ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2023; 14(1); 19-32

Examining Measurement Invariance in Bayesian Item Response Theory Models: A Simulation Study

Merve AYVALLI* Hülya KELECİOĞLU**

Abstract

The aim of the study is to determine a measurement invariance cut-off point based on item parameter differences in Bayesian Item Response Theory Models. Within this scope, the Bayes factor is estimated for testing measurement invariance. For this purpose, a simulation study is conducted. The data were generated in the R software for each simulation condition under the one-parameter logistic model for 10 binary (1-0 scored) items. The invariance test was performed for various group sizes (n=500, 1000, 1500 and 2000) and difficulty parameters (d_k=0, d_k=0.1, d_k=0.3, d_k=0.5 and d_k=0.7). The Bayesian analyzes were performed on the WINBUGS using the codes written in the R. A Bayes factor that provides evidence for measurement invariance was calculated depending on the parameter differences. The Savage–Dickey density ratio, one of the MCMC sampling schemas, was used to calculate the Bayes factor. As a result, if the item parameter difference is d_k=0.3 and group sizes are 1500 or larger, the measurement invariance cannot be achieved. However, for small sample sizes (n=1000 or less) measurement invariance interpretation should be done carefully. When the d_k=0.5, there are invariant items only in n=500. According to Bayes factor test results, evidence has been produced when d_k=0.7 that measurement invariance cannot be achieved.

Keywords: Measurement invariance, bayesian IRT models, bayes factor, random item effects modelling

Introduction

The frequency of tests applied in education and psychology to measure latent variables such as cognitive and affective characteristics in groups having different characteristics has been progressively increasing. These kinds of testing applications often include a comparison among specific groups. Especially in the international large-scale assessments which aim to make comparisons against time or among different groups in terms of their mathematics, science, or reading skills as well as other psychological structures such as attitude, motivation, or anxiety (Davidov et al., 2014). In order to make meaningful comparisons among groups the measured latent variable must be the same in all subgroups. The measurement invariance is an important prerequisite for making comparisons between individuals or groups with varying demographic characteristics, such as different cultures, genders, or regions, to which the measurement tool is applied, by considering these differences. This is important to ensure the generalizability of the measured structure in different groups (Brown, 2006).

There are several methods for testing measurement invariance (Millsap, 2011). These can be examined in two different groups. One of these methods is the confirmatory factor analysis-based method. In the confirmatory factor analysis-based methods, the measurement invariance is examined by testing the similarity of measurement models between groups. One of the most important advantages of this method is that measurement invariance can be examined in all aspects such as factor loadings, intercepts,

* PhD. Student, Hacettepe University, Faculty of Education, Ankara-Türkiye, merveayvalli@gmail.com, ORCID ID: 0000-0002-7301-0096
** Prof. Dr., Hacettepe University, Faculty of Education, Ankara-Türkiye, hulyaebb@hacettepe.edu.tr, ORCID ID: 0000-0002-0741-9934
The present study is a part of PhD Thesis conducted under the supervision of Prof. Dr. Hülya KELECİOĞLU and prepared by Merve AYVALLI.

To cite this article:
Ayvallı, M., & Kelecioğlu H. (2023). Examining measurement invariance in Bayesian Item Response Theory models: A simulation study. *Journal of Measurement and Evaluation in Education and Psychology, 14*(1), 19-32. https://doi.org/10.21031/epod.1101457

Received: 11.04.2022
Accepted: 3.10.2022

residual variances, factor variances and covariance, and the latent-means. The Measurement invariance is tested by making comparisons among nested models (Meredith, 1993; Steenkamp and Baumgartner, 1998). The Higher levels of invariance require strict parameter equality and constraints between groups, which is difficult to meet in real applications.

The other group of methods includes the Item Response Theory (IRT) based methods. In the IRT-based methods, measurement invariance is tested by item bias methods used to evaluate the values of item-level observations in subgroups. Unlike the confirmatory factor analysis (CFA) based methods assuming a linear relationship at the item level, in the IRT based methods, a non-linear relationship is revealed between the latent structure and the item level scores. The Lord's $\chi 2$, Raju's area measures, Wald statistics (likelihood ratio test), Mantel-Hanzel procedure are among the IRT based methods (Millsap, 2011). However, all of these methods have some limitations such as the inability to provide evidence for the measurement invariance hypothesis and requiring to identify anchor items before the analysis (Verhagen, Levy, Millsap, and Fox, 2016).

Verhagen and Fox (2013) suggested a Bayes Factor on the basis of the variance of the item parameters between groups in to compare the measurement invariance hypothesis in nested and large groups such as countries and schools. The calculation of the Bayes Factor provides evidence both in favour and against the measurement invariance hypothesis, unlike the frequentist methods. In addition, anchor items are not needed for this method (Verhagen et al., 2016). However, this method is not convenient to compare a small number of groups.

Verhagen et al. (2016) proposed a different Bayesian factor that allows the comparison of a small number of groups and testing measurement invariance. The presented Bayesian measurement model is discussed and presented within the framework of a one-parameter logistic model.

In a test with i person (i=1, …, N) and k binary items (k=1, …, K), the probability of a correct response in the one-parameter logistic model (θ: ability parameter, bk: item difficulty parameter) as shown Equation 1 (Wright, 1977):

$$P(Y_{ik} = 1|\theta_i, b_k) = \frac{e^{(\theta_i - b_k)}}{1+e^{(\theta_i - b_k)}} \qquad (1)$$

In the Bayesian IRT models, item parameters' priors determine the alterations of item characteristics. In the random item effect models, all test items are considered as a random sample of the item population. The item parameters' priors for all items show standard normal distribution with a common mean and variance (Janssen, Tuerlinckx, Meulders & De Boeck, 2000; De Boeck, 2008).

$$b_k \sim N(b_0, \sigma_{b_k}^2)$$

The posterior distributions for each parameter are the functions of the combination of the average percent accuracy of that item in all groups and a prior distribution, the b_0 and $\sigma_{b_k}^2$. The standard normal prior is selected for the prior distributions of the person parameters (Fox, 2010). In the measurement invariance test, the Bayesian IRT model is considered within the scope of multiple groups IRT because multiple group comparisons are made. The multiple-group IRT models allow differences in test scores and item characteristics among groups (Bock & Zimowski, 1997). Thus, in the Bayesian IRT model, a measurement model is created by considering the variation of group-specific item parameters between groups, as well as the variation among the items. The probability of a person's correct response in multiple group IRT one-parameter logistic model is shown in Equation 2 (j group, i person, θ_{ij}= group-specific person parameter, $\tilde{b}_{kj}$= Group-specific item parameter):

$$P(Y_{ijk} = 1|\theta_{ij}, \tilde{b}_{kj}) = \frac{e^{(\theta_{ij} - \tilde{b}_{kj})}}{1+e^{(\theta_{ij} - \tilde{b}_{kj})}} \qquad (2)$$

Group-specific person parameter θ_{ij}, which is hierarchically modeled, shows a normal distribution around group mean μj.

$$\theta_{ij} \sim N(\mu_j, \sigma_j)$$

In the multiple-group IRT models, it is assumed that group-specific item parameters have a multilevel structure for modeling the measurement variance (Fox, 2010). The Group-specific deviations have normal distributions with a mean of zero and $\sigma^2_{b_k}$ for all items. This variance component defines the variability of item functions between groups. When this variance is zero, the item is considered to be invariant because there is no variability. In addition, if a measurement invariance study is desired between a small number or fixed groups, it is more useful to use the fixed group model instead of the random group model, since it will be difficult to estimate the random item effects variance.

Verhagen et al. (2016) introduced a model that can test measurement invariance between two groups with Bayes factor. In this model, the group-specific item parameters are estimated separately for different groups. The item parameters are independent, and they do not provide information about each other. In such a case, a possible prior distribution for the group mean is the normal prior distribution with a large variance. A multivariate normal model is applied to group-specific item characteristics. In addition, covariance matrices which are based on the correlation between the item parameters of different groups are used.

The group-specific item parameters defined in the model are shown in Equation 3 (μ_j: mean of the item difficulty in group j, e_{kj}: error term):

$$\tilde{b}_{kj} = \mu_j + e_{kj} \qquad (3)$$

In the model, μ_j equals zero and e_{kj} equals the amount of deviation from the average item difficulty in the group.

These deviations are assumed to show multivariate normal distribution with covariance (Σb) for item difficulties consisting of item parameter variances for each group. The variance of item parameters may vary by group. This means that the variance in the item difficulty parameters of one group is higher than the other. Since the group-specific item difficulties are estimated independently for each group, the measurement invariance can be directly estimated based on the differences between the item difficulty parameters. The difference between the difficulty parameters among the two groups is shown in Equation 4 (k. item, groups j and j', j<j'):

$$d_{kjj'} = b_{kj} - b_{kj'} \qquad (4)$$

In the measurement invariance test for the two groups, the hypotheses are established based on the difference between the item difficulty parameters.

$$H_0 : d_k=0$$
$$H_1 : d_k\neq0$$

The Bayes factor which uses Bayesian hypothesis testing is very advantageous in that it provides direct information about items' measurement invariance in a whole test (Jeffreys, 1961). Additionally, it does not require items that have been proven measurement invariance before (Verhagen et. al., 2016). Furthermore, unlike frequentist statistics, it gives evidence for both H_0 and H_1.

When only evidence for H_0 is given without using the evidence given for H_1, it leads to exaggerated results against H_0 hypothesis that only the evidence for the null hypothesis is considered, especially in low-power studies (Rouder et al., 2009; Wagenmakers et al., 2017). In addition, providing evidence for both hypothesis tests is advantageous in terms of giving information about which hypothesis should be preferred.

In the measurement invariance test, the marginal likelihood of the H_1 hypothesis is weighted by the prior probability of the average likelihood on all possible values of the alternative hypothesis. This average likelihood value is equal to the integral of the likelihood function weighted by the prior density function of the parameters in the hypothesis. The Bayes factor includes the marginal likelihood ratio for both the null hypothesis and alternative hypothesis results.

The Bayes factor that provides relative evidence for the hypotheses tested is as in Equation 5 ($p_1(d_k)$: H_1 under Cauchy prior distribution)

$$BF_{01} = \frac{P(Y|H_0)}{P(Y|H_1)} = \frac{P(Y|d_k = 0)}{\int P(Y|d_k)p_1(d_k)dd_k} \qquad (5)$$

The increase in cross-cultural testing practices also increases the importance of measurement invariance. Traditional methods based on confirmatory factor analysis, which are frequently used in determining measurement invariance, require the comparison of different model fits. In addition, these methods are time-consuming as each of these models is expected to be fitted (White, 2000). Furthermore, additional restrictions are needed in the definition of these models (Reise, Widaman & Pugh, 1993).

Other invariance tests require at least one anchor item of which invariance has already been proven. The methods used to select an item that is invariant for many groups (Langer, 2008) lead to biased estimations if the item contains a certain level of bias (Woods, Cai, & Wang, 2012). Considering all these situations, more practical methods are needed to evaluate measurement invariance, especially in large-scale tests and cross-cultural studies. Unlike the frequentist methods, multiple hypotheses (H_0 and H_1) can be tested simultaneously with the Bayesian method that is used in the study. Thus, all items in the test can be evaluated simultaneously and measurement invariance estimation can be made directly. From both practical and theoretical perspectives, it is thought that the research results will be significant. It will be possible to have an idea about the measurement invariance based on the difference between item difficulty parameters. In addition, it is important that these cut-off points, which are limited by the conditions in the study, form an idea for both frequentist and Bayesian measurement invariance studies.

Purpose of the Study

The aim of the study is to determine a measurement invariance cut-off point based on item parameter differences. In accordance with this purpose, the Bayes factor which estimates invariance directly is used within the scope of Bayesian IRT models. The invariance test was performed when the difficulty parameter differences (d_k) are 0.0, 0.1, 0.3, 0.5, 0.7 at group sizes are 500, 1000, 1500, and 2000.

Method

The Simulation Conditions and Data Generation

The Data was generated in the R software (R Core Team, 2018) when evaluating measurement invariance using the Bayes factor. For each group, the sum of item thresholds (b_{kj}) and reference group's ability parameter ($\mu_{\theta j}$) are assumed to be zero. The sample sizes in groups were equally determined to be 500, 1000, 1500, and 2000. The previous studies suggested that the minimum sample size should be 500 for unbiased parameter estimation (Thompson, 2018, Asparouhov & Muthén, 2014; De Boeck, 2008; Stark et al., 2006). It has been found that by increasing the group size from 500 to 1,000 the Type I error rate decreased, but there was no significant difference in the Type I error when increasing the group size from 1,000 to 2,000 (Finch, 2016). And it was determined that the Bayes factor performed well in group sizes of 500 and more (Verhagen et al., 2016). Thompson (2018) noticed that Bayes Factor for measurement invariance has got higher power rate with larger sample sizes and suggested using at least 500 as a sample size. In addition, considering the real test applications such as PISA, TIMMS, and PIRLS, it is known that the minimum sample sizes are usually 500 and more. Based on these findings and real data applications the sample sizes in groups were equally determined to be 500, 1000, 1500, and 2000 in the study. The data were generated for each simulation condition under the one-parameter logistic model for 10 binary (1-0 scored) items. The difference between the difficulty parameters of the groups was determined as d_k=0, d_k=0.1, d_k=0.3, d_k=0.5 and d_k=0.7. d_k= 0.0 (there is no difference between item difficulty parameters) were considered as invariant items and the difference between the parameters gradually increased (Verhagen et al., 2016). Harwell et al. (1996) stated that 100 or fewer replications would have sufficient power in simulation studies and recommended at least 25 replications. In the current study, 100 replications were applied for each condition. The analyses were carried out for each data set. Item difficulty parameter values for each condition can be seen in Annex-1.

Data Analysis

Bayesian analyzes were performed on the WINBUGS using the codes written in the R. For the difference between the difficulty parameters of each item, a Bayes factor was created to provide evidence for the measurement invariance. In hypothesis testing, the ratio of the density of the null hypothesis under the prior and posterior distributions affects the Bayes factor test results. The Bayes factor test results depend on priors selected for the parameters to be evaluated. The priors can be selected based on the assumptions accepted for the parameter values. Since the Bayes factor is more likely to support measurement invariance when multivariate Cauchy prior is used. The difference between group-specific item parameters is equally distributed under the multivariate Cauchy prior. Thus, the analyses were performed using the multivariate Cauchy prior (Verhagen et al., 2016, Thompson, 2018). The Savage–Dickey density ratio, one of the MCMC sampling schemas, was used to calculate the Bayes factor.

This method is applied in nested models, and the calculation of the Bayesian factor for the parameter under test requires high posterior and prior distribution. Especially in complex models, such as nested structures, this method can be used for invariance testing. In this model, the null hypothesis is the hypothesis that the value of the parameter of interest is fixed, and the alternative hypothesis is the hypothesis that this parameter is released. Therefore, the null hypothesis is nested under the alternative hypothesis (Wagenmakers, Lodewyckx, Kuriyal, and Grasman, 2010). The difference between item difficulty parameters for any of the two groups is defined as:

$$d_k = b_{k1} - b_{k2} = 0$$

The Bayes factor reduces the H_0 to the prior and posterior distributions of the difference between parameters in the H_1, when evaluating the relative support of the H_0= d_k=0 according to the H_1= $d_k \neq 0$ hypothesis. In a simpler expression, it is obtained from the alternative hypothesis by setting it to H_0= 0.

$$BF_{01}=\frac{p_1\,(d_k=0\mid H_1,Y)}{p_1\,(d_k=0\mid H_1)}$$

Thus, the Bayes factor produces more evidence for the null hypothesis than for the alternative hypothesis (Verhagen et al., 2016).

The Bayes factor defines a relative estimation performance for the H_0 and H_1. In other words, it specifies a relative measure of the prediction quality of the hypothesis. For instance, if BF_{01}=5, it means that the data is 5 times more likely to be under H_0 than under H_1. However, the fact that the Bayes factor favors H_0 does not mean that H_0 predicts the data better (van Doorn, van den Bergh, Böhm et al., 2021). According to Jeffreys (1961), the Bayes factors between 1 and 3 produce equal evidence for the null hypothesis and the alternative hypothesis, and these values are accepted as weak evidence. A Bayes factor between 3 and 10 is considered sufficient evidence for the H_0 hypothesis. If the Bayes factor is greater than 10, it is accepted as strong evidence for the H_0 hypothesis. When the Bayes factor is between 0.33 and 0.10, it is accepted as sufficient evidence for the alternative hypothesis, and when it is less than 0.1, it is accepted as strong evidence for the H_1. In the current study, the cut-off point for the Bayes factor was determined as 3 if the invariance holds, and 0.33 if it does not hold. To complete the MCMC processes efficiently, the analysis was carried out with 3000 iterations with a 300 burn-in period.

Findings

The measurement invariance was tested when there is no difference between the difficulty parameters. The Results are shown in Table 1.

Table 1

The Bayes Factor Results for d_k=0.0

		BF_{01}			
		N=1000 (500 per group)	***N=2000 (1000 per group)***	***N=3000 (1500 per group)***	***N=4000 (2000 per group)***
d_k=0.0	**Item_1**	4.59773	6.30162	10.37240	18.34086
	Item_2	10.85618	10.0564	7.82427	16.83027
	Item_3	10.67626	11.04705	7.63375	11.0673
	Item_4	11.57481	9.56177	16.39697	15.62178
	Item_5	6.24642	11.08648	21.14449	14.84418
	Item_6	5.50773	6.36254	10.24370	19. 08634
	Item_7	10.97618	11.0004	8.84272	17.80372
	Item_8	12.67626	10.99905	7. 75363	12.00662
	Item_9	15.57481	9.59548	17.12697	16.16278
	Item_10	6.43642	13.08868	22.00443	15.73325

According to the results for d_k=0.0, in all sample sizes, it is seen that the BF_{01} values of the item parameters are greater than the cut-off point of 3. When the group sizes are 500, 1000, and 1500, the BF_{01} values of 4 items were greater than 3, and the BF_{01} values of 6 items were greater than 10. Since the group size is 2000, it is seen that BF_{01} values for all items are greater than 10, which provides strong evidence for the measurement invariance. The Bayes factor test results can be seen in Table 2 when the difference between parameters is dk=0.1 for each sample size.

Table 2

Bayes Factor Results for d_k=0.1

		BF_{01}			
		N=1000 (500 per group)	***N=2000 (1000 per group)***	***N=3000 (1500 per group)***	***N=4000 (2000 per group)***
d_k=0.1	**Item_1**	9.53619	13.53914	3.94931	9.53863
	Item_2	3.32114	10.49564	10.93055	4.85995
	Item_3	8.24058	8.81567	20.05434	6.65487
	Item_4	11.84632	9.00784	12.1836	16.94075
	Item_5	7.99764	12.38298	9.74235	6.89683
	Item_6	9.78869	15.91453	3.00491	9.63375
	Item_7	5.32114	10.56449	11.04056	4.99502
	Item_8	8.42058	7.99815	21.45434	7.48765
	Item_9	13.87632	9.78400	13.18360	17.93081
	Item_10	8.19765	12.29809	9. 35945	7. 96828

The results in Table 2 show that when the parameter differences are d_k=0.1, it is seen that BF_{01} values are greater than 3 in all sample sizes. It is seen that measurement invariance is obtained for all items. In cases where the difference between the parameters is d_k=0.3, the Bayes factor results calculated based on the difference between the item difficulty parameters are given in Table 3.

When the difference between item difficulties is 0.3, there are invariant items only n=500 and n=1000. The Bayes factor results for 6 items are invariant (n=500). However, there are 4 items producing equal evidence for both the null hypothesis and the alternative hypothesis. Therefore, those items cannot be interpreted as invariant. It was seen that if the group size was 1000 and the difference between the parameters was 0.3, the measurement invariance was obtained in 4 items. Invariance interpretation for 2 items cannot be made because of the equal evidence for the hypotheses (H_0 and H_1).

Table 3

Bayes Factor Results for d_k=0.3

		BF_{01}			
		N=1000 (500 per group)	***N=2000 (1000 per group)***	***N=3000 (1500 per group)***	***N=4000 (2000 per group)***
d_k=0.3	**Item_1**	7.19235	10.3738	2.61409	0.02789
	Item_2	12.04128	1.12155	0.29342	0.05877
	Item_3	8.74905	0.20997	0.32324	0.09858
	Item_4	1.92266	4.18721	0.01816	0.27239
	Item_5	1.32949	0.0207	0.01307	0.13994
	Item_6	8.19235	13.3738	2. 96104	0.02789
	Item_7	11.94128	1.15585	0.19388	0.05877
	Item_8	7.14901	0.19997	0.23564	0.09858
	Item_9	0.99266	4.72118	0.02817	0.27239
	Item_10	2.31742	0.20700	0.01509	0.13994

When the group size was 1500, there is not an invariant item. According to the Bayes factor results for n=1500, there are 2 items having equal evidence for both H_0 and H_1. That can be considered weak evidence for both hypotheses. It can be said that the values of BF_{01} for the remaining 8 items are less than 0.33, and the items are not invariant. The Bayes factor results are less than 0.33 for n=2000 which means none of the items were invariant. Table 4 shows the Bayes factor results calculated based on the difference between the item difficulty parameter which is d_k=0.5.

Table 4

Bayes Factor Results for d_k=0.5

		BF_{01}			
		N=1000 (500 per group)	***N=2000 (1000 per group)***	***N=3000 (1500 per group)***	***N=4000 (2000 per group)***
d_k=0.5	**Item_1**	0.15524	0.06085	0.00168	0.00016
	Item_2	0.18284	0.00891	0.01364	0.03911
	Item_3	4.09778	0.00085	0.00116	0.00415
	Item_4	0.35232	0.02586	0.06630	0.00147
	Item_5	4.31605	0.03253	0.00393	0.00135
	Item_6	0.23245	0.06857	0.01368	0.00012

Table 4

Bayes Factor Results for d_k=0.5 (Continued)

	Item_7	0.17294	0.00981	0.02264	0.01368
	Item_8	4.00038	0.00508	0.01015	0.00307
d_k=0.5	**Item_9**	0.29852	0.01258	0.06730	0.00112
	Item_10	3.93167	0.03534	0.00298	0.00129

As in Table 4, there are invariant items only for a group size of 500. In other group sizes, there are no items with a Bayes factor value greater than 3. When n=500, the 4 items are invariant, but the remaining items are not.

In all remaining group sizes (n=1000, n=1500, and n=2000), all BF_{01} values of the items are less than 0.10, which provides strong evidence in favor of the H_1 hypothesis. In Table 5, the Bayes factor results are shown for the cases where the difference between the difficulty parameters is d_k=0.7.

Table 5

Bayes Factor Results for d_k=0.7

		BF_{01}			
		N=1000 (500 per group)	***N=2000 (1000 per group)***	***N=3000 (1500 per group)***	***N=4000 (2000 per group)***
	Item_1	0.32990	0.01297	0.00030	0.00003
	Item_2	0.00047	0.00172	0.00156	0.00020
	Item_3	0.00328	0.00136	0.00077	0.00113
	Item_4	0.00232	0.02296	0.00175	0.00008
	Item_5	0.06634	0.00223	0.00026	0.00026
d_k=0.7	**Item_6**	0.26890	0.02129	0.00140	0.00002
	Item_7	0.00007	0.00147	0.00185	0.00009
	Item_8	0.00285	0.00163	0.00113	0.00126
	Item_9	0.00292	0.03295	0.00156	0.00017
	Item_10	0.07654	0.02019	0.00102	0.00032

According to the Bayes factor test results shown in Table 5, there is no evidence for measurement invariance in all group sizes. BF_{01}values for 2 items are less than 0.33 only when the group size is 500, in all other cases, it was determined that the BF_{01} value was less than 0.10 and produced strong evidence in favor of the H_1 hypothesis.

Conclusion

In the study, it was aimed to determine a cut-off point for measurement invariance based on the difference between parameters in different sample sizes and in cases where item parameters differed between groups with the Bayesian IRT model.

According to simulation results;

1. As predicted, it was determined that measurement invariance was achieved in all sample sizes when there was no difference between the difficulty parameters of the groups.
2. When the difference between item difficulty parameters is d_k=0.1, all items are invariant for all sample sizes.
3. Bayes factor results for d_k=0.3 shows that only a few items are invariant if the group sizes are 500 and 1000. It has been found that the number of invariant items decreases as the group size increases. When the group sizes are 1500 and 2000, the Bayes factor test results provide evidence for only the alternative hypothesis. Thus, there is no invariant item for these sample sizes.
4. When the difference between item difficulty parameters is d_k=0.5, there are invariant items only in n=500. Bayes factor results provide strong evidence in favor of H_1 for n=1000, n=1500, and n=2000.
5. There is no evidence in favor of invariant items when d_k=0.7 for all sample sizes.

When the results are evaluated, it is seen that no invariant item was detected independent of group size when the difference between the item difficulty parameters is d_k=0.7. In this situation, it is possible to state that if the item difficulty parameters difference between groups is 0.7, measurement invariance does not hold.

In cases where the d_k=0.5 and the sample size is 1000 or larger, it can be said that measurement invariance cannot be achieved. However, if the group size is n=500 or smaller, it is not possible to evaluate the invariance only based on the item parameter differences. For these sample sizes, it is recommended to perform a measurement invariance test at the item level.

For d_k=0.3 and n=2000 or larger, the measurement invariance is not achieved. But, for n= 1500 or smaller, it is not correct to make a final decision for measurement invariance based solely on the differences among the item parameters. To make a decision on the measurement invariance, the invariance test must be performed.

If there is no difference between difficulty parameters or d_k=0.1, it is possible to say that measurement invariance is achieved in all group sizes.

There are many studies related to the Bayesian approximate invariance and alignment optimization method to determine Bayesian measurement invariance. On the other hand, the studies are limited to investigating measurement invariance with the Bayes factor. In the literature, Verhagen (2013) stated that the Bayes factor performs well in detecting the measurement invariance when the difference between the item difficulty parameters is large (d_k>0.5). If there is a smaller difference (d_k=0.1 or d_k=0.3), the Bayes factor could not decide on invariance for most items. Also, Verhagen et al. (2016) have shown that the Bayes factor is a valid method for determining invariance when the difference between item difficulty parameters is 0.5 or more. It will be easier to detect measurement invariance with respect to the cut-off point, especially when group sizes are large. Thompson (2018) has shown that the Bayes Factor distinguishes invariant and non-invariant items.

In conclusion, providing measurement invariance is a prerequisite for meaningful comparisons, especially when comparisons between groups are required (Horn ve McArdle, 1992). The current study revealed that it is possible to have an idea about the measurement invariance based on item difficulty parameter differences and it creates a practical framework for doing meaningful comparisons across groups. Defining which items are most likely to be invariant with the difference between item difficulty parameters, provides pragmatic information for measurement invariance and differential item

functioning studies. The cut-off points presented in the study can be used in applications that are compatible with the conditions described in the research. In addition, as a secondary outcome, when an anchor item needs to be determined, the item selections can be made by considering the cut-off points d_k=0.0 and d_k=0.1

The current research is limited to the simulation conditions explained in detail in the method section, and binary items. Studies on polytomous items, unequal sample sizes, different simulation conditions, and real data applications can be conducted in future studies. Furthermore, the study focused on only the Bayes factor for detecting measurement invariance. In future applications, studies can be performed using not only the Bayes factor but also other Bayesian criteria such as Deviance Information Criteria (DIC).

Declarations

Author Contribution: Merve Ayvallı-Conceptualization, methodology, analysis, writing & editing, visualization. Hülya Kelecioğlu-Conceptualization, methodology, writing-review & editing, supervision.

Conflict of Interest: No potential conflict of interest was reported by the authors.

Ethical Approval: Simulated data were used in this study. Therefore, ethical approval is not required.

References

Asparouhov, T., & Muthén, B. (2014). Multiple-group factor analysis alignment. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(4), 495-508. https://doi.org/10.1080/10705511.2014.919210

Bock, R. D., & Zimowski, M. F. (1997). The multiple groups IRT. In Wim J. van der Linden, & Ronald K. Hambleton (Eds.), *Handbook of modern item response theory*. Springer-Verlag. https://doi.org/10.1007/978-1-4757-2691-6_25

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. Newyork: Guilford Publications.

Davidov, E., Meuleman, B., Cieciuch, J., Schmidt, P., & Billiet, J. (2014). Measurement equivalence in cross -national research. *Annual Review of Sociology*, *40*, 55-75. https://doi.org/10.1146/annurev-soc-071913-043137

De Boeck, P. (2008). Random item IRT models. *Psychometrika*, *73*(4), 533-559. https://doi.org/10.1007/s11336-008-9092-x

Finch, W. H. (2016). Detection of differential item functioning for more than two groups: A Monte Carlo comparison of methods. *Applied measurement in Education*, *29*(1), 30-45. https://doi.org/10.1080/08957347.2015.1102916

Fox, J. P. (2010). *Bayesian item response modeling: Theory and applications*. Springer Science & Business Media. https://doi.org/10.1007/978-1-4419-0742-4

Harwell, M., Stone, C. A., Hsu, T. C., & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied psychological measurement*, *20*(2), 101-125. https://doi.org/10.1177/014662169602000

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental aging research*, *18*(3), 117-144. https://doi.org/10.1080/03610739208253916

Janssen, R., Tuerlinckx, F., Meulders, M., & De Boeck, P. (2000). A hierarchical irt model for criterion-referenced measurement. *Journal of Educational and Behavioral Statistics*, *25*, 285–306. https://doi.org/10.3102/10769986025003285

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford University Press.

Langer, M. M. (2008). A reexamination of Lord's Wald test for differential item functioning using item response theory and modern error estimation (Doctoral dissertation, The University of North Carolina at Chapel Hill).

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*(4), 525-543. https://doi.org/10.1007/BF02294825

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. New York, NY: Routledge. https://doi.org/10.4324/9780203821961

R Core Team. (2018). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: two approaches for exploring measurement invariance. *Psychological bulletin*, *114*(3), 552.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16*, 225–237. https://doi.org/10.3758/PBR.16.2.225

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: toward a unified strategy. *Journal of Applied Psychology, 91*(6), 1292. https://doi.org/10.1037/0021-9010.91.6.1292

Steenkamp, J. B. E., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of consumer research*, *25*(1), 78-90. https://doi.org/10.1086/209528

Thompson, Y. T. (2018). Bayesian and Frequentist Approaches for Factorial Invariance Test (Doctoral dissertation, University of Oklahoma).

Van Doorn, J., van Den Bergh, D., Böhm, U., Dablander, F., Derks, K., Draws, T., Etz, A., Evans, N. J., Gronau, Q. F., Haaf, J. M., Hinne, M., Kucharský, Š. L., Marsman, A., Matzke, M., Gupa, D., R, A., Sarafoglou, A., Stefan, A., Voelkel, J. G., & Wagenmakers, E.-J. (2021). The JASP guidelines for conducting and reporting a Bayesian analysis. *Psychonomic Bulletin & Review*, 28(3), 813–826. https://doi.org/10.3758/s13423-020-01798-5

Verhagen, A. J., & Fox, J. P. (2013). Bayesian tests of measurement invariance. *British Journal of Mathematical and Statistical Psychology*, *66*(3), 383-401. https://doi.org/10.1111/j.2044-8317.2012.02059.x

Verhagen, J., Levy, R., Millsap, R. E., & Fox, J. P. (2016). Evaluating evidence for invariant items: A Bayes factor applied to testing measurement invariance in IRT models. *Journal of Mathematical Psychology*, *72*, 171-182. https://doi.org/10.1016/j.jmp.2015.06.005

Wagenmakers, E. J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage–Dickey method. *Cognitive Psychology*, *60*(3), 158-189. https://doi.org/10.1016/j.cogpsych.2009.12.001

Wagenmakers, E. J., Verhagen, J., Ly, A., Matzke, D., Steingroever, H., Rouder, J. N., & Morey, R. D. (2017). The need for Bayesian hypothesis testing in psychological science. In S. O. Lilienfeld & I. D. Waldman (Eds.). *Psychological science under scrutiny: Recent challenges and proposed solutions*, (pp. 123-138). https://doi.org/10.1002/9781119095910.ch8

White, H. (2000). A reality check for data snooping. *Econometrica*, *68*(5), 1097-1126. https://doi.org/10.1111/1468-0262.00152

Woods, C. M., Cai, L., & Wang, M. (2012). The Langer-improved Wald test for DIF testing with multiple groups: Evaluation and comparison to two-group IRT. *Educational and Psychological Measurement, 73*, 532–547. https://doi.org/10.1177/0013164412464875

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of educational measurement*, *14*(2) 97-116. https://www.jstor.org/stable/1434010

Annex 1

Difference between parameters	Items	Group 1 N= 500	Group 2 N= 500	Group 1 N= 1000	Group 2 N= 1000	Group 1 N= 1500	Group 2 N= 1500	Group 1 N= 2000	Group 2 N= 2000
d_k=0.0	**Item_1**	-1.51524	-1.51524	1.35099	1.35099	-0.14897	-0.14897	0.61867	0.61867
	Item_2	1.15225	1.15225	0.15923	0.15923	0.25329	0.25329	-1.20775	-1.20775
	Item_3	-0.58729	-0.58729	-1.04742	-1.04742	-0.13534	-0.13534	-0.08178	-0.08178
	Item_4	-1.17103	-1.17103	-2.31015	-2.31015	1.17636	1.17636	-0.95174	-0.95174
	Item_5	-0.19137	-0.19137	1.41825	1.41825	-0.44912	-0.44912	-1.66571	-1.66571
	Item_6	1.00173	1.00173	-0.8556	-0.8556	-0.85863	-0.85863	1.11071	1.11071
	Item_7	0.92033	0.92033	-0.22294	-0.22294	0.10577	0.10577	-0.18306	-0.18306
	Item_8	-0.19002	-0.19002	-0.1573	-0.1573	-0.77499	-0.77499	-0.09381	-0.09381
	Item_9	0.13939	0.13939	1.84483	1.84483	0.7638	0.7638	1.10559	1.10559
	Item_10	0.44124	0.44124	-0.17991	-0.17991	0.06784	0.06784	1.34887	1.34887
d_k=0.1	**Item_1**	0.11130	0.21130	1.47472	1.57472	-1.14341	-1.04341	-1.25033	-1.15033
	Item_2	-1.42535	-1.52535	-0.72646	-0.82646	0.64185	0.54185	-0.98874	-1.08874
	Item_3	1.0506	1.1506	1.28423	1.38423	-0.23042	-0.13042	0.46737	0.56737
	Item_4	0.61857	0.51857	0.72544	0.62544	-0.14181	-0.24181	0.28588	0.18588
	Item_5	-1.17563	-1.07563	-1.6525	-1.5525	0.33662	0.43662	-0.4916	-0.3916
	Item_6	0.85522	0.75522	-0.6442	-0.7442	1.94682	1.84682	1.52708	1.42708
	Item_7	-0.72103	-0.62103	-0.2051	-0.1051	0.04659	0.14659	-0.2995	-0.1995
	Item_8	0.67932	0.57932	1.21579	1.11579	-0.46611	-0.56611	-1.12467	-1.22467
	Item_9	-1.13605	-1.03605	-0.84939	-0.74939	0.98127	1.08127	0.72206	0.82206
	Item_10	1.14304	1.04304	-0.62251	-0.72251	-1.97141	-2.07141	1.15243	1.05243
d_k=0.3	**Item_1**	-0.83200	-0.53200	1.39128	1.69128	-0.24920	0.05080	0.31251	0.61251
	Item_2	2.23221	1.93221	-1.75713	-2.05713	-0.48799	-0.78799	-0.34488	-0.64488
	Item_3	1.19172	1.49172	-0.58153	-0.28153	0.23554	0.53554	0.28424	0.58424
	Item_4	-0.15327	-0.45327	1.3534	1.0534	0.28922	-0.01078	0.39478	0.09478
	Item_5	-0.85613	-0.55613	-0.49663	-0.19663	-1.10461	-0.80461	-1.6253	-1.3253
	Item_6	-1.24592	-1.54592	1.27965	0.97965	-0.53536	-0.83536	0.34952	0.04952
	Item_7	0.30058	0.60058	-0.97319	-0.67319	0.65091	0.95091	1.91602	2.21602

Annex 1 (Continued)

d_k=0.3	**Item_8**	-0.37292	-0.67292	-0.18041	-0.48041	0.55902	0.25902	-0.05588	-0.35588
	Item_9	-0.49208	-0.19208	-0.22857	0.07143	1.61207	1.91207	-1.14867	-0.84867
	Item_10	0.22780	-0.07220	0.19312	-0.10688	-0.96959	-1.26959	-0.08233	-0.38233
d_k=0.5	**Item_1**	1.01674	0.51674	0.69355	0.19355	0.38955	-0.11045	0.91923	0.41923
	Item_2	-0.78838	-0.28838	1.6497	2.1497	-0.76968	-0.26968	-0.24888	0.25112
	Item_3	-0.08844	-0.58844	-0.18656	-0.68656	1.59203	1.09203	-1.80859	-2.30859
	Item_4	-0.82591	-0.32591	0.92077	1.42077	0.83879	1.33879	-0.5192	-0.0192
	Item_5	0.88156	0.38156	0.36453	-0.13547	1.4836	0.9836	3.13911	2.63911
	Item_6	0.4986	0.9986	-0.801	-0.301	0.43809	0.93809	-1.70057	-1.20057
	Item_7	-0.47966	-0.97966	-0.60329	-1.10329	-0.74047	-1.24047	0.49179	-0.00821
	Item_8	-0.3249	0.1751	0.09697	0.59697	-1.30876	-0.80876	-0.14801	0.35199
	Item_9	-1.01617	-1.51617	-0.97367	-1.47367	-0.76574	-1.26574	0.23068	-0.26932
	Item_10	1.12655	1.62655	-1.16101	-0.66101	-1.15741	-0.65741	-0.35558	0.14442
d_k=0.7	**Item_1**	-0.02227	-0.72227	-0.06218	-0.76218	0.99252	0.29252	1.78028	1.08028
	Item_2	0.29443	0.99443	2.12972	2.82972	0.43886	1.13886	0.3036	1.0036
	Item_3	-0.88435	-1.58435	-0.77638	-1.47638	0.32672	-0.37328	0.74875	0.04875
	Item_4	0.98951	1.68951	0.12361	0.82361	-0.37344	0.32656	-0.46008	0.23992
	Item_5	0.00746	-0.69254	-0.58109	-1.28109	0.21024	-0.48976	-0.46803	-1.16803
	Item_6	-1.44076	-0.74076	-0.1527	0.5473	0.35615	1.05615	-1.92719	-1.22719
	Item_7	-0.21819	-0.91819	-1.40278	-2.10278	-0.45817	-1.15817	0.59344	-0.10656
	Item_8	2.1256	2.8256	0.84659	1.54659	-1.23148	-0.53148	0.20459	0.90459
	Item_9	-0.54605	-1.24605	0.42517	-0.27483	-0.31452	-1.01452	0.34729	-0.35271
	Item_10	-0.30538	0.39462	-0.54997	0.15003	0.05311	0.75311	-1.12264	-0.42264