academicJournals

Vol. 9(17), pp. 642-649, 10 September, 2014
DOI: 10.5897/ERR2014.1729
Articles Number: 4F689D746945
ISSN 1990-3839

http://www.academicjournals.org/ERR

Educational Research and Reviews

Full Length Research Paper

Assessing model data fit of unidimensional item response theory models in simulated data

İbrahim Alper KÖSE

Abant İzzet Baysal University, Turkey.

Received 27 January, 2014; Accepted 25 July, 2014

The pupose of this paper is to give an example of how to assess the model-data fit of unidimensinal IRT models in simulated data. Also, the present research aims to explain the importance of fit and the consequences of misfit by using simulated data sets. Responses of 1000 examinees to a dichotomously scoring 20 item test were simulated with 25 replications. Also, data were simulated to fit the 2-PL model. 4-step procedure has been used for model-data fit and BILOG was used as software. Results were discussed in the frame of the literature.

Key words: Item response theory, unidimensionalty, model-data fit, invariance.

INTRODUCTION

Researchers in the field of educational assessment are continually developing new approaches to improve the efficiency of assessments. They are often concerned with methodologies that can extract the most useful and accurate information from students' responses to test items (Wu and Adams, 2006). With the help of improved mathematical models and computer technologies, new theories have been developing in the field of educational and psychological assessment.

In the measurement history, the leading theory to explain latent trait underlying exaimee's test performance is Classical Test Theory (CTT). CTT is a simple model which states that the observed score on a test is the sum of the true score and measurement error. CTT is based on weak assumptions, that is, the assumptions can be met easily by most data sets, and therefore, the models can and have been applied to a wide variety of test development and test score analysis problems (Hambleton and Swaminathan, 1989). Group dependency of test and item characteristics, providing information about examinee performance from whole test and having no information about examinee's performance on a single test item are crucial shortcomings of CTT (Hambleton et al., 1991). In psychometrics, CTT was the dominant statistical approach to testing data until Lord and Novick (1968) placed it in context with several other statistical theories of mental test scores, notably item response theory (IRT) (Sijtsma and Junker, 2006; Seungho-Yang, 2007).

One of the most imortant improvements the last century is IRT, also known as latent trait theory, in psychological measurement. IRT is a modern test theory which explains examinee's ability level by using responses to test items with strong assumptions against CTT's weak assumptions

E-mail: i.alper.kose@gmail.com. Tel: +905052384914.

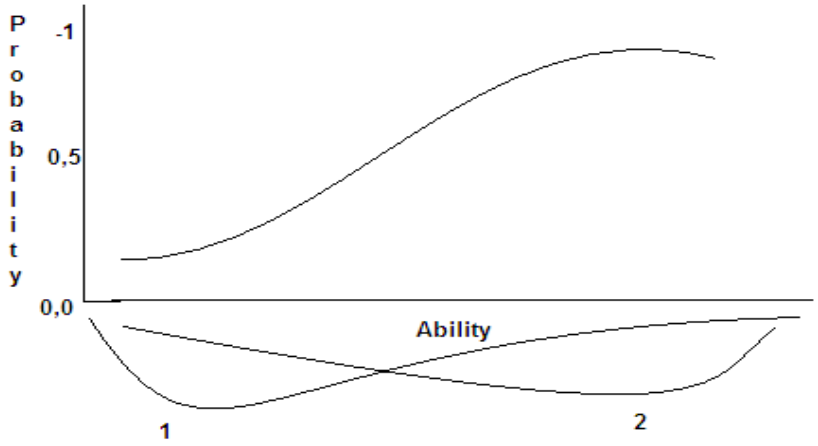

Figure 1. An item characteristic curve and Distirbutions of ability for two groups of examinees

with mathematical models (Bobcock, 2009). According to Embretson and Reise (2000), IRT is a "modern" test theory utilizing a set of propositions or mathematical models related to individuals' responses to items, providing a probabilistic way of linking observable data to theoretical constructs, with the ability to statistically adjust scores for properties of test items such as difficulty, discriminating power, and liability to guessing.

The origin of IRT can be traced to the efforts of Thurstone (1925; cited in Bock et al., 1997) and others to scale the tasks that made up the Binet test of intelligence. The objective of their work was the criterion of quantitative scales, not unlike those of physical science, on which the strength of a trait could be expressed. From this beginning, developments have continued to focus on the measurement of psychological constructs assumed to underlie persisting individual differences in behaviour (Bock et al., 1997).

Item response theory (IRT) makes a sharp distiction between the observable scores of a respondent on a set of items and the scale on which the unobservable psychological construct is measured. The construct can be a personality trait, a cognitive ability, an educational achievement, an attitude, or an opinion, in short, a latent trait. Typical of IRT measurement is that interest almost always lies with the respondent's position on the latent trait scale to be denoted theta (θ) (Sijtsma and Hemker, 2000).

IRT has a number of advantages over CTT methods. The most important advantages of IRT are placing the ability of the respondent and the difficulty of the item on the same measurement scale (Spencer, 2004). Additionally, the estimated item parameters are invariant with regard to who is sampled from the population, the estimated proficiency level remains constant regardless of which items are administrated and also IRT can estimate examinee ability with more precision of measurement and less measurement errors (Lee, 2007). CTT statsitics such as item difficulty, item discrimination and reliability are contingent on the sample of respondents to whom the questions were administered. In addition, CTT yields only a single estimate of reliability and corresponding standard error of measurement, whereas IRT models measure scale precision across the underlying latent variable being measured by the instrument (Cooke and Michie, 1997).

The property of invariance of item and ability parameters is the cornerstone of IRT and its major distinction from classical test theory. This property implies that the parameters that characterize an item do not depend on the ability distribution of the examinees and the parameter that characterizes an examinee does not depend on the set of test items.

When an IRT moel fits the test data of interest, several desirable features are obtained. Ability estimates obtained from different sets of items will be the same (except for measurement errors). In IRT, item and ability parameters are said to be invariant. Figure 1 shows distributions of ability for two groups of examinees. Examinees of the same ability have the same probability of giving a correct response to the item, regardless of wheather they are from Group 1 or Group 2 (Hambleton et al., 1991). In other words, if you pick different samples and estimate the item characteristic curves (ICC), you should get the similar values of a, b and c, that is you get same ICC (Drasgow, 1982). The property of invariance is only present when the IRT model fits the test data, and when model parameters are estimated properly (Hambleton et al., 1991). The property of invariance or item free measurement and sample free measurement allows for generalization beyond the specific test (Kreiter, 1993). The inavriant property of IRT makes it possible to solve

problems in measurement and testing that are difficult to solve in CTT, namely test equating, item banking, item bias, and the use of computer adaptive testing-CAT (Hambleton et al., 1991).

A variety of IRT models have been deveopled for dichotomous and polytomous data. The most commonly used models for dichotomous items are the logistic models (e.g. two parameter model and three parameter model). Samejima's graded response model is applied to polytomous data, where options are ordered along continuum (e.g. likert scales). The early IRT applications involved primariliy unidimensional IRT models. However, several multidimensional IRT models have been developed. These models usually are direct extensions of unidimensional models (Liu, 2007).

Before one uses any statistical model for any purpose, it is obviously necessary to insure the model chosen is appropraite for the data (Kingston and Stocking, 1986). Statistical models, such as item response theory, are based on assumptions. There are three assumptions of the most commonly used IRT models: undimensionality, local independence and particular shape of the item response function.

A common assumption of IRT models is that only one ability is measured by a set of items in a test. What is required for the unidimensionality assumption to be met adequately by a set of test data is the presence of a "dominant" component or factor that influences test perfromance. This dominant component or factor is referred to as the ability measured by the test (Hambleton et al., 1991). Any violation of this assumption would result in inadequacy of the model in describing the data and hence unreliable estimation of the examinee's ability. Therefore, the correct specification of the number of the latent dimensions is directly tied to the construct validity of the test (Sheng, 2005).

Local independence, which is the second assumption, means that the response of an individual to an item from the test is not influenced by his or her responses to the other items that the same test or by other traits that theta (Sijtsma and Hemker, 2000). This assumption is necessary in IRT to assure the independence of items and hence their multiplicative property when ascertaining likely abilities for response patterns (Pomplun, 1988). According to McDonald (1982), a theory of unidi-mensionality should be based on the assumption of local independence.

The third assumption of IRT is item characterictic curve (ICC). An ICC is defined completely when its general mathematical form is specified and when the parameters of the curve are chosen. In current popular IRT models the general form is a cumulative logistic ogive. The current popular models differ on the number of parameters for the curve (Pomplun, 1988).

Once an IRT model has been applied to a set of data, its approprietness should be investigated with data-model fit analysis. Otherwise, the researcher is under the risk of drawing incorrect conclusions regarding the scientific problem of interest. Substantial lack of fit should result in the replacement or extension of the model if possible (Sinharay, 2005). Traditional methods are most widely used to assess model fit; especially, the likelihood ratio chi-square goodness of fit statistics and these are provided in the most popular current software packages, such as BILOG, BILOG-MG and PARSCALE (Zhao, 2008). The most common criticism about the chi-square likelihood statistics is that they are sensitive to sample size (Hambleton and Swaminathan, 1989). When the sample size is large, the statistical test rejects just about every model since with large sample sizes (Zhao, 2008).

The assessment of data-model fit is important because the application of an IRT model can be justified only when data fit the model. The most general approach for assessing model-data fit of IRT models is to compare an observed score distribution with an expected score response distribution across dicrete ability levels for each item (Seungho-Yang, 2007; Kreiter, 1993; Dodeen, 2004).

In IRT models, various approaches were suggested for investigating model-data fit (Glass and Falcon, 2003; Hambleton, 1994; Stone, 2000; Stone and Zhang, 2003; Sinharay, 2005; Yen, 1981). These studies summarize the discrepancy between observed values and the values expected under an IRT model.

Zhao (2008) has recomended that judgments about the fit of the model to the test data be based on four steps of evidence:

1. Choosing software and initial classical analysis,
2. Checking basic assumptions of unidimensionality and local independence,
3. Assessing model data fit,
4. Checking model parameter invariance; item parameters invariance and ability parameter inavriance.

The pupose of this article is to assess the goodness of fit of unidimensinal IRT models in simulated data. In many IRT applications model data fit have not been investigated adequately. As a result, less is known about the appropriateness of particular IRT models. This study aims to give an example how to evaluate model-data fit, explain the importance of model-data fit and the consequences of misfit.

METHODS

This study was conducted based on a simulated data set. There are two major steps in the simulation; data generation and data calibration. A computer program WINGEN was used to simulate the item response data. Responses of 1000 examinees to a dichotomously scoring 20 item test were simulated with 25 replications. Also, data were simulated to fit the 2-PL model.

Data were simulated to have two normally distributed levels of item disrimination (a) and item difficulty (b). The person's ability (θ)

Table 1. Estiamted item paarmeters for simulated data.

	a	b		a	b
1	-1,925	0,115	11	1,757	-2,481
2	1,124	-0,403	12	0,690	1,399
3	0,473	0,016	13	0,359	-0,916
4	-0,135	-0,447	14	0,203	0,447
5	1,074	-0,863	15	-1,479	-0,672
6	0,295	0,091	16	-0,755	-0,334
7	0,334	-1,843	17	0,550	0,462
8	1,711	-0,375	18	0,558	0,412
9	0,659	-0,200	19	0,293	0,002
10	0,210	-3,220	20	-0,170	-1,540

was simulated to be normally distributed with a mean of 0 and a standard deviation of 1. Next, 2PL model was used to calibrate model parameters. Table 1 shows estimated item parameters of item difficulty and item discrimination.

RESULTS

In this section, it is aimed to provide an example of procedures for investigating model data fit using simulted 20 dichotomous items from responses of 1000 examinees. The first step in the research is choosing software and providing classical analysis.

Classical item analysis can assist in choosing IRT models. In the dichotomous case, the level of variation in item discrimination indices provides an indication about whether or not a discriminating parameter is needed. A wide range of classical item discrimination indices may suggest the need for the discriminating parameter in an IRT model, otherwise considerabe information would be lost and model fit would be poorer. The level of difficulty of multiple-choose items provides an indication of the need of a “guessing parameter” in the IRT model. If items are easy, the guessing parameter may not be necessary (Zhao, 2008).

This research was conducted by simulated data sets and data were generated to fit the 2-PL model. Item discrimination and difficulty parameters were simulated to be normally distributed. For these reasons, classical item analysis is not necessary in this step. BILOG, BILOG-MG and PARSCLE are three main programs for IRT analysis. In this research BILOG was used.

The second step in the research is to determine the dominance of the first factor (unidimensionality), and check the other model assumptions. Unidimensionality assumotion can be checked by exploratory factor analysis or confirmatory factor analysis. Cook et al. (2009) do not recommend CFA for checking unidimensionality. They argued that CFA’s fit values are sensitive and not reliable to decide the factors. For this reason, a linear factor analysis procedure is a popular approach to investigating the unidimensionality assumption. For this aim, factor analysis has been used for checking unidimensionality assumption. Figure 2 shows the dominance of the first factor.

From the scree plot, the largest eigenvalue is easily distinguishable from the smaller ones. This plot serves as a baseline for interpreting the dimensionality of the simulated data. As a result, these findings show the unidimensionality of the data. Local independence, which is the second assumption, means that the response of an individual to an item from the test is not influenced by his or her responses to the other items of the same test or by other traits of theta (Sijtsma and Hemker, 2000). This assumption is necessary in IRT to assure the independence of items and hence their multiplicative property when ascertaining likely abilities for response patterns (Pomplun, 1988). According to McDonald (1982), a theory of unidimensionality should be based on the assumption of local independence.

The third step in the research is the assessment of model data fit. Because of the binary response format of the items, either the one-parameter logistic IRT (Rasch) model, the two parameter logistic (2PL) IRT model or the three parameter logistic (3PL) IRT model may be appropriate for the data. Goodness of fit statistics can be used to test for the amount of improvement in model fit to the data. IRT models are nested models. The degree of freedom for the test of the difference between the goodness of fit statistics and the nested models is the difference the additional parameters needed to be estimated for the more complex model. In BILOG, -2Log likelihood ratio is commonly used to check the goodness of model fit.

Model differences in these values (-2Log likelihood ratio) may be evaluated as chi-square statistics to evaluate the improvements made by the successively more complex models (Embretson and Reise, 2000). Improvement made by the 2-PL model over Rasch model is evaluated as follows;

X^2= -2Log likelihood$_{RASCH}$ – (-2Log likelihood$_{2\text{-}PL}$) = 12142,537 - 10857,662 = 1284,875

At 20 degrees of freedom, the X^2 of 1284,875 is very unlikely. Thus, 2-PL model fits significantly better than the Rasch model. Similarly, the difference between the 3-PL and the 2-PL models can be evaluated by;

X^2= -2Log likelihood$_{2\text{-}PL}$ – (-2Log likelihood$_{3\text{-}PL}$) = 10857,662 - 10853,048 = 4.614

The resulting value of 4.64 is significant at 20 degrees of freedom; therefore, the 2-PL model fits significantly better than the 3-PL model.

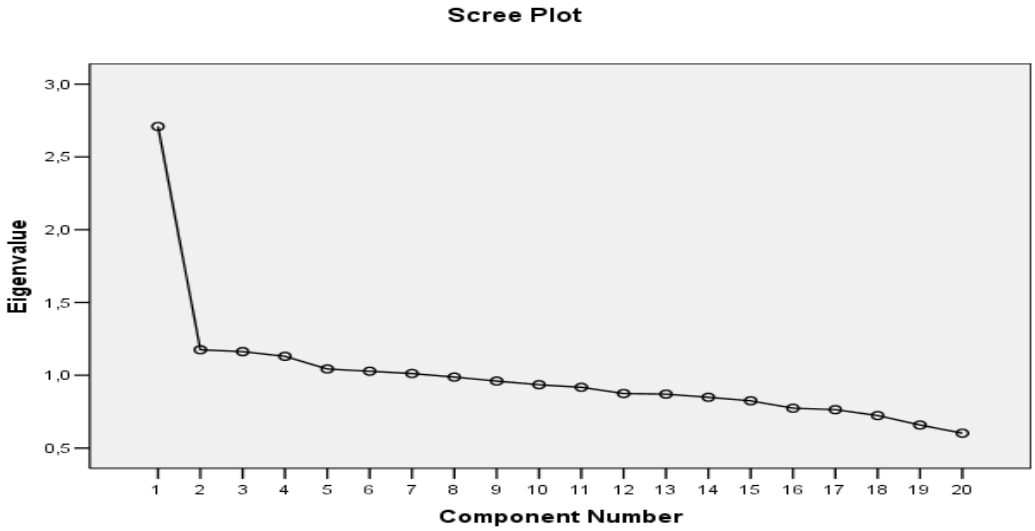

Figure 2. Scree plot.

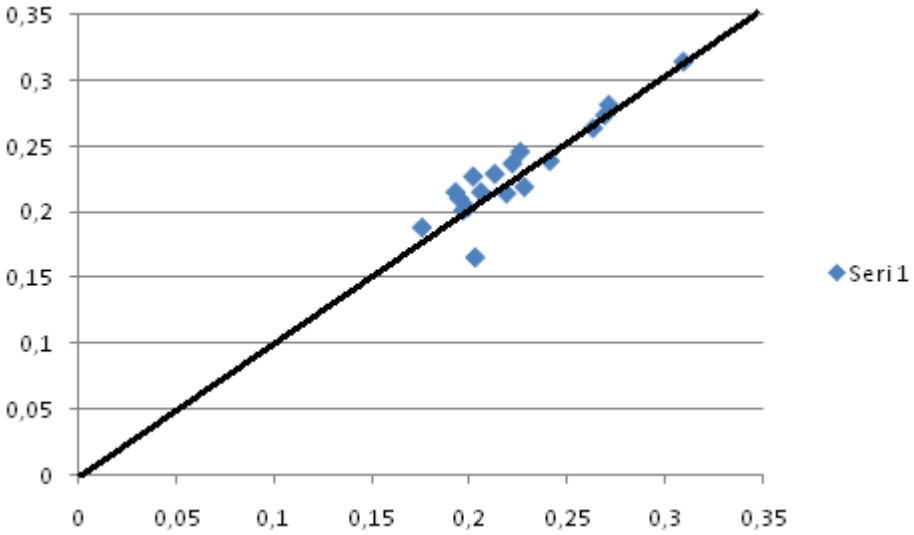

Figure 3. Item discrimination values based on two groups of examinees.

The most common criticism about the chi-square like statistics is that they are sensitive to sample size (Hambleton and Swaminathan, 1989). In this research sample size was 1000. For this reason invariance of item and ability patameters should be checked.

The next step was to investigate the invariance of the model parameters for the 2-PL model. The sample of 1000 examinees was split into two randomly equivalent groups of 500. In these groups, two ability groups were formed: the top half of the distribution and the bottom half of the distribution. After that, item difficulty and item discrimination parameters were estimated for each ability groups. These plots show high relationships between the sets of “b” and “a” values in the two samples. Figure 3 and 4 indicate that item parameter invariance is present.

Invariance of ability parameters over randomly equivalent test forms (e.g. ability estimates based on examinee performance on the odd-numbered items and the even numbered items) indicates the variability due to the sampling of test items. A more rigorous test of invariance would be a comparison of ability estimates over tests consisting of the easiest and hardest items in the item bank (Hambleton et al., 1991).

Invariance of ability parameters across different

Table 2. Item difficulty parameters of easy and hard items.

Easy Items		Hard Items	
Item	**β**	**Item**	**β**
10	-3,220	16	-0,334
11	-2,481	9	-0,200
7	-1,843	19	0,002
20	-1,540	3	0,016
13	-0,916	6	0,091
5	-0,863	1	0,115
15	-0,672	18	0,412
4	-0,447	14	0,447
2	-0,403	17	0,462
8	-0,375	12	1,399

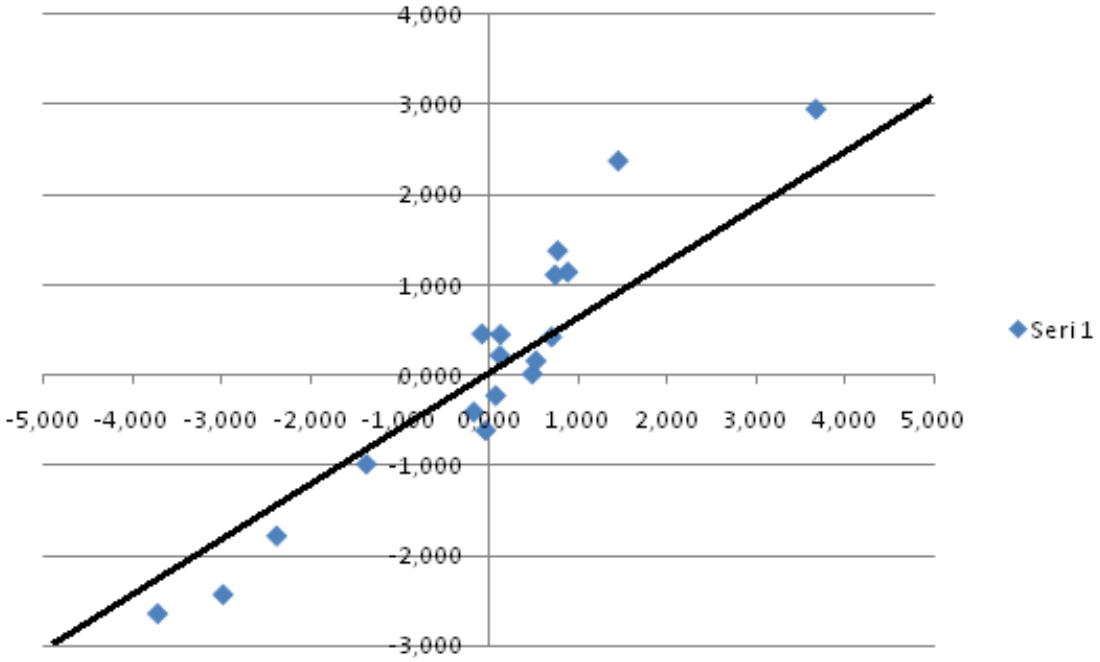

Figure 4. Item difficulty values based on two groups of examinees.

samples of items was investigated. For this aim, items were divided into two groups (10 hard and 10 easy items) (Table 2).

The ability parameter of exaiminees was calculated for easy and hard items. The scatter dot graph is obtained for abilities.

Figure 5 shows that abilities are estimated from two testlets on the line. This result provides evidence of the invariace of ability parameters over tests of varying difficulty. Based on the plots and other findings, it is obvious that simulated data were fit for the 2-PL model.

Conclusion

The pupose of this article is to assess the goodness of fit of unidimensional IRT models in simulated data. In this study an example of how to evaluate model-data fit was given. Assessment of model data fit is a stepwise procedure. Zhao (2008) suggested four step procedures for model data fit. With the help of simulated data, assessment of model data fit was examplified by using these steps for unidimensional IRT models.

Assessing model data fit is an important part of the test

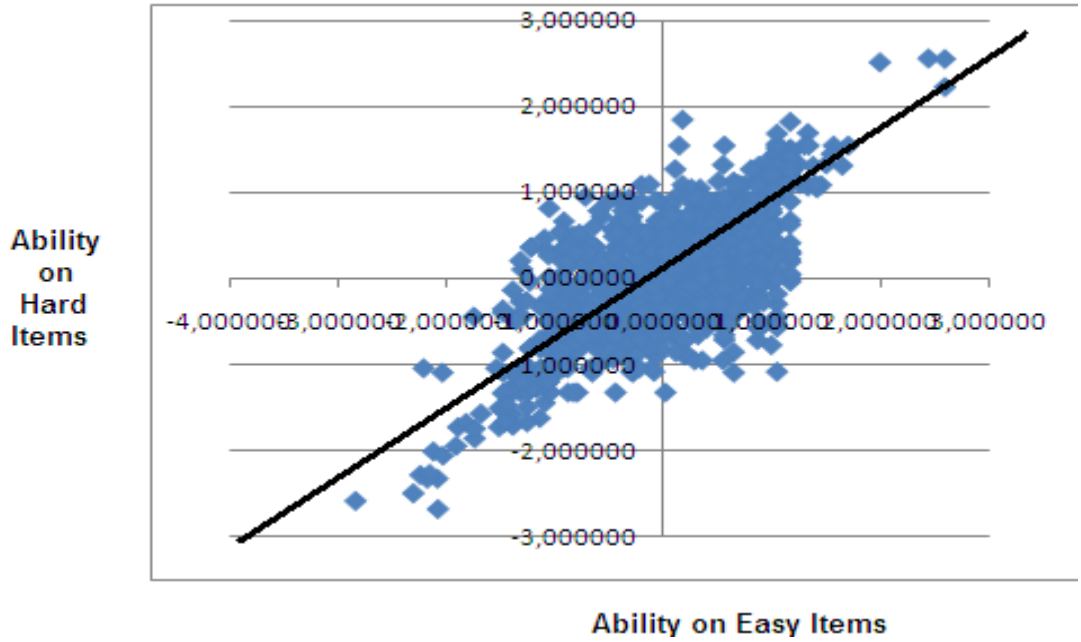

Figure 5. Invariance of ability parameters across different samples.

validation process. Assessing IRT model fit to item response data is one of the crucial steps before an IRT model can be applied with confidence to estimate proficiency or ability levels of exaimeinees (Stone and Zhang, 2003).

The assessment of fit of IRT models usually involves collecting of a wide variety of diagnostic evidences for model fit and then making an informed judgement about model fit and usefulness of a model with a particular set of data (Hambleton, 1994). Besides these, model-data misfit may be attributed to violation of model assumptions or the specific parameterization for the IRT model (number of parameters). For example, exclusion of relevant item or ability parameters may influence the appropriateness of IRT model. However, IRT model fit studies have not received the attention they deserve among test practitioners. Possible reasons for this neglect are the complexity of assessing fit, the lack of understanding of the fit statistics and the absence of comprehensive model fit software (Zhao, 2008).

This study can be replicated for different distributions, sample sizes and test lengths in simulated and real data. Also asessing model-data fit can be investigated in polytomous IRT models and multidimensional IRT models.

Conflict of Interests

The authors have not declared any conflict of interests.

REFERENCES

Bobcock BGE (2009). Estimating a Noncompensatory IRT Model Using a modified Metropolis algorithm. Unpublished Doctoral Dissertation.The University of Minesota.

Bock RD, Thissen D, Zimowski MF (1997). IRT estimation of domain scores. J. Educa. Meas. 34:197-211

Cook KF, Kallen MA, Amtmann D (2009). Having a fit: Impact of number of items and distribution of data on traditional criteria for assessing IRT's unidimensionalty assumption. Q. Life Res. 18(4):447-460.

Cooke DJ, Michie C (1997). An item response theory analysis of the Hare Psychopathy Checklist-Revised. Psychol. Assessment. 9:3-14.

Dodeen H (2004). The relationship between item parameters and item fit. J. Educa. Meas. 41(3):261-270.

Drasgow F (1982). Choice of test model for appropriateness measurement. Appl. Psychol. Meas. 6:297-308.

Embretson SE, Reise SP (2000). Item Response Theory for Psychologists. Mahwah, NJ:Erlbaum.

Glass CAW, Falcon JCS (2003). A comparison of the item-fit statistics for the three-parameter logistic model. Appl. Psychol. Meas. 27:87-106.

Hambleton RK, Swaminathan H (1989). Item Response Theory. Principles And Applications. Kluwer-Nijhoff Publishing. Boston-USA.

Hambleton RK, Swaminathan H, Rogers H (1991). Fundamentals of Item Response Teory**.** Newbury Park CA: Sage.

Hambleton RK (1994). Item response theory: A broad psychometric framework for measurement advances. Psicothema 6(3):535-556.

Kingston NM, Stocking ML (1986). Psychometric issues in IRT-based test construction. Paper presented at the annual meeting of the American Psychological Association (Washington, DC, August 22-26-1986) http://files.eric.ed.gov/fulltext/ED300456.pdf

Kreiter CD (1993). An emprical investigation of compensatory and noncompensatory test items in simulated and real data. Unpublished Doctoral Dissertation. The University of Iowa.

Lee SH (2007). Multidimensional Item Response Theory: A SAS MDIRT MACRO and Emprical Study of PIAT MATH Test. Unpublished Doctoral Dissertation. The University of Oklahoma.

Liu J (2007). Comparing Multidimensional and Unidimensional Computer Adaptive Strategies in Psychological and Health Assessment. Unpublished Doctoral Dissertation. University of Pittsurg.

McDonald RP (1982). Linear versus models in item response theory Appl. Psychol. Measurement. 6:379-396.

Pomplun MR (1988). Effecets of local dependence in achievement tests on IRT ability estimation. Unpublished Doctoral Dissertation. The Florida State University.

Seungho Yang MA (2007). A Comparison of Unidimensional and Multidimensional Rasch Models Using Parametrer Estimates and Fit İndices When Assumption of Unidimensionality is Violated. Unpublished Doctoral Dissertation. The Ohio State University

Sheng Y (2005). Bayesian Analysis of Hierarchical IRT models: Comparing and Combining the Unidimensional and Multidimensional IRT models. Unpublished Doctoral Dissertation. University of Missouri-Columbia.

Sijtsma K, Hemker BT (2000). A taxonomy of IRT models for ordering persons and items using simple sum scores. J. Educa. Behavioral Statistics. 25 (4):391-415.

Sijtsma K, Junker BW (2006). Item response theory: Past performance, present developments and future expectatitons. Behaviormetrika, 33 (1):75-102.

Sinharay S (2005). Assessing fit of unidimensional item response theory models using a Bayesian approach. J. Educ. Meas. 42(4):375-394.

Spencer GS (2004). The Strength of Multidimensional Item Response Theory in Exporing Consrtuct Space That is Multidimensional and Corralated. Unpublished Doctoral Dissertation. Brigam Young University.

Stone CA (2000). Monte-Carlo based null distribution for an alternative fit statistic. J. Educ. Meas. 37:58-75.

Stone CA, Zhang B (2003). Asessing goodness of fit of item response theory models: A comparison of traditional and altenative procedures. J. Educa. Meas. 40(4):331-352.

Wu M, Adams R (2006). Modelling mathematics problem solving item responses using a multidimensional IRT model. Math. Educ. Res. J. 18(2):93-113.

Yen WM (1981). Using simultaneous results to choose a latnt trait model. Appl. Psychol. Meas. 5:245-262.

Zhao Y (2008). Approaches for addressing the fit of iItem response theory models to educational test data. Unpublished Doctoral Dissertation. University of Massachusetts Amberst.

.