Available Online at http://iassr.org/journal
2013 (c) **EJRE** published by
International Association of Social Science Research - IASSR

ISSN: 2147-6284
European Journal of Research on Education, 2014, 2(1), 1-6

European Journal of Research on Education

Comparison of classical test theory and item response theory in terms of item parameters[1]

Neşe Güler [a] *, Gülden Kaya Uyanık [b], Gülşen Taşdelen Teker [c]

[a]*Assist. Prof., Sakarya University, Faculty of Education, Hendek, Sakarya, Turkey*
[b]*Res. Ass., Sakarya University, Faculty of Education, Hendek, Sakarya, Turkey*
[c]*Lecturer, Sakarya University, Faculty of Education, Hendek, Sakarya, Turkey*

Abstract

There are two currently popular statistical frameworks for addressing measurement problems - namely Classical Test Theory (CTT) and Item Response Theory (IRT). This study is aimed empirically to examine the similarities and differences in the parameters estimated using these two frameworks. For this purpose, firstly the model assumptions and the goodness of fit analyses were checked for IRT. The results suggested that the most appropriate model which fit the data was achieved by three parameter logistic IRT model. Then item difficulty and item discrimination parameters were calculated in terms of both IRT and CTT. Lastly parameters which were obtained by CTT and IRT analyses were compared. In the study, the answers given to the 25-item Turkish test by randomly chosen 1250 students from the group of 5989 students who had taken the high schools entrance exam (HSEE) in 2003 were used. In consequence, it was found that the highest correlations were available between CTT and 1-parameter IRT model (0.99) in terms of item difficulty parameters, and between CTT and 2-parameter IRT model (0.96) in terms of item discrimination parameters. Besides, although the 3-parameter model was identified as the most congruous one in terms of model-data fit, the lowest level of correlation was found between the 3-parameter model and CTT. In the light of these findings, it may be said that there is not much difference between using 1 or 2-parameter IRT model and CTT. However, in cases where the probability of guessing is high, there is a significant difference between 3-parameter model and CTT.

Keywords: Item response theory, classical test theory, item difficulty, item discrmination , guessing

1. Introduction

The basic aim of test development is to construct a test of desired quality by choosing the appropriate items, no matter what type of tool is used. There are two main theories called classical test theory (CTT) and item response theory (IRT) can be used in test development. CTT was used over the majority of 20th century (Demirtaşlı, 2002; Traub, 1997) and is still used in test development (Bechger et al., 2003). The main advantage of CTT are its relatively weak theoretical assumptions make it easy to apply in many testing situations (Hambleton & Jones, 1993). Although CTT's major focus is on test-level information, item statistics (i.e., item difficulty and item discrimination) are also an important part of the CTT model (Fan, 1998).

The CTT served very well for years in test development but it has some restrictions in use. The majority of the restrictions are coming from group dependent items and test statistics (Hambleton and Swaminathan, 1985). In other words, in CTT the person statistic (i.e., observed score) is item dependent, and the item statistics such as item

[1] Part of this study was presented in European Conference on Social Science Research in 19-21 June 2013 in Marmara University, Istanbul, Turkey.

* *E-mail address*: gnguler@gmail.com

difficulty and item discrimination are examinee dependent which poses some theoretical difficulties in CTT's application in some measurement situations such as test equating, computerized adaptive testing, identification of biased items, linking and building item banks (Önder, 2007; Demirtaşlı, 2002; Fan, 1998). On the other hand, limitations of CTT are overcome by the use of IRT which has witnessed an exponential growth in recent decades.

As its name indicates, IRT primarily focuses on the item-level information in contrast to the CTT which primarily focuses on test-level information (Fan, 1998). IRT is a measurement approach that relates the probability of a particular response on an item to overall examinee ability (Camilli & Shepard, 1994). Therefore, in IRT ability parameters estimated are not test dependent and item statistics estimated are not group dependent (Hambleton and Swaminathan, 1985). IRT presents a mathematical model on how examinees with different ability levels answered the items (Crocker and Algina, 1986). Ability derived from tests which are developed in accordance with this theory can be obtained independently from the group (Kelecioğlu, 2001). In other words, when an acceptable compatibility is attained between the set of data and the model chosen, IRT models ensure that we obtain constant item parameters and ability predictions (Çelen, 2008).

The scores received by individuals vary according to the difficulty level of the test in CTT (Lord and Novic, 1968). The IRT, on the other hand, claims to be able to do sample free constant parameter predictions (Hambleton, Swaminathan, & Rogers, 1991). Although the only coefficient obtained in the CTT means that reliability does not change at different levels of ability, on examining the reliability coefficients that are calculated with repetitive measurements, it is found that they are higher for individuals with high levels of the property that is measured. This shows that the measurement tool cannot have the same level of reliability for individuals with differing levels of ability (Nartgün, 2002). Several researchers presented some of the advantages of IRT over CTT (Camilli & Shepard, 1994; Fan, 1998; Hambleton et al., 1991; Özdemir, 2004). However, it is not possible to have these advantages unless model data fit is achieved for IRT (Fan, 1998; Hambleton et al., 1991).

IRT is said to have two basic assumptions which are unidimensionality and local item independence (Hambleton and Swaminathan, 1985). The assumption of unidimensionality assumes that items of a test measure only one ability. Since individuals' cognitive and personal characteristics, their levels of motivation, test anxiety, and behaviours such as running into contradictions as to the correct answers and discussing about it influence test performance and cannot often be controlled; it is not always possible to meet this assumption. We can talk about the unidimensionality of a test only when we consider that there is just one dominant ability in it (Hambleton et al., 1991). The other important assumption, local item independence, can be defined as the independence of the probability of being answered correctly of any two items in a tool of measurement, depending on the individual's level of ability (Embretson & Reise, 2000; Hambleton & Swaminathan, 1985; Yen, 1993).

The IRT puts forward three different models as 1-, 2-, and 3-parameter models. A one-parameter model sets up relations between the parameter of item difficulty (bi) and an individual's level of ability. In a two-parameter model discrimination parameter (ai) is added to the parameter of item difficulty. And in the third model, the parameter of guessing (ci) is added (Hambleton & Swaminathan, 1985).

Taking the above mentioned differences into consideration, a comparison was made between CTT and IRT in terms of item parameters in this article. First the model assumptions then goodness of fit analyses for IRT were checked. Then parameters - item difficulty and item discrimination- were calculated with both IRT and CTT. Lastly parameters which were obtained with CTT and IRT analyses were compared.

In Turkey there are lots of nationwide exams that are very important for examinees. Therefore parameters of items are also considerable. Just because of this in this study 2003 High School Entrance Exam (HSEE) - Turkish subtest- was used. Its parameters were estimated in the way of both IRT and CTT and compared. The main purpose of this study is to investigate the model-data fit of Turkish subtest of HSEE 2003 for the IRT models and to determine the correlation level between parameters estimated by CTT and IRT.

2. Method

2.1. Study Group

The study group was composed of 1250 students randomly chosen from the group of 5989 students who had taken the Turkish subtest of HSEE 2003.

2.2. Instruments

The data containing the answers given by the 1250 students to the 25-item Turkish subtest in the HSEE in 2003 were used for our purposes.

2.3. Analysis of the Data

In order to meet the assumption of unidimensionality of a structure measured -the most important supposition for the implementation of the IRT, factor analysis was performed on the SPSS. The item parameters were derived with the ITEMAN Programme in the analyses of the CTT while the MULTILOG Programme was used for the analyses of IRT. Also, the compatibility of the data with 1-, 2- and 3-parameter models was examined. To do this, the 2-loglikelihood differences were found and then it was tested with chi-square value. Finally, the Excel Programme was used in calculating the correlations between item parameters estimated by CTT and IRT.

3. Results

3.1. Checking Model Assumptions

In any application of the IRT model, it is important to assess to what extent the IRT model assumptions are valid for the given data and how well the testing data fit the IRT model selected for use in that particular situation. The violation of IRT model assumptions, or misfit between the IRT model used and the testing data, may lead to erroneous or unstable IRT model parameter estimates (Fan, 1998).

Firstly, in order to analyse unidimensionality -the most important assumption common for all IRT models- factor analysis was performed on the SPSS for the 25-item Turkish subtest included in the HSEE which was administered in 2003. Eigenvalues and scree plot obtained was investigated in order to determine whether there was a *dominant* factor. The two items (8 and 13) which spoiled unidimensionality were removed from the test, and following the factor analysis for the second time, unidimensionality was achieved. Thus, in the following stages, the analyses were done with 23 items.

If the items are not statistically independent for constant level of ability, the test scores of some individuals are expected to be higher than those of other individuals. In consequence, it will be possible to account for the individual's test performance with more than one ability. The violation of this assumption also means the violation of unidimensionality assumption. Hence, explanation of the relations in a set of items should be connected to only one ability (Hambleton and Swaminathan, 1985). Thus, the 23 items in the HSEE Turkish subtest achieving the assumption of unidimensionality were also regarded as achieving the assumption of local independence.

3.2. Analysis of Model-Data Fit

Model-data fit test was performed on the MULTILOG programme for the data of responses given by 1250 students to the 23 items. The -2 Log Likelihood values for the 1-, 2-, and 3-parameter logistic models were derived, as is shown in Table 1.

Table 1.The -2 Log Likelihood values of 1-, 2-, and 3-parameter logistic models

Model	**-2 Log Likelihood Values**
1PLM	16712
2PLM	16489
3PLM	16308

In the Chi-square test, the table value was found as 38.88 in the 23 degrees of freedom at 0.05 significance level. Thus, considering the -2 Log likelihood values found for the 1-, 2-, and 3-parameter logistic models, the following results were obtained:

- Because 1PLM - 2PLM= 16712- 16489= 223 > 38.88, the 2-parameter model is more significant than the 1-parameter model.
- Because 2PLM - 3PLM= 16489- 16308= 181 > 38.88 , the 3-parameter model is more significant than the 2-parameter model.

In consequence, it was concluded that the model with the most compatibility to the data was the 3-parameter IRT model.

3.3. Prediction of Item Parameters and Analysis of the Correlations

The item parameters for 23 items were calculated with the MULTILOG programme for IRT and with the ITEMAN programme for CTT. The obtained difficulty parameter (b) for the 1-parameter IRT model, the difficulty (b) and discrimination (a) parameters for the 2- and 3-parameter IRT models, and difficulty (p) and discrimination (r) parameters for CTT are given in Table 2.

Table 2. Item parameters calculated by IRT and CTT

Item	IRT					CTT	
	1PLM	2PLM		3PLM			
	b	a	b	a	b	r	p
1	0.19	0.94	0.19	1.45	0.79	0.55	0.46
2	-0.71	1.65	-0.53	1.44	-0.07	0.71	0.64
3	*-2.07*	0.99	*-2.02*	0.56	*-2.1*	0.51	*0.85*
4	0.71	0.63	1	0.39	0.96	0.45	0.36
5	*0.99*	0.68	*1.3*	1.22	*1.42*	0.46	*0.31*
6	-0.43	1.11	-0.4	0.77	-0.1	0.6	0.58
7	-0.57	1.47	-0.45	1.14	-0.09	0.68	0.61
8	0.21	0.98	0.2	1	0.67	0.57	0.45
9	-1.51	0.85	-1.67	0.48	-1.72	0.49	0.77
10	0.41	1.3	0.32	1.36	0.63	0.67	0.42
11	0.06	1.07	0.05	0.7	0.17	0.6	0.49
12	-1.16	0.81	-1.32	0.49	-1.29	0.48	0.72
13	0.32	0.83	0.36	1.3	0.96	0.52	0.43
14	0.72	0.67	0.96	0.58	1.23	0.46	0.36
15	-0.22	0.78	-0.25	0.47	-0.24	0.5	0.54
16	-1.23	1.52	-0.86	1.26	-0.29	0.66	0.71
17	-0.55	0.85	-0.6	0.7	0.09	0.52	0.61
18	-0.96	1.71	-0.7	1.78	-0.56	0.71	0.68

19	0.23	1.14	0.2	1.09	0.71	0.62	0.45
20	-0.58	0.98	-0.57	0.63	-0.36	0.57	0.61
21	0.65	0.95	0.66	1.33	1.01	0.56	0.37
22	-0.28	0.67	-0.36	0.97	0.86	0.46	0.55
23	0.36	0.78	0.42	1.28	1.04	0.51	0.43

In table 2 it is clear that the easiest item for 1 PLM, 2PLM, 3PLM and CTT is the third one the values of which are -2.07; -2.02; -2.01 and 0.85 respectively. Although it is the same item for all of the models, value of difficulty parameters are different. Also the most difficult item is the same for all models and its fifth one. Values of this item's difficulty parameters are 0.99 for 1PLM; 1.3 for 2PLM ; 1.42 for 3PLM and 0.31for CTT. There are also item discrimination parameters in table 2. The most distinctive item is 18th one for 2 PLM; 3PLM and CTT and its value is respectively 1.71; 1.78 and 0.71. Also the worst distinctive item is fourth one for all models and its value is 0.63 for 2PLM; 0.39 for 3PLM and 0.45 for CTT.

Since the 1-parameter IRT model assumes fixed item discrimination and no guessing for all items, the model only provides estimates for item parameter of difficulty. Therefore, only the correlation between item difficulty parameters of 1-parameter IRT model and CTT estimated. Between 2- and 3-parameter IRT models and CTT, the correlation coefficients were estimated for both the item difficulty and discrimination parameters. Moreover, since there is no estimated guessing parameter which can be estimated by 3-parameter IRT model, any correlation computed between any models. Correlations between the results of item parameters given in Table 2, which were derived with each model of the IRT according to the CTT, were calculated with the Excel programme, and are shown in Table 3.

Table 3. The correlation coefficients of item parameters between IRT and CTT

1PL IRT-CTT (b – p)	2PL IRT-CTT (a- r)	2PL IRT-CTT (b - p)	3PL IRT-CTT (a - r)	3PL IRT-CTT (b - p)
-0.99806	0.965921	-0.98464	0.525264	-0.93891

As the tabled results indicate, for the one-, 2- and 3-parameter IRT models, the relationship between CTT and IRT-based item difficulty estimates is almost perfect since the majority of the coefficients were above .93 under all conditions. Moreover, because item difficulty parameter estimates of 1-parameter IRT model were almost perfectly related to CTT-based item difficulty indexes (both original and normalized), it appears that the one-parameter model provides almost the same information as CTT with regard to item difficulty. As is clear from Table 3, correlations obtained from item difficulty parameters (b and p) are marked as also negative. The reason is evident: item difficulty index (b) moves from the smaller to the bigger and item difficulty moves from the easier to the more difficult in item response theory; whereas item difficulty index (p) moves from the smaller to the bigger but item difficulty moves from the more difficult to the easier in the classical test theory; therefore a negative correlation holds here.

Because the 1-parameter IRT model assumes fixed item discrimination for all items, no correlation between CTT discrimination and 1-parameter IRT model item discrimination which is a constant--could be computed. As can be seen in Table 3, correlations between the 2-parameter IRT model and CTT is very high (0.96). On the other hand the coefficient between 3-parameter IRT model and CTT is low compared to the former one. However, both of the coefficients were positive.

4. Conclusion

The item difficulty parameters from CTT were very comparable with those from all IRT models and especially from the 1-parameter IRT model. Compared with item difficulty parameters, the item discrimination parameters from CTT were somewhat less comparable with those from 3-parameter IRT model. Although under the majority of

the conditions, the comparability was moderately high to high, there were a few cases where the comparability was very low.

The item that item difficulty parameter had the highest correlation with the classical test theory was the 1-parameter model of the item response theory (0.99), but the item that item discrimination parameter had the highest correlation with the classical test theory was the 2-parameter model of the IRT (0.96). However, the lowest correlation was found between the 3-parameter model -the most appropriate model in terms of data-model fit – and CTT.

All these results show that there is not much difference between using 1 or 2-parameter model according to IRT and using CTT on data obtained from a sufficient number of samples. Moreover, there are no significant differences between using a 2-parameter model according to the IRT and using the CTT in cases when the IRT discrimination parameter is important and needs calculation. Yet, when there is accidental success, or chance guessing, and when the item parameters are calculated considering this, then there is a significant difference between using the 3-parameter model according to the IRT and using the CTT.

Of course, the present empirical study, like many other research studies, had its share of limitations that may potentially undermine the validity of its findings. First of all, the characteristics of the test items used in the study may be somewhat out of date. However, since the aim of the study is just to compare the theories, the data only used as a research instrument.

The second shortcoming of the investigation is the somewhat limited item pool used in the study. Although the examinee pool is large enough in the sense that a variety of different samples can be drawn from it, the same cannot be said about the item pool which consists of just 25 items. Ideally, the test item pool should be larger and more diverse in terms of item characteristics so that items can be sampled from the pool to study the behaviors of CTT and IRT item statistics under different conditions of item characteristics (Fan, 1998).

References

Bechger, T.M., Maris, G., Verstralen, H.H.F.M., & Beguin, A.A. (2003). Using classical test theory in combination with item response theory. *Applied Psychological Measurement*, 27 (5), 319-334.

Camilli, G., & Shepard, L.A. (1994). *Methods for identifying biased test items* (vol. 4). Thousand Oaks, CA: Sage.

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

Çelen, Ü. (2008). Klasik Test Kuramı ve Madde Tepki Kuramı Yöntemleriyle Geliştirilen İki Testin Geçerlilik ve Güvenirliğinin Karşılaştırılması. *İlköğretim Online*, 758-768.

Demirtaşlı, N.Ç. (2002). A study of raven standard progressive matrices test's item measures under classical and item response models: An empirical comparison. *Ankara Üniversitesi Eğitim Bilimleri Fakültesi Dergisi*, 35 (2), 71-79.

Embretson, S. E., & Reise, S. P. (2000). *Item Response Theory for Psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person parameters. Educational and Psychological Measurement, 58, 357–381.

Hambleton, R. K., & Swaminathan, H. (1985). *Item Response Theory: Principles and Applications*. USA: Kluwer-Nijhoff Publishing.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. California: SAGE Publications.

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), 253-262.

Kelecioğlu, H. (2001). Örtük Özellikler Teorisindeki b ve a Parametreleri ile Klasik Test Teorisindeki p ve r İstatistikleri Arasındaki İlişki. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, 104-110.

Lord, F. M., & Novick, M. R. (1968). *Statistical Theories of Mental Scores*. Addison-Wesley Publishing Company.

Nartgün, Z. (2002). *Aynı Tutumu Ölçmeye Yönelik Likert Tipi Ölçek ile Metrik Ölçeğin Madde ve Ölçek Özelliklerinin Klasik Test Kuramı ve Örtük Özellikler Kuramına Göre İncelenmesi*. Yayınlanmamış Yüksek Lisans Tezi, Ankara: Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü.

Önder, İ. (2007). Model Veri Uyumunun Araştırılması. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, 32, 210-220.

Özdemir, D. (2004). Çoktan seçmeli testlerin klasik test teorisi ve örtük özellikler teorisine gore hesaplanan psikometrik özelliklerinin iki kategorili ve ağırlıklandırılmış puanlanması yönünden karşılaştırılması. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, 26, 117-123.

Traub, R.E. (1997). Classical test theory in historical perspective. *Educational Measurement: Issues and practice*, 8-14.

Yen, W. M. (1993). Scaling Performance Assessment: Strategies for Managing Local Item Dependence. *Journal of Educational Measurement, 30*, 187-213.