandam, 2012; Tabatabaee-Yazdi, Motallebzadeh, Ashraf, & Baghaei, 2018).

Continuous Rating Scale Model

The Continuous Rating Scale Model **(**Müller, 1987, CRSM) assumes a response mechanism where a latent response variable, originally unbounded and following a normal distribution, is doubly truncated to fit the response format constraint. Specifically, Müller (1987) considered the rating scale as a straight line segment of midpoint *c* and length *d*. In keeping with Samejima (1973), the end points of that scale, that is, $c \pm d/2$, are assumed to be defined (e.g., by labels such as "extremely positive" and "extremely negative"), and the person is allowed to mark any point along the line segment.

In CRSM there is a dispersion parameter λ. The CRSM posits a uniform density of thresholds along the latent interval $[-\lambda d, \lambda d]$. Hence, λ parameterizes the range of the threshold distribution. The dispersion parameter indicates the degree to which the thresholds increase in a strictly monotonic fashion along the continuous rating scale. If respondents actually use the continuous scale in a continuous manner, the dispersion parameter takes on positive values greater than zero. Conversely, violations of model assumptions are indicated by negative values of the dispersion parameter. In the special case of $\lambda = 0$ the uniform threshold distribution degenerates into a single threshold $\tau = 0$ (the "degenerate case"; Müller, 1987/1999). The model has been used by Eckes (2011) in research.

Partial Credit Model

Masters' (1982) partial credit model, on the other hand, is less restrictive than Andrich's (1978) RSM and equidistant model in that it does not require equal distances between the steps neither within items nor across items. Therefore, each item has a unique rating scale structure. That is, the distances between the steps can vary for all the items and within each single item and even the number of steps can vary. This property of PCM makes it the model of choice for analyzing educational tests where the assumption of equal step difficulties across items is very unrealistic. Masters' partial credit model, which is the least restrictive model in terms of the distances

between the steps within and across items, is the most appropriate model to analyse a reading comprehension test or any other educational test in which several items are based on a prompt. The model has been used by Baghaei, Hohensinn, and Kubinger (2014) in research.

Equidistant Model

Andrich (1982) proposed a model called 'equidistant model'. This model assumes that the distances between the thresholds within the items are equal but not necessarily across the items. The model was especially suggested to account for local dependency in educational tests where several items are based on one prompt by forming super-items. The model has been used by Baghaei (2010b) in research and the results are compared with partial credit model and the rating scale model.

Many-Facet Rasch Measurement

Many-Facet Rasch Measurement (MFRM, Linacre, 1989) refers to a class of measurement models that extend the basic Rasch model by incorporating more variables (or facets) than the two that are typically included in a test (i.e., examinees and items), such as raters, scoring criteria, and tasks. In the analysis of performance assessments, MFRM allows the inclusion and estimation of the effects of additional facets that may be of interest besides test takers and items, such as raters, criteria, tasks, and assessment occasions. Within each facet, MFRM represents each element (i.e., each individual test taker, rater, criterion, task, etc.) by a separate parameter. The parameters denote distinct attributes of the facets involved, such as proficiency or ability (for test takers), severity or harshness (for raters), and difficulty (for scoring criteria or tasks). In most instances, the attribute of primary interest refers to test taker proficiency, such as when proficiency measures are used to inform decisions on university admission, placement, or graduation. As a critical MFRM feature, the measures of test taker proficiency compensate for variation in rater severity; that is, these measures are adjusted for the differences in the level of severity characterizing the raters who assigned the ratings. For applications of the model see Eckes, (2005/2008/2012), Myford and Wolfe (2003/2004), Wind and Schumacker (2017).

Mixed Rasch Model

Mixed Rasch model (MRM) or mixture distribution Rasch model (Rost, 1990) identifies latent classes of persons for whom the Rasch model holds separately. MRM is a combination of Rasch model and latent class analysis. The idea is that the Rasch model holds for classes of persons within a population with differing difficulty order for items in the latent classes. The model is a unidimensional model, however, the intended dimension changes across the classes. Under the standard unidimensional Rasch model item difficulty estimates should remain constant for different groups of people. MRM can account for data when difficulty patterns of items consistently differ in classes of population. MRM allows item parameters to vary across classes of population, i.e., when the unidimensional RM does not fit for the entire population (Rost, 1990; Rost & von Davier, 1995; Yamamoto, 1987). MRM has been used in personality testing to identify latent classes differing in the use of response scale. Mixed Rasch models can detect examinee heterogeneity and the associated item profiles, the latent score distribution and the size of latent classes. It can also help to test the fit of unidimensional Rasch models (Rost, 1990). The model is used in Baghaei and Carstensen (2013); Pishghadam, Baghaei, and Seyednozadi (2017); and Baghaei, Kemper, Reichert, and Greif (2019).

Linear logistic test model

LLTM is an extension of the Rasch model (RM, Rasch, 1960/1980) which decomposes item parameters into a linear combination of several basic parameters that are defined a priori. In other words, the item parameters is decomposed into a weighted sum of basic parameters. In laymen terms, LLTM assumes that the Rasch model item difficulty parameters are composed of the difficulty of several cognitive components or item characteristics which linearly add up and lead to the overall estimated difficulty parameter. According to Gorin (2005) characteristics of an item can be classified as *radicals* and *incidentals*. Radicals are substantive components of items which are responsible for their difficulty, i.e., characteristics which can be manipulated to change the cognitive processing needed to solve the item. Incidentals are surface characteristics which are not expected to affect item difficulty and the processing load of items. For example, in math word problems the names of objects and people are incidentals. LLTM helps us quantify the difficulty of radicals and incidentals, if we hypothesize those incidentals also affect difficulty.

The major motivation behind the development of LLTM was the need in educational settings to break down learning materials into smaller manageable units for learners to master (Baghaei & Kubinger, 2015; Baghaei & Ravand, 2015; Baghaei & Hohensinn, 2017; Ghahramanlou, Zohoorian, & Baghaei, 2017; Fischer, 1973/2005; Hohensinn & Baghaei, 2017).

Rasch Poisson Counts Model

The Rasch Poisson Counts Model (RPCM, Rasch, 1960/1980) is a member of the family of Rasch models which was developed for tests where counts of errors or successes on a number of tasks are modeled instead of replies to individual items. Modeling the number of errors might be the only option when the number of potential successes is not well defined, say in classic oral reading tests where examinees are to read a passage aloud and the test administrator counts the number of errors (Rasch, 1960/1980). In such conditions, the total scores or the total number of errors on each block are assumed to be the realization of a Poisson process (e.g., Ross, 1983). That is, the number of correct checks (or errors) on each block for each person is assumed to be Poisson distributed and is the unit of analysis Baghaei & Doebler, 2018).

In most item response theory (IRT) models, the unit of analysis is the individual item. In such models, the probability that a person correctly answers an item or endorses certain categories is modeled. However, common IRT models need at least one parameter per item (any many more on polytomous IRT models), so they are relatively complex for situations where the same task or many simple tasks are given to examinees and aggregation of hits/misses is conducted. Such testing conditions arise in psychomotor testing (Spray, 1990), the testing of attention/processing speed (Baghaei, Ravand, & Nadri, 2019; Doebler & Holling, 2016; Nadri, Baghaei, & Zohoorian, 2019), oral reading errors (Jansen, 1997; Rasch, 1960/1980; Verhelst & Kamphuis, 2009), reading comprehension (Verhelst & Kamphuis, 2009), and divergent thinking (Forthmann et al., 2016). In these tests, examinees usually have to solve an unlimited (or at least very large) number of relatively easy items within a fixed period of time. Another example is identifying correctly spelled words in a long list of words. In such testing situations, the total scores (raw counts) or the total numbers of errors on the tasks are modeled instead of the individual attempts.

Multidimensional Rasch Model

Multidimensional IRT models (MIRT) account for multiple dimensions in a dataset and estimate students' abilities separately on the dimensions involved. Fitting a unidimensional model to a multidimensional test results in loss of information and disappearance of subscales. As a consequence we cannot investigate possible relationships among dimensions (Adams, Wilson, & Wang, 1997; Brandt, 2012; Höhler, Hartig, & Goldhammer, 2010). Reckase (1997) considers MIRT as a special case of FA because both try to detect hypothetical scales or factors on the basis of a matrix of observed scores. However, the focus of the two approaches is quite different; while FA attempts to reduce the data to a minimum number of underlying factors, MIRT tries to parameterize items and persons on a common scale so that one can predict the chances of success of a person with a known ability parameter on an item with a known difficulty parameter. That is, in MIRT models we want to model the interaction between persons and items in order to understand the characteristics of persons and items and the nature of their interaction (Reckase, 1997). The more dimensions we extract from the data the more precise our understanding of the nature of the interaction will be. Therefore, one of the basic distinctions between MIRT models and FA is that MIRT models do not focus on reducing the data to a minimum number of underlying factors. MIRT accounts for profiles of proficiency rather than overall proficiency and, as was mentioned above, items can measure one or more latent dimensions. Adams et al. (1997) summarize the advantages of analyses based on multidimensional models as follows:

1. They take care of the intended structure of the test in terms of the number of subscales.
2. They provide estimates of the relationships among the dimensions.
3. They make use of the relationships among the dimensions to produce more accurate item and person estimates.
4. They are single rather than multistep analyses.
5. They provide more accurate estimates than the consecutive approach.
6. Unlike the consecutive approach they can be applied to tests that contain items which load on more than one dimension.

The model has been used in research in validation and scaling (Baghaei, 2013; Baghaei, 2016; Baghaei & Aryadoust, 2015; Baghaei & Grotjahn, 2014a; Baghaei & Grotjahn, 2014b; Shoahosseini & Baghaei, 2020). It should be noted that cognitive diagnostic models (de la Tore,

) are also a type of multidimensional item response theory models (Effatpanah, Baghaei, &Boori, 2019; Ravand & Baghaei, 2019).

Testlet Rasch Model

Local dependence produced by passage dependence is the only source of LID that has widely been addressed in educational testing. Items that share the same passage or stimulus—usually called testlets—can produce LID. However, other sources of LID, their impact, and strength have been ignored in educational testing. To solve the problem of local dependence due to shared passages, Wang and Wilson (2005) advanced the Rasch testlet model, which is an extension of the standard Rasch model (Rasch, 1960) with an extra random effects parameter, γ, to model the dependence among items. The core of these models is the introduction of an extra random effects parameter, to the standard Rasch model to account for local dependency within each testlet. In this model, LID is considered a random effects parameter or a person characteristic that is common to all item responses within a testlet but different across testlets (Wang & Wilson, 2005). Modeling LID due to shared stimuli with a random effects parameter makes Rasch testlet model a special case of multidimensional Rasch model (Baghaei, 2012) or more precisely a bifactor IRT model where each item loads on two dimensions: an overall ability dimension and a testlet-specific dimension (Wang & Wilson, 2005). That is, each testlet is modeled as a latent trait whose items load on a general factor and a testlet-specific factor. The model has been used in several studies in language assessment (Baghaei & Ravand, 2016; Baghaei & Ravand, 2019; Eckes & Baghaei, 2015).

Rasch Model and Validation

A major threat to construct validity that obscures score meaning and its interpretation, according to Messick (1989), is *construct under-representation*. This refers to the imperfectness of tests in accessing all features of the construct. Another threat to construct validity is referred to as construct-irrelevant variance by Messick. There are always some unrelated sub-dimensions that creep into measurement and contaminate it. These sub-dimensions are irrelevant to the focal construct and in fact we do not want to measure them, but their inclusion in the measurement is inevitable. It is argued here that misfitting items are indications of construct irrelevant variance and gaps along the unidimensional continuum are indications of construct underrepresentation (Baghaei, 2008).

The formulation of the Rasch model (RM) provides a mechanism to check the two conditions for validity as delineated by Borsboom, et al. (2004). When the RM fits, it means that the most important assumption of the model, i.e., conditional independence holds. Conditional independence stipulates that conditioning on the latent trait the item residuals should be uncorrelated. In other words, after the shared variances among the items (observed variables) are captured the unique variances (residuals or errors) should be independent. The latent trait is incorporated in the covariation among the items and when it is extracted the relationship between the items is eliminated. Therefore, it is logical to assume that the latent trait is the cause of their covariations although other reasons cannot be ruled out (Baghaei & Tabatabaee-Yazdi, 2016). This is interpreted as the existence of a construct underlying all the item responses which causes their variations. Therefore, when the RM holds the first condition for validity is satisfied.

In the Rasch model formula, if θn, i.e., the latent trait, increases the probability of a correct response increases and if it decreases the probability of a correct response decreases. In other words, levels of the construct, indicated by the person parameter θ, causally determine the observed score Xni. Therefore, the causal relationship between the latent trait and the test scores can be tested. Note that in the latent trait models the latent variable is linked to the probability of the observed data not the data themselves. If a test is valid, Equation 1 should correctly predict the probability of a correct response to each item. This can be checked by examining the item characteristic curves and the item fit values (Baghaei & Shoahosseini, 2019).

Bibliography

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

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*, 561-573.

Andrich, D. (1982). An extension of the Rasch model for ratings providing both location and dispersion parameters, *Psychometrika, 47*, 105-113.

Baghaei, P., Ravand, H., Nadri, M. (2019). Is the d2 test of attention Rasch scalable? Analysis with the Rasch Poisson Counts Model. Advance online publication. *Perceptual and Motor Skills*, *126*, 70-86.

Baghaei, P. (2012). The application of multidimensional Rasch models in large scale assessment and validation: An empirical example. *Electronic Journal of Research in Educational Psychology, 10,* 233-252.

Baghaei, P. (2008). The Rasch model as a construct validation tool. *Rasch Measurement Transactions, 22,* 1145-1146.

Baghaei, P., & Tabatabaee-Yazdi, M. (2016). The logic of latent variable analysis as validity evidence in psychological measurement. *The Open Psychology Journal, 9,* 168-175.

Baghaei, P., & Shoahosseini, R. (2019). A note on the Rasch model and the instrument-based account of validity. *Rasch Measurement Transactions, 32,* 1705-1708.

Baghaei, P. (2010). A comparison of three polychotomous Rasch models for super-item analysis. *Psychological Test and Assessment Modeling, 52,* 313-323.

Baghaei, P., Kemper, C., Reichert, M., & Greif, S. (2019). Mixed Rasch modeling in assessing reading comprehension. In Aryadoust, V. & Raquel, M. (Eds.), *Quantitative Data Analysis for Language Assessment* (Vol. II) (pp.15-32). New York: Routledge.

Baghaei, P. (2016). Modeling multidimensionality in foreign language comprehension tests: An Iranian example. In Aryadoust, V., & Fox, J. (Eds.), *Trends in Language Assessment Research and Practice: The View from the Middle East and the Pacific Rim* (pp.47-66). New Castle, England: Cambridge Scholars.

Baghaei, P., & Grotjahn, R. (2014). The validity of C-Tests as measures of academic and everyday language proficiency: A multidimensional item response modeling study. In R. Grotjahn (Ed.). *Der C-Test: Aktuelle Tendenzen/The C-Test: Current trends* (pp. 163-171.). Frankfurt/M.: Lang.

Baghaei, P. (2014). Development and validation of a C-Test in Persian. In R. Grotjahn (Ed.). *Der C-Test: Aktuelle Tendenzen/The C-Test: Current trends* (pp.299-312). Frankfurt/M.: Lang.

Baghaei, P. (2010). An investigation of the invariance of Rasch item and person measures in a C-Test. In R. Grotjahn (Ed.). *Der C-Test: Beiträge aus der aktuellen Forschung/ The C-Test: Contributions from Current Research* (pp.100-112). Frankfurt/M.: Lang.

Baghaei, P., & Ravand, H. (2019). Method Bias in Cloze Tests as Reading Comprehension Measures. *SAGE Open*.

Baghaei, P. & Doebler, P. (2018). Introduction to the Rasch Poisson Counts Model: An R tutorial. *Psychological Reports*. Advance online publication.

Baghaei, P., & Hohensinn, C. (2017). A method of Q-matrix validation for the linear logistic test model. *Frontiers in Psychology, 8*: 897.

Baghaei, P., & Ravand, H. (2016). Modeling local item dependence in cloze and reading comprehension test items using testlet response theory. *Psicológica, 37*, 85-104.

Baghaei, P., & Ravand, H. (2015). A cognitive processing model of reading comprehension in English as a foreign language using the linear logistic test model. *Learning and Individual Differences, 43,* 100-105.

Baghaei, P. & Kubinger, K. D. (2015). Linear logistic test modeling with R. *Practical Assessment, Research & Evaluation, 20*, 1-11.

Baghaei, P., & Aryadoust, V. (2015). Modeling local item dependence due to common test format with a multidimensional Rasch model. *International Journal of Testing, 15*, 71–87.

Baghaei, P., & Cassady, J. (2014). Validation of the Persian translation of the Cognitive Test Anxiety Scale. *Sage Open, 4*, 1-11.

Baghaei, P., Hohensinn, C., & Kubinger, K.D. (2014). The Persian adaptation of the foreign language reading anxiety scale: A psychometric analysis. *Psychological Reports, 114*, 315-325.

Baghaei, P., & Grotjahn, R. (2014). Establishing the construct validity of conversational C-Tests using a multidimensional Item Response Model. *Psychological Test and Assessment Modeling, 56*, 60-82.

Baghaei, P., & Carstensen, C. H. (2013). Fitting the mixed Rasch model to a reading comprehension test: Identifying reader types. *Practical Assessment, Research & Evaluation, 18,* 1-13.

Baghaei, P. (2013). Development and psychometric evaluation of a multidimensional scale of willingness to communicate in a foreign language. *European Journal of Psychology of Education, 28,* 1087-1103.

Baghaei, P. (2011). Do C-Tests with different number of gaps measure the same construct? *Theory and Practice in Language Studies, 1*, 688-693.

Baghaei, P. (2011). Optimal number of gaps in C-Test passages. *International Education Studies, 4*, 166-171.

Baghaei, P. (2009). A Rasch-informed standard setting procedure. *Rasch Measurement Transactions, 23*, 1214.

Baghaei, P., Monshi Toussi, M.T., & Boori, A. A. (2009). An Investigation into the validity of conversational C-Test as a measure of oral abilities. *Iranian EFL Journal, 4*, 94-109.

Baghaei, P. (2008). The effects of the rhetorical organization of texts on the C-Test construct: A Rasch modelling study. *Melbourne Papers in Language Testing, 13*, 32-51.

Baghaei, P. (2008). An attempt to fit the Rasch model to a C-Test. *Iranian EFL Journal, 2,* 6-15.

Baghaei, P. (2007). Applying the Rasch rating-scale model to set multiple cut-offs. *Rasch Measurement Transactions. 20,* 1075-76.

Baghaei, P., Yanagida, T., & Heene, M. (2017b). A new fit statistic for the dichotomous Rasch model: Modified cut-off values. *Rasch Measurement Transactions, 31, 1621-1624.*

Baghaei, P. (2009). *Understanding the Rasch model*. Mashhad: Mashhad Islamic Azad University Press.

Baghaei, P., Yanagida, T., & Heene, M. (2017a). Development of a descriptive fit statistic for the Rasch model. *North American Journal of Psychology, 19,* 155-168.

Baghaei, P., & Dourakhshan, A. (2016). Properties of single-response and double-response multiple-choice grammar items. *International Journal of Language Testing, 6,* 33-48.

Baghaei, P., & Amrahi, N. (2011a). The effects of the number of options on the psychometric characteristics of multiple choice items. *Psychological Test and Assessment Modeling, 53,* 192-211.

Baghaei, P., & Amrahi, N. (2011b). Validation of a multiple choice English vocabulary test with the Rasch model. *Journal of Language Teaching and Research, 2,* 1052-1060.

Baghaei, P. (2011). Test score equating and fairness in language assessment. *Journal of English Language Studies, 1,* 113-128.

Baghaei, P., & Amrahi, N. (2009). Introduction to Rasch measurement. *Iranian EFL Journal, 5,* 139-154.

Baghei, P. (2007). Local dependency and Rasch measures. *Rasch Measurement Transactions. 21,* 1105-1106.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Lawrence Erlbaum.

Borsboom D., Mellenbergh G.J., & van Heerden J. (2004). The concept of validity. *Psychological Review, 111,* 1061-71.

Brandt, S. (2012). Robustness of multidimensional analyses against local item dependence. *Psychological Test and Assessment Modeling, 54,* 36–53.

Doebler, A., & Holling, H. (2016). A processing speed test based on rule-based item generation: An analysis with the Rasch Poisson Counts model. *Learning and Individual Differences, 52*, 121-128.

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly, 2*, 197–221.

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing, 25*, 155–185.

Eckes, T. (2012). Operational rater types in writing assessment: Linking rater cognition to rater behavior. *Language Assessment Quarterly, 9*, 270–292.

Eckes, T., & Baghaei, P. (2015). Using testlet response theory to examine local dependency in C-Tests. *Applied Measurement in Education, 28*, 85–98.

Eckes, T. (2011). Item banking for C-tests: A polytomous Rasch modeling approach. *Psychological Test and Assessment Modeling, 53*, 414-439.

Effatpanah, F., Baghaei, P., Boori, A. (2019). Diagnosing EFL learners' writing ability: A Diagnostic Classification Modeling Analysis. *Language Testing in Asia*.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37*, 359-374.

Fischer, G.H. (2005). Linear logistic test models. In *Encyclopedia of Social Measurement, 2*, 505-514.

Forthmann, B., Gerwig, A., Holling, H., Celik, P., Storme, M., & Lubart, T. (2016). The be-creative effect in divergent thinking: The interplay of instruction and object frequency. *Intelligence, 57*, 25-32.

Ghahramanlou, M., Zohoorian, Z., & Baghaei, P. (2017). Understanding the cognitive processes underlying performance in the IELTS listening comprehension test. *International Journal of Language Testing, 7*, 62-72.

Gorin, J. S. (2005). Manipulating Processing Difficulty of Reading Comprehension Questions: The Feasibility of Verbal Item Generation. *Journal of Educational Measurement, 42*, 351-373.

Hohensinn, C., & Baghaei, P. (2017). Does the position of response options in multiple-choice tests matter? *Psicológica, 38*, 93-109.

Höhler, J., Hartig, J., & Goldhammer, F. (2010). Modeling the multidimensional structure of students' foreign language competence within and between classrooms. *Psychological Test and Assessment Modeling, 52*, 323-340.

Jansen, M. G. H. (1997b). Rasch's model for reading speed with manifest exploratory variables. *Psychometrika, 62*, 393-409.

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

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149- 174.

Masters, G. N., & Wright, B. D. (1984). The essential process in a family of Rasch models. *Psychometrika, 49*, 529-544.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13–103). New York, NY: Macmillan.

Müller, H. (1987). A Rasch model for continuous ratings. *Psychometrika, 52*, 165-181.

Müller, H. (1999). CRSM: A Fortran program for the analysis of continuous rating scale data according to a Rasch model for continuous responses (Version 1.3) [Computer software]. Department of Psychology, University of Vienna, Vienna, Austria.

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement, 4*, 386–422.

Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement, 5*, 189–227.

Nadri, M., & Baghaei, P., & Zohoorian, Z. (2019). Analysis of the Ruff 2 & 7 Test of Attention with the Rasch Poisson Counts Model. *The Open Psychology Journal, 12*, 7-11.

Pishghadam, R., Baghaei, P., Bazri, E, & Ghaviandam. S. (2012). Using Rasch to validate a measure of English language teacher prejudice. *Journal of the Teaching English Language and Literature Society of Iran (TELL), 6*, 25-47.

Pishghadam, R., Baghaei, P., Shahriari Ahmadi, H. (2011). Development and validation of an English language teacher competency test using Item Response Theory. *The International Journal of Educational and Psychological Assessment, 8*, 54-68.

Pishghadam, R., Baghaei, P., Shams, M.A., & Shamsaee, S. (2011). Construction and validation of a narrative intelligence scale with the Rasch rating scale model. *The International Journal of Educational and Psychological Assessment, 8*, 75-90.

Pishghadam, R., Baghaei, P., & Seyednozadi, Z. (2017). Introducing emotioncy as a potential source of test bias: A mixed Rasch modeling Study. *International Journal of Testing, 17*, 127-140.

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research, 1960. (Expanded edition, Chicago:

The university of Chicago Press, 1980).

Rasch, G. (1977). *On specific objectivity: An attempt at formalizing the request for generality and validity of scientific statements*. Danish Yearbook of Philosophy, 14, 58-93.

Ravand, H., & Baghaei, P. (2019). Diagnostic classification models: Recent development, practical issues and prospects. *International Journal of Testing*. Advance online publication. Doi: 10.1080/15305058.2019.1588278

Reckase, M. D. (1997). The past and the future of multidimensional item response theory. *Applied Psychological Measurement, 21*, 25-36.

Ross, S. M. (1983). *Stochastic processes*. New York, NY: Wiley.

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement, 14*, 271 – 282.

Rost, J., & von Davier, M. (1995). Mixture distribution Rasch models. In G. H. Fischer, & I. W. Molanaar (Eds.), *Rasch models: Foundations, recent developments and applications* (pp. 257-268). New York: Springer Verlag.

Samejima, F. (1973). Homogeneous case of the continuous response model. *Psychometrika, 38*, 203-219.

Shoahosseini, R., & Baghaei, P. (2020). Validation of the Persian translation of the Children's Test Anxiety Scale: A multidimensional Rasch model analysis. *European Journal of Investigation in Health Psychology and Education, 10*, 59–69.

Spray, J. A. (1990). One-parameter item response theory models for psychomotor tests involving repeated independent attempts. *Research Quarterly for Exercise and Sport, 61*, 162-168.

Tabatabaee-Yazdi, M., Motallebzadeh, K., Ashraf, H., & Baghaei, P. (2018). Development and Validation of a Teacher Success Questionnaire Using the Rasch Model. *International Journal of Instruction, 11*, 129-144.

Verhelst, N. D., & Kamphuis, F. H. (2009). *A Poisson-Gamma model for speed tests*. Measurement and Research Department. Reports 2009-2. Arnhem, the Netherlands: Cito.

Wang, W.-C., & Wilson, M. (2005). The Rasch testlet model. *Applied Psychological Measurement, 29*, 126-149.

Wind, S. A., & Schumacker, R. E. (2017). Detecting measurement disturbances in rater-mediated assessments. *Educational Measurement: Issues and Practice, 36*, 44–51.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago: MESA Press.

Wright, B. D., & Panchapakesan, N. (1969). A procedure for sample-free item analysis. *Educational and Psychological Measurement, 29*, 23-48.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago: Mesa Press.

Yamamoto, K. Y. (1987). *A model that combines IRT and latent class models*. Unpublished doctoral dissertation, University of Illinois Urbana Champaign.