The Effects of Q-Matrix Misspecification on Item and Model Fit

Seçil Ömür Sünbül [i]
Mersin University

Semih Aşiret [ii]
Mersin University

Abstract

In this study it was aimed to evaluate the effects of various factors such as sample sizes, percentage of misfit items in the test and item quality (item discrimination) on item and model fit in case of misspecification of Q-matrix. Data were generated in accordance with DINA model. Q-matrix was specified for 4 attributes and 15 items. While data were generated, sample sizes as 1000, 2000, 4000, s and g parameters as low and high discrimination index were manipulated. Three different misspecified Q-matrix (overspecified, underspecified and mixed) was developed considering the percentage of misfit items (%20 and %40). In the study, $S-\chi^2$ was used as item fit statistics. Furthermore absolute (abs(fcor), max(χ^2)) and relative (-2 log-likelihood, Akaike's information criterion (AIC) and Bayesian information criterion (BIC)) model fit statistics were used. Investigating the results obtained from this simulation study, it was concluded that $S-\chi^2$ was sufficient to detect misfit items. When the percentage of misfit items in the test was much or Q-matrix was both underspecified and overspecified, the correct detection of both abs(fcor) and max (χ^2) statistics was approximately 1 or 1. In addition, the correct detection rates of both statistics was high under other conditions, too. AIC and BIC were successful to detect model misfit in the cases where the Q-matrix underspecified, whereas, they were failed detect model misfit for other cases. It can be said that the performance of BIC was mostly better than other relative model fit statistics to detect model misfit.

Key Words: Cognitive diagnostic assessment, item fit, model fit, misspecification of Q-matrix

DOI: 10.29329/ijpe.2018.157.12

----

[i] **Seçil Ömür Sünbül,** Assist. Prof. Dr., Mersin University, Measurement and Evaluation at Education, Mersin/Turkey.

[ii] **Semih Aşiret,** Mersin University, Science of Education, Measurement and Evaluation at Education, Mersin/Turkey.

Correspondence: semihasiret@gmail.com

Introduction

Cognitive Diagnostic Models (CDMs) are latent discrete models which enable to recognize the presence (mastery) or absence (nonmastery) of many skills or processes which are required to solve the problem in a test (de la Torre, 2009). Generally, CDM put on the attributes and latent skills which the examinee must have to respond the item correctly (DiBello, Roussos, & Stout, 2007; Rupp & Templin, 2008). Attributes refer to various latent characteristics such as cognitive processes, skills (Dimitrov & Atanasov, 2012).

The DINA Model (deterministic-input, nosiy-and-gate) is the most common and known model of CDM. The DINA model is a non-complementary and has a conjective condensation rule (Rupp, Templin, & Henson, 2010). In the DINA model, examinees are classified into two different mastery classes which the examinee masters the all attributes and examinee does not masters at least one required attribute for an item. Examinee is deemed sufficient if only he has mastered all attributes which are specified for an item in Q-matrix (Rupp, Templin, & Henson, 2010). However, it is assumed that the examinee will give wrong responses to the item when examinee has not mastered at least one required attribute.

Equation of DINA model has three basic components (Rupp, Templin, & Henson, 2010). The first component is latent variable (η_{ic}) which is defined for item i and examinees in latent class c. This latent variable indicates that whether examinees in latent class c have mastered ($\eta_{ic} = 1$) all attributes for item i or not ($\eta_{ic} = 0$)

$$\eta_{ic} = \prod_{k=1}^{K} \alpha_{ck}^{q_{ik}} \qquad (1)$$

The examinees with $\eta_{ic} = 1$ will give correct response to an item as long as there is no slipping. For this reason, the probability of giving correct response to the item equals to probability of not slipping $(1 - s_i)$. If $\eta_{ic} = 0$, examinees will give incorrect response to the item, therefore, the probability of giving correct response to the item will equal to guessing parameter. The probability that examinees in latent class c respond the item i correctly is given in Equation 2

$$\pi_{ic} = P(X_{ic} = 1|\eta_{ic}) = (1 - s_i)^{\eta_{ic}} g_i^{1-\eta_{ic}} \qquad (2)$$

If the examinee has mastered all required attribute for the item, $\eta_{ic} = 1$. Then, the probability that examinees in latent class c respond the item i correctly (π_{ic}) is given by $(1 - s_i)^{\eta_{ic}} g_i^{1-\eta_{ic}} = (1 - s_i)^1 g_i^{1-1}$=$(1 - s_i)$. If the examinee has not mastered at least one required attribute, $\eta_{ic} = 0$ and the probability that examinees in latent class c respond the item i correctly (π_{ic}) is given by $(1 - s_i)^{\eta_{ic}} g_i^{1-\eta_{ic}} = (1 - s_i)^0 g_i^{1-0}$=$g_i$.

Q-matrix and Misspecification of Q-matrix

In cognitive diagnostic assessment, one of the most important steps is specification of Q-matrix. Q-matrix associate the attributes and items in the test. (Li, 2016; Rupp & Templin, 2008; Tatsuoka, 1983). The diagnostic power of the CDM is based on the specification of Q-matrix which is supported empirically (Lee ve Sawaki, 2009). CDM could estimate latent attribute vectors for each examinee in observed data with specified Q-matrix. In Q-matrix composed of items listed in row and attributes listed in column. If test has *k* attributes and *i* items, Q-matrix consists of *i x k* 1-0 data. If item i requires attribute *k, i*. row and *k*. column of Q-matrix is 1, otherwise, it is 0. If it is known which attributes are required for each item and which attributes are mastered by examinees, responses of examinees to items are estimated.

The Q-matrix can be developed more accurately if attributes are well-defined and valid and the items are constructed along with these attributes. However, Q-matrix can be misspecified due to many different reasons such as underspecification, overspecification or both under and overspecification (mixed) of Q-matrix (Rupp & Templin, 2008). In underspecified Q-matrix, at least one 1 is inaccurately specified 0 in matrix whereas at least one 0 is specified 1 incorrectly under the overspecified Q-matrix. Estimation of parameters can be insufficient due to these misspecifications of Q-matrix. Furthermore, more than enough attributes could be specified in Q-matrix and these attributes are separated into many more detailed categories. Hence, it can be needed large sample size to estimate of item parameters. If required attributes are not specified in Q-matrix, it will cause to low score and fail to diagnose other attributes (Li, 2016).

Model-Data and Item Fit

One of the main problems in CDMs is model-data fit. Rupp and et al. (2010) stated that when model-data fit was weak, statistical inferences are not significant. Model-data fit in CDMs is evaluated in two different approaches; absolute and relative. Both relative and absolute model fit statistics were examined in this study.

Absolute model fit statistics evaluate the misfit between model and data. In this study, absolute value of the deviations of Fisher transformed correlations (abs(fcor)) and maximum of all χ^2 (max(χ^2)) as absolute model fit statistics were used. Equation of abs(fcor) statistic is given by

$$\left|Z\left[Corr(X_j, X_{j\prime}\right] - Z\left[Corr(\widehat{X_j}, \widehat{X_{j\prime}}\right]\right| \qquad (3)$$

where X_j is observed response of item j, $\widehat{X_j}$ is estimated response of item j, Corr; Pearson correlation coefficient and Z; Fisher transformation. As abs(fcor) value is close to 0, the model-data fit increases (Chen et al., 2013).

Max(χ^2) statistic is the maximum value of all χ^2 of all item pairs. Max(χ^2) is defined as:

$$\chi^2_{jj\prime} = \sum_{K=0}^{1} \sum_{k=0}^{1} \frac{(n_{jj\prime,ll\prime} - e_{jj\prime,ll\prime})^2}{e_{jj\prime,ll\prime}} \qquad (4)$$

where k is attribute kth, $n_{jj\prime,ll\prime}$ is observed frequency and $e_{jj\prime,ll\prime}$ is expected frequency.

Relative model fit statistics enable to select the model fits better to data. Although absolute model fit statistics are often preferred in CDM studies, relative model fit statistics should be used as the first step by eliminating possible models before conducting absolute model fit statistics. -2 log-likelihood, Akaike's information criterion (AIC) and Bayesian information criterion (BIC) relative model fit statistics were used in this study. -2 log-likelihood statistic is defined as -2LL=ln(ML). In equation ML refers to maximum likelihood value. Rupp et al. (2010) stated that the most commonly used relative fit statistics were AIC and BIC. The general equation for both AIC and BIC I as follows:

$$information\ criteria = -2ln(L) + ki \qquad (5)$$

In equation 5, L is log-likelihood of the model, i is the number of items and k is total number of structural parameters in model. Both statistics differ according to k. Always k=2 is for AIC so that AIC is given AIC = -2ln (L) + 2i. For BIC, k=ln(n) so that, BIC is calculated by BIC = -2ln (L) + ln (n)i. It is desired that the value is small for both AIC and BIC.

$S-\chi^2$ statistic developed by Orlando and Thissen (2000) was used as item fit statistic in this study. using χ^2statistic. The observed and expected responses obtained from the summed score are compared using χ^2statistic. $S-\chi^2$ statistic is computed as follows:

$$S - \chi^2_j = \sum_{s=1}^{I-1} N_s \frac{(O_{is} - E_{is})^2}{E_{is}(1 - E_{is})} \sim \chi^2(I - 1 - m) \qquad (6)$$

In this equation, i, denote the number of the items, s is group score, N_s is the number of the examinees in groups, O_{is} and E_{is} are observed and estimated responses of item i corresponding to group s.

There have been few research studies in the literature on item fit in CDMs Wang, Shu, Shang & Xu, 2015; Sorrel, Olea, Abad, de la Torre, & Barrada, 2017; Sinharay & Almond,2007; Oliveri & von Davier,2011; Kunina-Habenicht, Rupp, & Wilhelm, 2012; Choi, Templin, Cohen, & Atwood, 2010). Wang et al. (2015) noted that item fit analysis was not studied satisfactorily and they developed item fit statistic for DINA model. Classic fit index based on EM algorithm and PPMC (posterior predictive model checking) method based on MCMC estimation were evaluated for model fit in their studies. As a result of Wang et al.'s (2010) study, classic item fit index had higher fit detection rate than PPMC method. Sorrel et al. (2017) used inferential item fit statistics such as $S-\chi^2$,, likelihood ratio, Wald test and Lagrange multiplier in their study. In the study, tha factors item quality (0.40, 0.60 and 0.80), sample sizes (500, 1000) correlational structure (unidimensional, bidimensional), test length (12, 24, 36) and generated model (DINA, ACDM, DINO) were manipulated to evaluate the performances of the item fit statistics. The number of attributes is 4 and fixed. They concluded that $S-\chi^2$, statistic had sufficient Type I error rate but the power ratio was insufficient. Furthermore, it was stated that likelihood ratio and Wald test were more preferable than the LM test in terms of Type I error and power ratios. In addition, it was reported that all item fit statistics were influenced by item quality and the Type I error and power ratios of the item fit statistics were acceptable with few exceptions only if item quality was high. Sinharay and Almond (2007) used Bayesian residual plots and chi-square statistics to evaluate item fit statistics in their study using real data. They pointed out that Bayesian residual plots were simple but powerful to detect model-data misfit, beside this, item fit plots were quite good at detecting misfit items. Oliveri and von Davier (2011) used RMSEA statistic to evaluate the item fit by fitting PISA data to general diagnostic models (GDM). Kunina-Habenicht et al. (2012) examined the type I error and power ratios of MAD and RMSEA item fit statistics by conducting simulation study. As a result of Kunina-Habenicht et al.'s (2012) study, classification accuracy was significantly reduced when the Q-matrix was incorrectly specified. Furthermore, they concluded that item fit statistics were more sensitive in overspecification of Q-matrix than with underspecification of Q-matrix and AIC and BIC relative fit statistics were sufficiently sensitive in both overspecification and underspecification of Q-matrix.

In this study, it is aimed to evaluate the effect of various sample sizes, percentage of misfit items in the test and item discrimination levels on item and model fit with misspecification of Q-matrix. Investigating the relevant literature, it was studied that the effect of sample size and number of misfit items in the test on the performance of item fit (Wang, Shu, Shang, & Xu, 2015; Lai, Gierl, Cui, & Babenko, 2017; de la Torre & Lee, 2013) and model fit (Chen, de la Torre, & Zhang, 2013; Hu, Miller, Huggins-Manley, & Chen, 2016; Galeshi & Skaggs, 2014; Kunina-Habenicht, Rupp, & Wilhelm, 2012; Liu, Tian, & Xin, 2016) statistics in case of misspecification of Q-matrix. It is foreseen that item quality (item discrimination) could affect the performance of item and model fit statistics in case of misspecification of Q-matrix. Therefore, both different factor levels were considered and the item quality was included and manipulated and the effects of these factors on item and model fit statistics were evaluated together. Accordingly, it is expected that this study will contribute to the field.

Method

Simulation Design

Sample Size (N): Sample sizes of 1000, 2000 and 4000 were used in this study.

Number of Attributes and Items: Number of attributes was fixed at 4 and number of items was set at 15 accordingly.

Levels of s and g parameters: In the study, s and g parameters were generated as item quality was low and high. While g and s parameters were generated from uniform distribution U(0.10, 0.20) and U(0.10, 0.40) for low quality items, both g and s parameters were generated from uniform distribution U(0.05, 0.10) for high quality items.

Percentage of misfit items: The percentage of misfit item was set at %20 and %40

Misspecification of Q-matrix: Misfit in CDM is mostly due to the nature of the attributes, construct of the attribute, Q-matrix or selected cognitive diagnostic model (Chen, de la Torre and Zhang, 2013). In this study, only misfit source due to Q-matrix misspecification was examined. Specification of Q-matrix mostly criticized because of subjective (Rupp and Templin, 2008). Therefore, Q-matrix misspecification is one of the possible misfit sources. In the study, Q-matrix was misspecified in different three ways: underspecification, overspecification and mixed. Correctly specified Q-matrix, misspecified Q-matrix and misfit items used in this study are presented in Table 1

Table 1: Correctly specified Q-matrix, misspecified Q-matrix and misfit items used in this study

	Specified				Underspecified				Overspecified				Mixed			
Item	k1	k2	k3	k4	k1	k2	k3	k4	k1	k2	k3	k4	k1	k2	k3	k4
1	1	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0
2	0	1	0	0	0	1	0	0	0	1	0	0	0	1	0	0
3	0	0	1	0	0	0	1	0	0	0	1	0	0	0	1	0
4	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	1
5	1	1	0	0	1	0	0	0	1	1	1	0	0	1	1	0
6	1	0	1	0	0	0	1	0	1	1	1	0	0	1	1	0
7	1	0	0	1	1	0	0	0	1	1	0	1	1	0	1	0
8	0	1	1	0	0	1	0	0	0	1	1	1	0	1	0	1
9	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1
10	0	0	1	1	0	0	0	1	0	1	1	1	0	1	0	1
11	1	1	1	0	1	1	1	0	1	1	1	0	1	1	1	0
12	1	1	0	1	1	1	0	1	1	1	0	1	1	1	0	1
13	1	0	1	1	1	0	1	1	1	1	1	1	0	1	1	1
14	0	1	1	1	0	1	0	1	1	1	1	1	1	0	0	1
15	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

When Q-matrix was underspecified and if the percentage of misfit item was %20, one attribute of 5.,10. and 14. items and if the percentage of misfit item was %40, one attribute of 5., 7., 8., 10., 13., and 14. items were transformed from 1 to 0. When Q-matrix was overspecified, one attribute of same items was transformed from 0 to 1. When Q-matrix was both underspecified and overspecified, one attribute was translated from1 to 0 and another attribute was translated from 0 to 1 but number of measured attributes didn't change.

Manipulated factors and their levels in this study are presented in Table 2.

Table 2. Manipulated factors and their levels in this study

Factor	Factor levels	Values
Sample size	3	1000,2000,4000
item quality	2	Low and high
Percentage of misfit items	2	%20, %40
Q-matrix misspecification	3	Underspecification, overspecification, mixed

Data Generation

In this study, it was aimed to evaluate the effect of different sample sizes, percentage of misfit items in the test and level of item discrimination (item quality) on item and model fit in case of misspecification of Q-matrix in cognitive diagnostic models. Data were generated in accordance with the DINA model and the Q-matrix were defined for 4 attributes and 15 items. In data generation, g and s parameters were manipulated to produce low and high-quality items. For low quality items, g and s parameter were generated from U(0.10, 0.20) and U(0.10, 0.40) uniform distribution and for high quality items, both g and s parameters were generated from U(0.05, 0.10) uniform distribution. In the study, sample sizes of 1000, 2000 and 4000 were used. Q-matrix was misspecified in three different way (underspecification, overspecification and mixed) by considering the percentage of misfit items (%20(3 items) and %40 (6 items))

Data Analysis

In the study, $S-\chi^2$ was used as item-fit statistic due to having better performance than other statistic in previous studies. Correct detection rate for related items was calculated to evaluate the performance of model-fit and item-fit statistics. To calculate the correct detection rate, when number of misfit item was three, if p value of misfit items was less than 0.05, the value of detection was assigned 0, if it was larger than 0.05, it was assigned 1. This process replicated 100 times and value of detection was summed for each replication. Lastly, it was averaged of total value of detection. Similar process was applied when misfit item was six

Absolute and relative model-data fit statistics was used to evaluate the fit of generated data to model. In the study, maximum of all χ^2 (max (χ^2)) and absolute value of the deviations of Fisher transformed correlations abs(fcor) statistics were used as absolute model-fit statistics and -2loglike, AIC and BIC were used as a relative model fit statistic. The correct detection rate was calculated to evaluate model fit. When Q-matrix was misspecified, if the p value of max (χ^2) and abs(fcor) statistics was less than 0.05, it was assigned 0 to detection value, if it was larger than 0.05, it was assigned 1 to detection value for 100 replication and in each replication, detection value was added total detection value. Lastly, it was averaged of total detection value. In calculation of relative model fit, -2LL, AIC and BIC fit statistics of GDINA and DINA model were calculated and compared to each other.

Findings

Evaluation of Item Fit Statistic for Q-matrix Misspecification

The results of correct detection rates of $S-\chi^2$s statistic for misspecified Q-matrix are shown in Table 3.

Table 3: Correct detection rates of $\mathbf{S}-\boldsymbol{\chi}^2$ statistic for Q-matrix misspecification.

			3 item			6 item		
Statistic	IQ	N	u	o	m	u	o	m
$S-\chi^2$	LQ	1000	0,15	0,15	0,25	0,08	0,07	0,09
		2000	0,36	0,39	0,58	0,11	0,08	0,12
		4000	0,68	0,71	0,89	0,23	0,13	0,22
	HQ	1000	0,90	0,87	0,99	0,42	0,39	0,75
		2000	1,00	1,00	1,00	0,72	0,61	0,92
		4000	1,00	1,00	1,00	0,90	0,80	0,99

Note. N = sample size; IQ = Item Quality; LQ = Low Quality; HQ = High Quality; u = under-specified; o = over-specified; m = mixed

Investigating Table 3 shows that as sample size increased, the correct detection rate of $S-\chi^2$ statistic increased across all conditions. In addition, It can be seen in Table 3 that S-X2 had larger correct detection rate when the item quality was higher than when the item quality was low for all Q-matrix misspecification (underspecification, overspecification, mixed). This difference was extremely too much when sample size was small. As the percentage of misfit items in the test increased, the correct detection rate of $S-\chi^2$ statistic decreased. As is shown in Table 3, was when sample size was 1000, item quality was low, percentage of misfit items was %40 and Q-matrix was overspecified, $S-\chi^2$ had the lowest correct detection rate. Furthermore, it can be said that when item quality was high, $S-\chi^2$ had the lowest correct detection rate with overspecified Q-matrix. Correct detection rate with mixed misspecified Q-matrix was higher than When Q-matrix was under or over misspecified in all conditions.

Evaluation of Model-Data Fit for Q-matrix Misspecification

The results of correct detection rates of model-data fit of absolute model fit statistics for misspecified Q-matrix are presented in Table 4.

According to Table 4, as the percentage of misfit item increased, the correct detection rate of both max((χ^2) and abs(fcor) statistics increased. Furthermore, max ((χ^2) statistics had higher correct detection rates than abs(fcor) statistic in most cases. It can be seen in Table 4 that when the percentage of misfit items was %40, max((χ^2) statistic detected model misfit correctly almost in all conditions. When the percentage of misfit items was 20 and item quality was low with under and over misspecified Q-matrix, as sample size increased, correct detection rates of both of two statistics increased. However, when item quality was high and sample size was 1000, the correct detection rate decreased. When Q-matrix was mixed misspecified, correct detection rate of max((χ^2) and abs(fcor) was approximately 1 or 1.

Table 4: Correct detection rates of absolute model fit statistics for Q-matrix misspecification.

			3 item			6 item		
Model	IQ	N	u	o	m	u	o	m
MAX(χ^2)	LQ	1000	0,87	0,94	1,00	1,00	1,00	1,00
		2000	0,94	0,97	1,00	1,00	1,00	1,00
		4000	0,96	0,98	1,00	1,00	1,00	1,00
	HQ	1000	1,00	1,00	1,00	1,00	1,00	1,00
		2000	0,91	0,96	1,00	1,00	1,00	1,00
		4000	0,95	0,98	1,00	1,00	1,00	1,00
ABS(fcor)	LQ	1000	0,72	0,84	0,99	0,98	0,96	1,00
		2000	0,83	0,92	1,00	0,99	0,98	1,00
		4000	0,89	0,94	1,00	0,99	0,99	1,00
	HQ	1000	1,00	1,00	1,00	1,00	1,00	1,00
		2000	0,81	0,89	0,99	0,99	0,97	1,00
		4000	0,86	0,93	1,00	0,99	0,98	1,00

Note. N = sample size; IQ = Item Quality; LQ = Low Quality; HQ = High Quality; u = under-specified; o = over-specified; m = mixed; MAX(χ^2) = maximum χ^2; ABS (fcor) = Absolute deviation of Fisher transformed item pair correlation

The results of correct detection rates of model-data fit of relative model fit statistics for misspecified Q-matrix are presented in Table 5.

Table 5: Correct detection rates of relative model fit statistics for Q-matrix misspecification.

			3 items			6 items		
Model	IQ	N	u	o	m	u	o	m
-2LL	LQ	1000	0,00	0,00	0,00	0,00	0,00	0,00
		2000	0,00	0,00	0,00	0,06	0,00	0,00
		4000	0,01	0,00	0,00	0,28	0,00	0,00
	HQ	1000	0,16	0,00	0,00	0,52	0,00	0,00
		2000	0,67	0,00	0,00	0,64	0,00	0,00
		4000	0,97	0,00	0,00	0,79	0,00	0,00
AIC	LQ	1000	1,00	0,00	0,00	0,98	0,00	0,00
		2000	0,90	0,00	0,00	0,95	0,00	0,00
		4000	0,90	0,00	0,00	0,96	0,00	0,00
	HQ	1000	0,99	0,00	0,00	0,90	0,00	0,00
		2000	0,98	0,00	0,00	0,86	0,00	0,00
		4000	0,98	0,00	0,00	0,90	0,00	0,00
BIC	LQ	1000	1,00	1,00	1,00	1,00	1,00	0,83
		2000	1,00	1,00	0,77	1,00	0,02	0,00
		4000	1,00	0,00	0,00	1,00	0,00	0,00
	HQ	1000	1,00	0,00	0,02	1,00	0,00	0,00
		2000	1,00	0,00	0,00	1,00	0,00	0,00
		4000	1,00	0,00	0,00	1,00	0,00	0,00

Note. N = sample size; IQ = Item Quality; LQ = Low Quality; HQ = High Quality; u = under-specified; o = over-specified; m = mixed; AIC = Akaike's information criterion; BIC = Bayesian information criterion; -2LL = *-2 log-likelihood*

Investigating Table 5, AIC and BIC statistics detected relative misfit at high rates in all conditions when Q-matrix was underspecified. The correct detection rates of-2LL statistic was approximately 0 or 0 when item quality was low with underspecified Q-matrix whereas the correct detection rate of -2LL increased as the sample size increased when item quality was high with underspecified Q-matrix. However, these rates were lower compared to detection rates of AIC and BIC. It can be seen in Table 5 that when Q-matrix was underspecified, BIC statistic had the highest detection rates. As is shown in Table 5, correct detection rates of AIC and BIC statistic was approximately 0and 0 when Q-matrix was over or mixed specified. It is clear from the Table 5 when item quality was low and sample size was 1000, the detection rate of misfit model for BIC was high.

Discussion, Conclusion and Recommendation

In this study, it was aimed that investigating the effect of factors such as sample sizes, percentage of misfit items in the test and item quality, on performance of model-data and item fit statistics with Q-matrix misspecification. For this purpose, simulation study was conducted and results of this simulation study were analyzed.

The main condition in this study is Q-matrix misspecification. Specification of Q-matrix is one of the most important steps in cognitive diagnostic assessment. However, specification of Q-matrix is subjective so that it can be specified in many different ways. In the study, when Q-matrix was misspecified, the effects of various factors on item fit was examined first. $S-\chi^2$ was used as a item-fit statistic in this study. The correct detection rate of misfit items for $S-\chi^2$ was high as sample size and item quality increased and percentage of misfit items in the test decreased. Similarly, Sorel and et al. (2017) reported that Type I error of $S-\chi^2$ was sufficient however power of this statistic was poor. These results are identical to the findings from this study. $S-\chi^2$ statistic had the lowest correct detection rate when Q-matrix was over misspecified and item quality was high. This may be due to the fact that it is easier to detect misfit items because of small number of estimated parameters in cases where Q-matrix was under or mixed misspecified.

In the second part of the study, the effect of the factors in the study on model data fit was investigated when Q-matrix was misspecified. For this purpose, it was examined both relative and absolute model-data fit. The AIC, BIC and -2LL statistics were used to determine the relative model-to-data fit. When the Q-matrix was underspecified, the AIC and BIC statistics detected almost all of the relative model data misfit correctly, however these statistics tended to select GDINA when Q-matrix was over or mixed misspecified along with especially high item quality and large sample size. Chen et al. (2013) pointed out that) AIC and BIC statistics tend to select the saturated model between saturated model (i.e., the G-DINA) and a misspecified reduced model (i.e., the DINA) regardless of which Q-matrix was used. . In this regard, this result is consistent with Chen et al. (2013). Hu et al. (2016) suggested using the AIC and BIC statistics for the DINA model in case of Q-matrix underspecification. Furthermore, Hu et. al (2016) reported that when Q-matrix was underspecification and also overspecification along with small sample size and small number of misfit items in the test, BIC had high correct detection rates. Conversely, when the Q-matrix was overspecified along with large sample size and the number of misfit items i, the BIC statistic could not almost detect the correct model. The results from the studies of the results from the studies of are parallel with the results of this study. Similarly, when Q-matrix was specified correctly, the relative model fit statistics had high detection rate in both studies. Unlike other studies, the factor used in this study is the item quality. According to the findings of this study, as item quality increased, correct detection rates of relative model fit statistics for DINA model decreased. Galeshi and Skaggs (2014) stated that detection rate of AIC and BIC statistics increased as sample size increased for CRUM model. However, in this study, there was no significant change in performance of AIC and BIC statistics as the sample size increased. In addition, Galeshi and Skaggs (2014) concluded that AIC and BIC had a similar performance however when Q-matrix was overspecified with small sample size BIC had more accurate results. The findings from studies of Galeshi and Skaggs (2014) are similar to findings of this study.

In the last part of study, Absolute model-data fit was examined. Abs(fcor and max(χ^2) were used as absolute model fit statistics. In the study, as the number of misfit items increased, the correct detection rates of both statistics was increased. Hu et al. (2016) stated that when Q-matrix was specified correctly, both statistics selected the correct model (DINA). Similar to this finding of Hu et al. (2016), both statistics selected the correct model (DINA) in this study. Moreover, Hu et al. (2016) abs(fcor) and max(χ^2) statistics mostly detected the misspecified Q-matrix except for the small samples. Similarly, in this stud, when item quality was low, correct detection rate of both statistics was high and increased as sample size increased. When item quality was high, the correct detection rates of both statistics higher across all sample size.

Ultimately, it can be said that $S-\chi^2$ succeeds in detecting misfit items when item quality is high, sample size is large and the number of misfit items are small. One conclusion from this study which is consistent with other research is that AIC and BIC, relative model fit statistics, detect the misfit when Q-matrix is underspecification and they fail to detect misfit when Q-matrix is mixed or overspecification. Absolute model fit statistics are successful to detect misspecified Q-matrix. These statistics are more successful when sample size is large and numbers of misfit items is much. However max(χ^2) had higher detection rates than abs(fcor) in almost all conditions. Therefore, max(χ^2) is more preferable than abs(fcor).

In this study, some factors such as sample size, item quality and percentage of misfit item were used at various levels to examine the effect of these factors on the model-data and item fit with Q-matrix misspecification. The same study could be repeated at different levels for these factors. Moreover, different factors, such as correlation between attributes, test length and number of attributes could be included in further studies. In this study, only the $S-\chi^2$ was used as item-fit statistic. Similar studies with different item fit statistics can be conducted. Beside this This study was constrained with DINA model. In further studies, different cognitive diagnostic models can be evaluated with CDM and Q-matrix misspecification. There are not enough studies related to the effect of item quality on model-data and item fit in the literature. Therefore, it is recommended that same study can be repeated by using different item parameters or including different factors. Simulated data was generated for this study. Same study could be conducted with real data.

References

Chen, J., de la Torre, J., & Zhang, Z. (2013). Relative and absolute fit evaluation in cognitive diagnosis modeling. Journal of Educational Measurement, 50, 123-140.

Choi, H.-J., Templin, J. L., Cohen, A. S., & Atwood, C. H. (2010, April). The impact of model misspecification on estimation accuracy in diagnostic classification models. Paper presented at the meeting of the National Council on Measurement in Education (NCME), Denver, CO.

De La Torre, J. (2009). A cognitive diagnosis model for cognitively-based multiple-choice options. Applied Psychological Measurement, 33, 163–183.

De La Torre, J., & Lee, Y. S. (2013). Evaluating the Wald test for item-level comparison of saturated and reduced models in cognitive diagnosis. Journal of Educational Measurement, 50, 355-373.

DiBello, L., Roussos, L. A., & Stout, W. F. (2007). Review of Cognitively Diagnostic Assessment and a Summary of Psychometric Models. In C. R. Rao & S. Sinharay (Eds.), Handbook of Statistics, 26, 979-1030.

Galeshi & Skaggs (2014). Traditional fit indices utility in new psychometric model: cognitive diagnostic model. International Journal of Quantitative Research in Education, 2, 2.

Kunina-Habenicht, O., Rupp, A. A., & Wilhelm, O. (2012). The impact of model misspecification on parameter estimation and item-fit assessment in log-linear diagnostic classification models. Journal of Educational Measurement, 49, 59-81.

Lee, Y.W., & Sawaki, Y. (2009). Cognitive diagnostic approaches to language assessment: An overview. Language Assessment Quarterly, 6(3), 172-189. doi: 10.1080/15434300902985108

Li, H. (2016). Estimation of Q-matrix for DINA Model Using the Constrained Generalized DINA Framework. (doctoral dissertation). Coulumbia University.

Liu, Y., Tian, W., & Xin, T. (2016). An application of M2 statistic to evaluate the fit of cognitive diagnostic models. Journal of Educational and Behavioral Statistics, 41, 3-26.

Oliveri, M. E., & von Davier, M. (2011). Investigation of model fit and score scale comparability in international assessment. Psychological Test and Assessment Modeling, 53, 315-333.

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous Item Response Theory models. Applied Psychological Measurement, 24(1), 50–64.

Robitzsch, A., Kiefer, T., George, A. C., & Uenlue, A. (2015). CDM: Cognitive Diagnosis Modeling (R package Version 6.4-23). Retrieved from http://CRAN.R-project.org/package=CDM

Rupp, A. A., & Templin, J. L. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. Measurement,6(4), 219-262.

Rupp, A. A., Henson, R. A., & Templin, J. L. (2010) Diagnostic measurement : theory, methods, and applications. Guilford Press

Sinharay, S., & Almond, R. G. (2007). Assessing fit of cognitive diagnostic models: A case study. Educational and Psychological Measurement, 67, 239-257.

Sorrel, Abad, Olea, de la Torre, & Barrada. (2017). Inferential Item-Fit Evaluation in Cognitive Diagnosis Modeling. Applied Psychological Measurement, 2017, 41, 614–631

Sorrel, M. A., Olea, J., Abad, F. J., de la Torre, J., Aguado, D., & Lievens, F. (2016). Validity and reliability of situational judgement test scores: A new approach based on cognitive diagnosis models. Organizational Research Methods, 19, 506-532.

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. Journal of Educational Measurement, 20, 345-354.

Wang, C., Shu, Z., Shang, Z., & Xu, G. (2015). Assessing item level fit for the DINA model. Applied Psychological Measurement, 39, 525-538.