Eurasian Journal of Educational Research
www.ejer.com.tr

A Comparison of Difficulty Indices Calculated for Open-Ended Items According to Classical Test Theory and Many Facet Rasch Model

Mustafa ILHAN [1], Nese GULER [2]

ARTICLE INFO

Article History:
Received: 23 Jan. 2018
Received in revised form: 09 Mar. 2018
Accepted: 26 Mar. 2018
DOI: 10.14689/ejer.2018.75.6

Keywords
open ended question, item difficulty index classical test theory, many facet rasch model

ABSTRACT

Purpose: This study aimed to compare difficulty indices calculated for open-ended items in accordance with the classical test theory (CTT) and the Many-Facet Rasch Model (MFRM). Although theoretical differences between CTT and MFRM occupy much space in the literature, the number of studies empirically comparing the two theories is quite limited. Therefore, this study is expected to be a substantial contribution to the literature.

Research Methods: The research data were collected through three teachers rating the answers given by 375 eighth grade students to ten open-ended questions in a mathematics test. The difficulties of the items in the test were calculated according to CTT and MFRM by using the obtained data, and the consistency between the difficulty indices estimated based on the two theories was tested. While the Microsoft Excel program was used in the analyses for CTT, the FACETS package was employed in the analyses for MFRM.

Findings: The research findings showed that CTT and MFRM yielded similar results in terms of difficulty indices of open-ended questions. It was found that, according to both theories, the ten items in the achievement test were ranked as I2, I3, I1, I4, I7, I6=I8, I5 and I9, from easiest to most difficult.

Implications for Research and Practice: It may be said that estimating item difficulties according to either CTT or MFRM will not cause any notable differences in terms of the items to be included or excluded in the development of an achievement test with open-ended questions.

[1] Dicle University, TURKEY, e-mail: mustafailhan21@gmail.com, ORCID: orcid.org/0000-0003-1804-002X
[2] İzmir Democracy University, TURKEY, e-mail: gnguler@gmail.com, ORCID: orcid.org/0000-0002-2836-3132

Introduction

The majority of constructs intended to be measured in education and in psychology are abstract and cannot be directly observed. For this reason, stimuli to transform the constructs into observed outcomes are needed to measure such constructs. The stimuli expected to uncover certain types of responses in individuals are called *items*, and the process of selecting the items to stimulate only the properties of individuals intended to be measured on the basis of certain criteria is referred to as *test development*. In this context, selecting the appropriate items in the test development process is the pre-requisite to accurately measuring a property. The appropriate items are selected through item analysis. Item difficulty and discrimination indices are calculated, and efforts are realized to determine the functioning of items in item analysis. Item analysis is a common procedure in all test development processes. However, the following stages can differ according to the theory of measurement used. There are two main theories used to estimate item statistics: the Classical Test Theory (CTT) and the Item Response Theory (IRT).

Classical Test Theory (CTT)

CTT, which is also called the True Score Theory, is described with the concepts of true score, observed score, and random error. According to CTT, a value found in consequence of a measurement operation represents the observed score for the measured property, and the score is composed of the true score and random error (Kline, 2005). Therefore, whether the measured property reaches its true value depends on if the random error in measurement is zero. Nevertheless, it is inevitable for measurements to contain a certain amount of error, no matter how carefully the measurement is performed. Therefore, it is impossible to reach the true score in measurement activities, and true scores is estimated by means of observed scores. The estimation is based on the assumptions that true scores and error scores are uncorrelated, that there are no systematic patterns between the error scores obtained from the parallel applications of the same measurement tool, and that the expected value of the error scores is zero (Hambleton & Jones, 1993). CTT does not have many assumptions, which is considered as an advantage as it is easier to apply CTT to several test situations (Kelecioğlu, 2001). In addition, the mathematical operations it requires are not difficult, and it can be used with small samples (Schumacher, 2010). In addition to its positive aspects, which make it possible to use CTT in a wide range of areas, the theory also has certain limitations forcing researchers to search for new methods. First, item parameters are dependent on the group to which a test is administered, and the individuals' ability levels are dependent on the items available in a test in CTT (Hambleton, 2004). In addition, it is not possible to make an evaluation on individuals' performance at the level of items in CTT, since it is dependent on the total scores received from a test (Hambleton, Swaminathan, & Rogers, 1991). Other weaknesses of CTT include yielding only one standard error value for all individuals to whom a test is administered, difficulties performing measurements of high reliability with a small number of items, and that the reported data are in the ordinal scale (Embretson & Reise, 2000). These limitations of CTT have paved the way for new methods; thus, IRT was proposed, claiming that it could offer solutions to the abovementioned limitations.

Item Response Theory (IRT)

IRT is a theory of measurement that is based on the probabilistic relations between responses to items in a test and the construct a test aims to measure (Schultz & Whitney, 2005). The construct that is intended to be measured with a test but is not directly observed is called a *latent trait* in IRT. For this reason, the Latent Trait Theory is another name for IRT (De Ayala, 2009). Differing models were suggested throughout the historical development of IRT. The first model suggested within the framework of IRT was the Rasch model, which was developed for items rated in two categories and contains only difficulty parameter (DeVellis, 2003). A two-parameter model was developed with the inclusion of a discrimination parameter in the Rasch model, and a three-parameter model was developed with the inclusion of a guessing parameter in the two-parameter model (Furr & Bacharach, 2008). As can be understood, the first factor influential in the emergence of different models in the development process of IRT was the number of estimated item parameters. The second factor was the response categories in relation to items. IRT was first developed for items that were rated dichotomously. However, later, the use of the theory was not limited to dichotomously rated items and, thus, models for polytomous items (nominal response model, partial credit model and graded response theory) were also included in IRT (Harvey & Hammer, 1999; van der Linden, 2005). IRT is divided into two categories, parametric and non-parametric models, in terms of approaches considered in estimating the item characteristic curve. While parametric IRT models assume that the item characteristic curve has normal ogive or logistic properties, non-parametric models do not have an assumption limiting the item characteristic curve to a certain form (Takano, Tsunoda & Muraki, 2015). Another element distinguishing IRT models from each other is dimensionality. IRT is considered as unidimensional and multi-dimensional in this respect (Reckase, 2009). And finally, IRT can be considered in two groups, two-facet models and many-facet models, in terms of the number of variability sources, which can be influential in measurement results. The sources of variability that can affect measurement results are limited to -items and persons- in the two-facet model. On the other hand, other sources of variability (such as raters) apart from items and persons can also affect measurement results in the many-facet model. Today, IRT has only one model containing more than two sources of variability. The model, which is based on the Rasch analysis, is referred to as the Many-Facet Rasch Model (MFRM).

Many-Facet Rasch Model (MFRM)

MFRM was developed as an extension of the partial credit model by Linacre in 1989. MFRM is a model in which all sources of variability, such as raters, items and persons, which have the potential to influence measurement results, are considered simultaneously (Lunz & Stahl, 1990). In this respect, it differs from other IRT models that have two sources of variability labeled as items and persons, and it becomes a model that is primarily preferable in analyzing data from open-ended items (Mulqueen, Baker & Dismukes, 2000). Using MFRM in the analysis of subjectively rated tests enables researchers to compare all facets, such as raters, persons and items considered in analyses, on a common metric. It also enables researchers to detect rater errors (such as Halo effects, central tendency, bias, range restriction etc.) and assures that measurements for raters are also taken into consideration in estimations for item difficulty and individuals' ability levels (Lynch & McNamara, 1998). Furthermore, MFRM includes all the advantages common to the other IRT models. In other words, abilities can be estimated for individuals independently of item parameters, and item parameters can be estimated independently of individuals' ability levels in MFRM, similarly to other IRT models when the model-data fit has been attained (Sudweeks, Reeve, & Bradshaw, 2005). Additionally, the data in the ordinal scale are brought to the level of those in the interval scale in MFRM, and separate error values are reported for each element in facets of measurement, in contrast to CTT, which yields only one value of standard error (Prieto & Nieto, 2014). Therefore, MFRM offers more advantages than CTT in those aspects. Nevertheless, discussions on how the differences between MFRM and CTT are reflected into the analysis results based on the two theories still occupy a significant place in the literature of measurement and evaluation.

Studies Comparing CTT and MFRM

Although MFRM entered the literature approximately 30 years ago, empirical studies comparing MFRM with CTT started after the 2000s. The first study to compare the results obtained through CTT to those obtained through MFRM was performed by MacMillian (2000), and an increase of similar studies was observed in the following years. Studies concerning a comparison between CTT and MFRM available in the literature are summarized in Table 1.

Table 1

Studies Concerning a Comparison between CTT and MFRM available in the Literature

Study Tag	Purpose of Study
MacMillian (2000)	The study examines the consistency between results reported in CTT, MFRM and generalizability theory in rater reliability.
Haiyang (2010)	Reliability coefficients calculated for an English test including open-ended questions according to CTT and MFRM are compared.
Kadir (2013)	Findings on analyses on the basis of CTT, MFRM, and the generalizability theory are compared in assessing writing performance in English.
Huang, Guo, Loadman, and Low (2014)	Item difficulty parameters calculated in CTT and MFRM and reliability estimated according to the two theories are compared.
İlhan (2016)	Ability estimations made according to CTT and MFRM are compared in terms of relative agreement, absolute agreement, and criterion-related validity.

As is clear from Table 1, the issue most frequently considered in comparing CTT and MFRM is the extent to which reliability values calculated according to the two theories agree. It is evident that studies that were conducted more recently compared the two theories in terms of ability estimation and item difficulty parameters. Upon examining the studies comparing the reliability values reported in CTT and MFRM, studies were found to differ in terms of designs used. Accordingly, some of those studies (Güler & Gelbal, 2010; Haiyang, 2010) used crossed designs, in which all the students' responses were assessed by the same group of raters. Others (Huang et al., 2014; MacMillian; 2014), however, employed nested designs, in which different groups of raters were utilized in the assessment process. Thus, the available studies in the literature presented comprehensive information on the degree to which reliability values calculated on the basis of the two theories are in agreement. However, the same cannot be said about the comparison of calculated item difficulty indices in accordance with CTT and MFRM. This is because only one study comparing item difficulty in accordance with CTT and MFRM was found in the literature (Huang et al., 2014), and that study used a nested design. No studies comparing the item difficulties calculated according to the two theories by using a crossed design were encountered in relevant literature. In addition, while the study conducted by Huang et al. (2014) estimated item difficulty according to CTT by basing it on measurements for the top 25% and bottom 25% groups, the study used measurements for all individuals in estimating item difficulty according to MFRM. Such a difference is thought to be important for a study comparing the item difficulty values calculated according to both theories. In this sense, comparing item difficulty values calculated according to CTT and MFRM with different measurement conditions from those mentioned in the literature would be considered valuable.

Purpose and Significance of the Study

This study aimed to compare item difficulty indices calculated according to CTT with those calculated according to MFRM for open-ended questions. The study employed a crossed design in which students' responses to open-ended items were assessed by the same group of raters. Furthermore, since measurements for all individuals were taken into consideration in estimating item difficulty according to MFRM, item difficulty indices in CTT were also calculated by including all the individuals in the analysis, and not on the basis of the top and bottom groups. Thus, this study differs from Huang et al. (2014) in this respect. For this reason, it may be said that the study is original and could contribute to the literature. The fact that it can function as a resource calculating item difficulty for open-ended questions on the basis of CTT is another property of this study that is expected to be a valuable contribution. Calculation of item difficulty according to CTT is generally restricted to multiple-choice tests in studies in Turkish literature. This study offers a detailed description on calculating CTT-based item difficulty. Therefore, it is thought to serve as an important resource in calculating difficulty indices for open-ended items in measurement activities where one or more than one rater is available.

Method

Research Model

This study aimed to compare difficulty indices calculated for open-ended items according to two different theories of measurement, which qualifies it as basic research. Basic research is concerned with generating new knowledge, unlike applied studies, which focus on the use of knowledge (Bickman & Rog, 2009). Therefore, studies aiming to develop a theory or compare the theories available in the literature are defined as basic research (Connaway & Powell, 2010).

Study Group

This study was conducted in Diyarbakır city center in the spring semester of the 2016-2017 academic year. The participants were 375 eighth graders, of which 183 (48.80%) were girls and 192 (51.20%) were boys, and three mathematics teachers who rated the students' responses to open-ended mathematics questions.

Data Collection Tool

The data were collected through an achievement test of open-ended questions and a holistic rubric used to grade the students' responses to the test items. The achievement test used in the study contained ten open-ended mathematics questions and was developed by Ilhan (2016a). The rubric developed by Ilhan (2016b) was employed in marking the responses to the open-ended items. The rubric has four categories: *inadequate*, *needs improvement*, *good,* and *very good*. The students' responses to the items were rated between 1 and 4, based on the four categories listed in the rubric. After grading, analyses for the validity and reliability of the measurements were realized.

The arithmetic mean was calculated for the grades given by three raters for each item within the scope of CTT-based validity and reliability analyses. Then, exploratory factor analysis was executed and Cronbach's alpha internal consistency coefficient was calculated. Accordingly, it was found that the explained variance was 70.60% in the factor analysis, and that the test items had one factor with factor loads ranging between .68 and .93. Cronbach's alpha internal consistency coefficient was found to be .95. Inter-rater correlation coefficients were also calculated for the estimation of rater reliability according to CTT, and the correlation values were found as .75 (rater1-rater2), .65 for (rater1-rater3), and .60 (rater2-rater3).

The psychometric properties of the collected data were analyzed not only on the basis of CTT, but also on the basis of MFRM. Table 2 shows the findings reported for reliability and model-data fit in MFRM. As is apparent from Table 2, the infit and outfit statistics are in the suggested interval of .5 and 1.5 for all three person, item, and rater facets (Wright & Linacre, 1994). These values for fit indices demonstrate that the model-data fit was attained and that the measurements are valid.

Table 2

Findings Reported for Reliability and Model-Data Fit in MFRM

	Person	Item	Rater
Infit	.99	.99	.99
Outfit	1.02	1.02	1.02
Separation ratio	4.45	10.57	39.11
Reliability	.95	.99	1.00
df	374	9	2
Chi-square	6467.3*	1000.1*	3063.3*

$^{*}p$<001

According to Table 2, the chi-square value for the rater facet is significant and the reliability coefficient and the separation ratio are high. This result indicates that the raters differed in severity/leniency. Despite the differences mentioned, the values reported for item and person facets show that the measurements are reliable. This is clear from Table 2 that the chi-square values for person and item facets are significant, the reliability coefficients are above .80, and the separation ratios are above 2 (Linacre, 2012). In other words, the students' performance on different items of the test were marked independently of each other, and students with differing ability levels were distinguished from each other effectively.

Data Analysis

The data were analyzed at two stages. First, CTT-based item difficulties were found. The following formula was used in calculating item difficulty indices for open-ended items:

Difficulty Index = (x – y) / (z – y)

x: Mean scores received from the item
y: The minimum score receivable from the item
z: The maximum score receivable from the item

The formula can be directly used in cases in which there is only one rater. However, when there is more than one rater, certain procedures should be followed prior to using the formula. The first step taken here was to calculate the mean scores assigned by different raters to students' responses to each item. The second step was to divide the sum total of the scores students had received from the items into the number of participants (separately for each item) to attain mean scores for the items. After that, the abovementioned formula was used. In other words, the difference was found for each item by subtracting the mean scores received from an item from the minimum score receivable from an item, and then the difference was divided into the item score range to attain the CTT-based item difficulty indices. The Microsoft Office Excel program was used in all operations for estimating item difficulty according to CTT.

Second, the MFRM analysis was executed in a design containing three sources of variability as persons, items, and raters. Before interpreting the analysis outcomes, whether the assumptions of MFRM were met was tested. As was stated above under Data Collection Tool, the fit statistics suggested that the model-data fit was attained.

Another indicator that the model-data fit had been attained, in addition to the infit and outfit statistics, was the standardized residual reported in MFRM. A model is considered to fit the data when the number of standardized residuals remaining outside ±2 interval is not above 5% of the total number of data, and when the number of standardized residuals remaining outside ±3 interval is not above 1% of the total amount of data in consequence of analyses (Linacre, 2014). While the proportion of the standardized residual outside the ±2 interval to the total number of data was 0.22%, according to the MFRM outcomes [25 out of 11250 (375x10x3) data], there were no data found with standardized residuals remaining outside the ±3 interval. Accordingly, it may be said that there is a high fit between the model and the data. Since the Rasch analysis was based on unidimensional data, the high fit between the model and the data indicated that the assumption of unidimensionality was met. As the assumption of unidimensionality functioned in parallel to local independence (Hambleton et al., 1991), attaining unidimensionality indicated that local independence – another assumption of MFRM – was also met. Having found that the assumptions were met, the measurement reports for the item facet were examined to determine the difficulty indices calculated in MFRM. The FACETS package program was used in the analyses for MFRM in this study.

After calculating item difficulty indices according to CTT and MFRM, correlations between difficulty indices estimated according to both theories were checked. Furthermore, a chart for the correlation between item difficulty indices found for CTT and for MFRM was created to visually express the correlation. Microsoft Excel was used in operations for calculating correlation coefficients and in forming the chart as in CTT-based item difficulty analyses.

Results

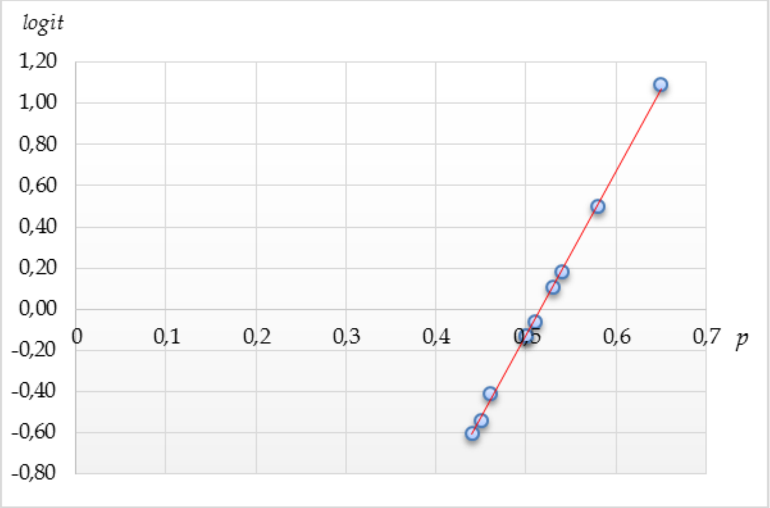
Difficulty indices calculated according to CTT and MFRM for the ten open-ended items in the achievement test used in this study are shown in Table 3. An item difficulty index close to zero in CTT demonstrates that an item is difficult, whereas a value close to 1 indicates that the item is easy. The way item difficulty indices are interpreted differs according to whether the item facet is positively or negatively oriented in MFRM. When an item facet is described as positively oriented, items become increasingly more difficult as one moves from the negative end of the logit scale to the positive end of it. On the contrary, when an item facet is described as negatively oriented, it is said that items with high logit values are easier and that the ones with low logit values are more difficult. Therefore, the item facet was defined as negatively oriented in the Rasch analysis to accurately compare the difficulty indices in CTT and in MFRM.

Table 3

Item Difficulty Indices Calculated in CTT and MFRM

Items	I1	I2	I3	I4	I5	I6	I7	I8	I9	I10
CTT	.54	.65	.58	.53	.45	.50	.51	.50	.44	.46
MFRM	.18	1.09	.50	.11	-.54	-.13	-.06	-.13	-.60	-.41

As is clear from Table 3, the items are ranked from easiest to most difficult as I2, I3, I1, I4, I7, I6=I8, I10, I5 and I9 in both CTT and MFRM. Thus, it may be said that there is a complete agreement between the item difficulties calculated according to both theories. This is also evident from the chart below showing the correlation between item difficulty indices calculated in CTT and MFRM.



r = .999, p<.001

Chart 1. Correlation between Item Difficulty Indices Calculated in CTT and MFRM

As is clear from Chart 1, there is a linear correlation between item difficulty indices estimated according to the two theories. Chart 1 includes 9 points, although there are ten items in the achievement test, because I6 and I8 have equal difficulty indices in both CTT and MFRM. Spearman's correlation coefficient shown at the bottom of Chart 1 suggests that there is a positive and perfect correlation between item difficulty indices estimated in CTT and MFRM.

Discussion and Conclusion

This study aimed to compare difficulty indices for open-ended items calculated according to CTT and MFRM. The results obtained in this study suggested that there was a high level of agreement between difficulty indices estimated according to the two theories. Upon ranking the items according to their difficulty, the rankings were found to be identical in both theories. This was a similar result as that obtained by Huang et al. (2014), which compared item difficulty indices in CTT with those in MFRM by using a nested design. In Huang et al. (2014), 124 competitive grant applications were rated according to a six-point graded rubric with 24 items. Sixty-four experts rated the proposals, and each of the 124 proposals was assessed by only three experts; therefore, each expert rated approximately six different proposals. However, this current study, instead used a crossed design in which all of the 375 student responses to the ten open-ended mathematics questions were assessed by the same group of raters. Therefore, upon considering the results obtained in Huang et al. (2014) and those obtained in this study, the CTT and the MFRM yielded similar results in terms of item difficulty indices for open-ended questions, no matter which design – crossed or nested – was used.

Upon comparing the findings obtained in Huang et al. (2014) with those obtained in this study, it may be inferred that item difficulty indices in CTT – whether they are calculated by comparing the top and bottom groups or by including all individuals in analyses – agree with those reported in MFRM. This is because the item difficulties in CTT were calculated by comparing the top and bottom groups in Huang et al. (2014), but they were estimated by including all the individuals in the analyses in this study. Despite this difference, the difficulty indices calculated according to CTT and MFRM were found to agree in both studies.

It may be stated, based on the results of this study, that estimating item difficulty according to CTT or MFRM does not cause a difference in terms of the items to be included or excluded in the development of an achievement test with open-ended items. In the Mead and Meade (2010) simulation study, it was concluded that test construction using either CTT or IRT procedures lead to empirically similar exams. Thus, other properties, such as ease of use and the comprehensiveness of the reported results, should be prioritized in making decisions on whether to use CTT or MFRM in developing an achievement test containing open-ended questions. For instance, the fact that CTT is a more frequently used theory and that the analyses for this theory can easily be performed by using Microsoft Excel can cause researchers/practitioners to consider CTT as a more practical way to develop an open-ended test. In spite of those positive characteristics, MFRM also has advantages compared to CTT. For example, synchronically calculating the validity and reliability of measurements, item difficulties, individuals' ability levels, and raters' severity and leniency; comparing all the facets used by putting them on the same logit; and analysis outcomes having test information function, category statistics and unexpected responses – all of which have no counterparts in CTT – make MFRM a more preferable model to CTT, even though it yields similar results in terms of item difficulty indices.

Recommendations

A review of the literature showed that the item parameters estimated in CTT and IRT were mostly restricted to using multiple-choice tests. This current study, however, compared difficulty indices of open-ended items calculated on the basis of CTT with those calculated on the basis of MFRM-model based on IRT. It is thought that the study will contribute to the literature in this respect. Nevertheless, this study – as all scientific studies – also has some limitations. The restrictions, which also imply suggestions for further research, are related to the external validity of the study. Studies comparing different theories can contain effects stemming from the data set (Engelhard, 1984) and limit the generalizability of the conclusions reached in the study. Therefore, conducting similar studies with different data sets is important in raising generalizability of the conclusions reached.

References

Bickman, L., & Rog, D.J. (2009). *Applied social research methods*. Los Angeles: Sage.

Connaway, L.S., & Powell, R.R. (2010). *Basic research methods for librarians*. Santa Barbara, CA: Libraries Unlimited.

De Ayala, R.J. (2009). *The theory and practice of item response theory*. New York: The Guilford.

DeVellis, R.F. (2003). *Scale development: Theory and applications*. Thousand Oaks, CA: Sage.

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

Engelhard, G. (1984). Thorndike, Thurstone, and Rasch: A comparison of their methods of scaling psychological and educational tests. *Applied Psychological Measurement*, *8*(1), 21-38. http://dx.doi.org/10.1177/014662168400800104

Furr, R.M., & Bacharach, V.R. (2008). *Psychometrics: An Introduction*. Thousand Oaks, CA: Sage.

Haiyang, S. (2010). An application of classical test theory and many facet Rasch measurement in analyzing the reliability of an English test for non-English major graduates. *Chinese Journal of Applied Linguistics*, *33*(2), 87–102.

Hambleton, R.K. (2004). Theory, methods, and practices in testing for the 21st century. *Psicothema*, *16*(4), 696-701.

Hambleton, R.K., & Jones, R.W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice,* *12*(3), 38-47. http://dx.doi.org/10.1111/j.1745-3992.1993.tb00543.x

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: SAGE Publications, Inc.

Harvey, R.J., & Hammer, A.L. (1999). Item response theory. *The Counseling Psychologist*, *27*(3), 353-383. http://dx.doi.org/10.1177/0011000099273004

Huang, T.W., Guo, G.J., Loadman, W., & Low, F.M. (2014). Rating score data analysis by classical test theory and many-facet Rasch model. *Psychology Research*, *4*(3), 222-231.

İlhan, M. (2016a). A comparison of the results of many-facet Rasch analyses based on crossed and judge pair designs. *Educational Sciences: Theory & Practice*, *16*(2), 579-601. http://dx.doi.org/10.12738/estp.2016.2.0390

İlhan, M. (2016b). A comparison of the ability estimations of classical test theory and the many facet Rasch model in measurements with open-ended questions. *Hacettepe University Journal of Education,* *31*(2), 348-358. http://dx.doi.org/10.16986/HUJE.2016015182

Kadir, K.A. (2013). Examining factors affecting language performance: A comparison of three measurement approaches. *Pertanika Journal of Social Sciences & Humanities*, *21*(3), 1149-1162.

Kelecioğlu, H. (2001). The relationship between b and a parameters in latent trait theory and p and r statistics in classical test theory. *Hacettepe University Journal of Education*, *20*, 104–110.

Kline, T.J.B. (2005). Classical test theory: Assumptions, equations, limitations, and item analyses. In T.J.B. Kline (Ed.), *Psychological testing: A practical approach to design and evaluation* (pp. 91–105). Thousand Oaks, CA: Sage. http://dx.doi.org/10.4135/9781483385693.n5

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

Linacre, J. M. (2012). *Many-facet Rasch measurement: Facets tutorial2*. Retrieved December 19, 2018, from http://www.winsteps.com/a/ftutorial2.pdf

Linacre, J.M. (2014). *A user's guide to FACETS Rasch-model computer programs*. Retrieved December 18, 2018, from http://www.winsteps.com/a/facets-manual.pdf

Lynch, B.K., & McNamara, T.F. (1998). Using G-theory and Many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing*, *15*(2), 158–180. http://dx.doi.org/10.1191/026553298674579408

Lunz, M.E., & Stahl, J.A. (1990, April). *Severity of grading across time periods*. Paper presented at the Annual Meeting of the American Educational Research Association, Boston. Retrieved January 11, 2018, from https://files.eric.ed.gov/fulltext/ED317602.pdf

MacMillan, P.D. (2000). Classical, generalizability and multifaceted Rasch detection of interrater variability in large sparse data sets. *The Journal of Experimental Education*, *68*(2), 167-190. http://dx.doi.org/10.1080/00220970009598501

Mead, A.D., & Meade, A.W. (2010, April). *Item selection using CTT and IRT with unrepresentative samples*. Paper presented at the twenty-fifth annual meeting of the Society for Industrial and Organizational Psychology in Atlanta, GA. Retrieved December 22, 2018, from http://mypages.iit.edu/~mead/Mead_and_Meade-v10.pdf

Mulqueen C., Baker D., & Dismukes, R.K. (2000, April). *Using multifacet Rasch analysis to examine the effectiveness of rater training*. Presented at the 15th Annual Conference for the Society for Industrial and Organizational Psychology (SIOP). New Orleans. Retrieved December 14, 2018, from https://www.air.org/sites/default/files/downloads/report/multifacet_rasch_0.pdf

Prieto, G., & Nieto, E. (2014). Analysis of rater severity on written expression exam using many faceted Rasch measurement. *Psicológica*, *35*, 385-397.

Reckase, M.D. (2009). *Multidimensional item response theory*. New York, NY: Springer.

Schumacker, R.E. (2010). *Classical test analysis*. Applied measurement associates LLC. Retrieved December 12, 2018, from http://appliedmeasurementassociates.com/ama/assets/File/CLASSICAL_TEST_ANALYSIS.pdf

Shultz, K.S., & Whitney, D.J. (2005). *Measurement theory in action: Case studies and exercises*. Thousand Oaks, CA: Sage.

Sudweeks, R.R., Reeve, S., & Bradshaw, W.S. (2005). A comparison of generalizability theory and many-facet Rasch measurement in an analysis of college sophomore writing. *Assessing Writing, 9*(3), 239–261. https://doi.org/10.1016/j.asw.2004.11.001

Takano, Y., Tsunoda, S., & Muraki, M. (2015). Mathematical optimization models for nonparametric ıtem response theory. *Information Science and Applied Mathematics*, *23,* 1-16.

van der Linden, W.J. (2005). Item response theory. In K. Kempf-Leonard (Ed.), *Encyclopedia of social measurement* (Vol. 2, pp. 379-387). San Diego, CA: Academic Press.

Wright, B.D. & Linacre, J.M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8*, 370. Retrieved November 23, 2018 from http://www.rasch.org/rmt/rmt83b.htm

Açık Uçlu Maddelerde Klasik Test Kuramı ile Çok Yüzeyli Rasch Modeline göre Hesaplanan Güçlük İndekslerinin Karşılaştırılması

Atıf:

Ilhan, M., & Guler, N. (2018). A comparison of difficulty ındices calculated for open-ended ıtems according to classical test theory and many facet Rasch model. *Eurasian Journal of Educational Research*, 75, 99-114, DOI: 10.14689/ejer.2018.75.6

Özet

Problem Durumu: Klasik test kuramı (KTK) ve çok yüzeyli Rasch modeli (ÇYRM) arasındaki kuramsal farklılıklar alanyazında geniş bir yer tutmasına rağmen bu iki kuramı ampirik açıdan karşılaştıran araştırmaların oldukça sınırlı olduğu görülmektedir. KTK ve ÇYRM'nin karşılaştırılmasına yönelik çalışmalarda üzerinde en fazla durulan konu iki kurama göre hesaplanan güvenirlik değerlerinin ne derece tutarlı olduğudur. Daha yakın zamanda yapılan araştırmalarda ise iki kuramının yetenek kestirimleri ile madde güçlük parametreleri açısından karşılaştırıldığı anlaşılmaktadır. KTK ve ÇYRM'de rapor edilen güvenirlik değerlerinin

karşılaştırıldığı araştırmalar incelendiğinde, bu çalışmaların kullanılan desen açısından farklılık gösterdiği saptanmıştır. Çalışmaların bir kısmında açık uçlu maddelere verilen öğrenci cevaplarının tamamının aynı puanlayıcı grubu tarafından değerlendirildiği çapraz bir desen kullanılmıştır. Bazılarında ise değerlendirme sürecinde birbirinden farklı puanlayıcı gruplarının görev aldığı yuvalanmış bir desen tercih edilmiştir. Dolayısıyla, konu ile ilgili alanyazındaki mevcut çalışmalar iki kurama göre hesaplanan güvenirlik değerlerinin ne derece tutarlı olduğuna ilişkin kapsamlı bir bilgi sunabilmektedir. Ancak aynı şeyi KTK ile ÇYRM'de hesaplanan madde güçlük indekslerinin karşılaştırılmasına yönelik araştırmalar için söylemek güçtür. Çünkü alanyazında KTK ve ÇYRM'de hesaplanan madde güçlüklerinin karşılaştırıldığı yalnızca bir araştırmaya rastlanmış ve bu çalışmada yuvalanmış bir desenin kullanıldığı belirlenmiştir. İki kurama göre hesaplanan madde güçlüklerinin çapraz bir deseninin kullanıldığı ölçme koşulları altında karşılaştırıldığı bir çalışmaya ise alanyazında rastlanmamıştır. Ayrıca alanyazındaki sözü edilen araştırmada, KTK'ya dayalı madde güçlükleri %25'lik alt ve üst gruba ait ölçümler esas alınarak kestirilirken; ÇYRM'ye ilişkin madde güçlük kestiriminde tüm bireylere ait ölçümler kullanılmıştır. Böylesi bir farkın iki kurama göre hesaplanan madde güçlüklerinin karşılaştırıldığı bir çalışma için önemli olabileceği düşünülmektedir. Bu anlamda, ölçme koşulları açısından alanyazındaki bahsi geçen araştırmadan farklılık gösteren bir çalışma ile KTK ve ÇYRM'de hesaplanan madde güçlüklerinin karşılaştırılması önemli görülmektedir.

Araştırmanın Amacı: Bu araştırmada, açık uçlu maddelerde klasik test kuramı (KTK) ile çok yüzeyli Rasch modeline (ÇYRM) göre hesaplanan güçlük indekslerinin karşılaştırılması amaçlanmıştır.

Araştırmanın Yöntemi: Araştırmanın verileri, sekizinci sınıfa devam eden 375 öğrencinin açık uçlu 10 maddeye verdiği cevabın üç öğretmen tarafından puanlanmasıyla elde edilmiştir. Puanlamalarda dörtlü derecelemeye sahip bütüncül bir rubrik kullanılmıştır. KTK'ya dayalı madde güçlüklerinin hesaplanmasındaki ilk adım, öğrencilerin her bir maddeye verdikleri cevaplar için farklı puanlayıcılar tarafından atanan puanların ortalamasının alınması olmuştur. İkinci adımda tüm maddeler için ayrı ayrı olmak üzere, öğrencilerin maddelerden aldıkların puanların toplamı çalışmadaki katılımcı sayısına bölünmüş ve bu şekilde maddelere ilişkin puan ortalamaları hesaplanmıştır. Daha sonra her bir madde için, ilgili maddeden alınan puanların ortalaması ile maddeden alınabilecek en düşük puan arasındaki fark bulunmuştur. Bulunan bu farkın madde puan ranjına bölünmesiyle KTK dayalı madde güçlük parametrelerine ulaşılmıştır. Madde güçlüklerinin KTK'ya göre hesaplanmasında Microsoft Excel'den yararlanılmıştır. KTK'ya ilişkin analizlerin ardından ÇYRM'ye yönelik analizlere geçilmiştir. Bu kapsamda, FACETS paket programı kullanılarak puanlayıcı, madde ve öğrenci şeklinde üç yüzeyli bir desen ile Rasch analizi gerçekleştirilmiştir. Analiz çıktılarında, madde yüzeyine ilişkin ölçüm raporları incelenerek ÇYRM'ye dayalı madde güçlük parametreleri elde edilmiştir. Madde güçlük indekslerinin KTK ve ÇYRM'ye göre hesaplanmasını takiben, iki kurama göre kestirilen güçlük değerleri arasındaki tutarlılığa bakılmıştır.

Araştırmanın Bulguları: Araştırmadan elde edilen bulgular, iki kurama göre kestirilen güçlük indeksleri arasında yüksek bir tutarlılık olduğunu göstermiştir. Maddeler güçlük düzeyleri açısından bir sıralamaya tabi tutulduğunda KTK ile ÇYRM'de

ulaşılan sıralamaların özdeş olduğu saptanmış ve iki kurama göre kestirilen güçlük indeksleri arasında pozitif yönde, güçlü ve anlamlı bir korelasyonun (r=.999, p<.001) bulunduğu belirlenmiştir. Her iki kurama göre de başarı testindeki 10 maddenin kolaydan zora doğru; M2, M3, M1, M4, M7, M6=M8, M10, M5 ve M9 şeklinde sıralandığı sonucuna ulaşılmıştır.

Araştırmanın Sonuç ve Önerileri: Araştırma sonuçlarından hareketle, açık uçlu maddeler içeren bir başarı testi geliştirme sürecinde, madde güçlüklerinin KTK veya ÇYRM'ye göre kestirilmiş olmasının teste alınacak ya da test dışında tutulacak maddeler ile ilgili bir farklılık yaratmayacağı söylenebilir. Dolayısıyla açık uçlu maddelerin bulunduğu bir başarı testi geliştirirken KTK ile ÇYRM'den hangisinin tercih edilmesi gerektiğine dair verilecek kararlarda kullanım kolaylığı ve rapor edilen sonuçların ne derece ayrıntılı olduğu gibi kuramlara ilişkin diğer özelliklerin ön plana çıkacağı düşünülmektedir. Örneğin, KTK'nın birçok kişinin daha aşina olduğu bir kuram olması ve bu kurama ilişkin madde analizlerinin Microsoft Excel'de kolaylıkla gerçekleştirilebilmesi araştırmacıların/uygulayıcıların açık uçlu test geliştirme sürecinde KTK'yı daha pratik bir yol olarak görmesine sebep olabilir. KTK'yı ÇYRM'ye göre daha kullanışlı hale getiren bu özelliklerine karşın ÇYRM'nin de KTK'ya kıyasla daha avatanjlı olduğu bazı yönleri bulunmaktadır. Ölçümlerin geçerliği ile güvenirliğinin, madde güçlüklerinin, bireylerinin yetenek düzeylerinin ve puanlayıcıların katılık/cömertliklerinin eş zamanlı olarak hesaplanması, analizde işlem gören tüm yüzeylerin ortak bir metrik (logit) üzerine yerleştirilerek birbiriyle karşılaştırılabilmesi ve analiz çıktıları arasında KTK'da karşılığı olmayan test bilgi fonksiyonunun, kategori istatistiklerinin ve beklenmedik yanıtların yer alması madde güçlük indeksleri açısından benzer sonuçlar üretmesine rağmen ÇYRM'yi KTK'ya göre daha tercih edilebilir bir model haline getirebilecek özelliklerdir.

Anahtar Kelimeler: Açık uçlu maddeler, madde güçlük indeksi, klasik test kuramı, çok yüzeyli Rasch modeli.