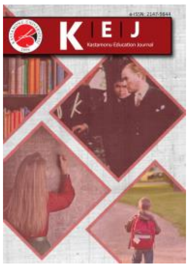
Kastamonu Education Journal, 2021, Vol. 29, No:5, 1024-1033
doi: 10.24106/kefdergi.836241



| Research Article / Araştırma Makalesi |

Impact of Aberrant Responses on Item Response Theory-Based Model Estimations

Normal Olmayan Yanıtların Madde Tepki Kuramına Dayalı Model Kestirimleri Üzerindeki Etkisi

Akif Avcu[1]

Keywords

1. Person-fit
2. Item response theory
3. University students
4. Generalized anxiety
5. R program

Anahtar Kelimeler

1. Kişi uyumu
2. Madde tepki kuramı
3. Üniversite öğrencileri
4. Yaygın kaygı
5. R programı

Received/Başvuru Tarihi
05.12.2020

Accepted / Kabul Tarihi
04.06.2021

Abstract

Purpose: Misfit individuals can bias model results at both the tests and item levels. Given the importance of detecting aberrant responses, the purpose of this study was to examine the effect of aberrant responses on item response theory-based model estimates.

Design/Methodology/Approach: This study is descriptive research, and simulated data was used. For this purpose, data were collected from 1104 university students enrolled in 8 different universities in Turkey using the Generalized Anxiety Disorder-7 scale. After parameter estimation based on the item response theory model, 100 different datasets were simulated using the item and person parameters obtained from these estimations. In this way, it was aimed at increasing the findings' generalizability. The R program analyzed the data using "PerFit" and "mirt" packages. Misfit persons were identified with Lz, U3, G and norm-based G person fit statistics.

Findings: The findings showed that misfit persons affected the model fit statistics, item fit statistics, item discrimination values, the amount of information provided by the items, the total amount of information provided by the scale, and empirical reliability levels across different levels of ability trait. In addition, in order to improve the results based on the item response theory, it was observed that removing the misfit persons detected based on the Lz statistic from the dataset was the least effective among the existing techniques. On the other hand, the G-fit statistic has been identified as the most effective technique.

Highlights: The obtained results should be interpreted with caution because the simulated data used in this study is based on parameters representing the dataset collected with a measurement tool aimed at measuring anxiety, and these results may not be generalizable to the measurement of different traits.

Öz

Çalışmanın amacı: Uyumsuz bireyler, model sonuçlarını hem test hem de madde düzeyinde bozabilir. Anormal yanıtların tespit edilmesinin önemi göz önünde alındığında, gerçekleştirilen bu çalışmanın amacı anormal yanıtın madde madde tepki kuramına dayalı kestirimler üzerindeki etkisinin incelenmesi olarak belirlenmiştir.

Materyal ve Yöntem: Gerçekleştirilen bu çalışma betimsel araştırmadır ve türetilmiş veriler kullanılmıştır. Bu amaçla Türkiye genelinde 8 farklı üniversiteye kayıtlı 1104 üniversite öğrencisinden Yaygın Kaygı Bozukluğu-7 Ölçeği kullanılarak veriler toplanmış ve madde tepki kuramı modeline dayalı parametre kestirimleri gerçekleştirildikten sonra elde edilen madde ve kişi parametreleri kullanılarak 100 adet veri seti türetilmiştir. Bu sayede elde edilen bulguların genellenebilirliğinin arttırılması amaçlanmıştır. Veriler R ortamında ”lavaan”, “perfit” ve “mirt” paketleri kullanılarak analiz edilmiştir. Uyumsuz kişiler Lz, U3, G ve norma dayalı G kişi uyumu istatistikleri ile tespit edilmiştir.

Bulgular: Elde edilen bulgular, uyumsuz kişilerin model uyumu istatistikleri, madde uyumu istatistikleri, madde ayırt edicilik değerleri, maddeler tarafından sağlanan bilgi miktarı, ölçeğin verdiği toplam bilgi miktarı ve kaygı özelliğinin farklı düzeyleri boyunca görgül güvenilirlik düzeyi üzerinde etkisi olduğunu göstermiştir. Ayrıca, madde tepki kuramına dayalı sonuçları iyileştirmek için lzpoly istatistiğine dayalı belirlenen uyum göstermeyen bireyleri veri setinden uzaklaştırmanın mevcut teknikler içerisinde en az etkilisi olduğu görülmüştür. Diğer taraftan, G istatistiği ise en etkilisi olarak belirlenmiştir.

Önemli Vurgular: Elde edilen sonuçlar dikkatle yorumlanmalıdır çünkü gerçekleştirilen bu çalışmada kullanılan türetilmiş veriler kaygının ölçümünü amaçlayan bir ölçüm aracı ile elde edilen veri setini temsil eden parametrelere dayalıdır ve elde edilen sonuçlar farklı özelliklerin ölçümüne genellenemeyebilir.

[1] **Corresponding Author,** Marmara University, Atatürk Faculty of Education, Educational Sciences Department, İstanbul, TURKEY; avcuakif@gmail.com, https://orcid.org/0000-0003-1977-7592

Citation/Alıntı: Avcu, A. (2021). Impact of Aberrant Responses on Item Response Theory-Based Model Estimations, Kastamonu Education Journal, 29(5), 1024-1033. doi: 10.24106/kefdergi.836241

INTRODUCTION

For any measurement instrument, validity is an essential concern for sound interpretation of test results and proper use of those results (American Educational Research Association, 1985). As Messick (1995) puts it, the validity of test scores can be examined at both the score level and the individual level because test scores are a function of the items or stimuli given and a function of respondents. When factors other than the latent trait being measured affect the response process, the subject's response behavior becomes abnormal (also referred to as aberrant or unexpected), and the resulting test score does not adequately reflect the level of the latent trait. This can lead to biased research results and erroneous decisions about individuals. Reasons such as the respondents' level of motivation, poor understanding of the instructions, inattentive reading of the items, inability to respond sincerely, and ignoring some response categories cause respondents to produce response patterns that are inconsistent with the underlying model of the trait (Meijer, 1996).

It is essential to identify individuals who do not fit the underlying model or give strange responses compared to the rest of the sample group, as the proper level of the latent trait may not be accurately estimated. Inaccurate trait estimation negatively affects individuals and organizations and leads to erroneous conclusions about test validity (Schmitt et al., 1999). Examining test score validity at the individual level could be done through a person-fit analysis. Person-fit assessment is concerned with identifying atypical test performance based on the item or test score patterns (Meijer & Sijtsma, 2001), and it is conducted by using one or more person-fit statistics. Currently, about forty person-fit statistics have already been developed in the relevant literature, although only four of them are designed explicitly for Likert-type polytomous items. For a detailed overview of these person-fit statistics, the reader is advised to look at Karabatsos (2003).

Various person-fit statistics can be divided into two broad families: group-based and IRT-based. Most group-based person-fit statistics are based on the Guttman model (Guttman, 1944), in which items are classified from easiest to the most complex and any inconsistency in the response vector is counted as a Guttman error. More specifically, on a dichotomously scored achievement test, a Guttman error occurs when a person gives a correct response to a difficult item while giving an incorrect response to a relatively easy item. In a test, if a person answers the easiest x items correctly while answering the remaining items incorrectly, this pattern is considered a perfect Guttman scaling with no error. On the other hand, answering a more difficult item correctly and answering an easier item incorrectly is considered a Guttman error. For example, on a five-item achievement test where items are ordered by difficulty, a response pattern of [1, 1, 1, 0, 0] contains zero errors, while a response pattern of [1, 0, 0, 1, 1] contains four Guttman errors because it contains four (0, 1) item pairs.

On the other hand, polytomous item formats are usually used in psychological assessment instruments. For polytomous items, the concept of item steps is used instead of difficulty (Sijtsma & Molenaar, 2002). Item level is the psychological threshold between ordered response options. For example, in a Likert-type item, the threshold between the "I disagree completely" option and the "I disagree" option is the first step for a respondent. If the person feels that they do not fully agree with this statement, they will cross the first threshold and choose between "I disagree" and "I have no idea." This process will continue until the respondent does not choose the next option. Using this approach, the proportion of people passing each step for any item can be determined by the proportion of respondents answering correctly in the previous example. Later, the steps can be ordered from lowest to highest popularity, and the polytomous item responses for each respondent can be compared to the order of the item steps. Taking a less popular item step will result in a Guttman error (Molenaar, 1997).

One of the fit statistics used based on Guttman error is G statistics. Polytomous extension of G statistics is named G-poly. G statistics reflect the number of Guttman errors. The minimum possible value for the G statistic is 0. This indicates that the Guttman error has not been observed, and the score vector follows perfect Guttman scaling. On the other hand, the maximum value that G can take varies depending on the number of items and the number of item categories.

For this reason, it is not possible to compare the G-poly values of different tests if they contain a different number of items or items' response categories are not the same. The normed G statistics, on the other hand, enable this comparison. Derived from G-poly statistics, the standardized version can take values between 0 and 1. The polytomous version of the normed G statistic is denoted as Gnormed-poly (Emons, 2008). U3 statistic (Van Der Flier, 1982) is also based on Guttman errors. It takes into account both item difficulty order and values of item difficulties. Accordingly, if the response vector is perfectly Guttman vector, the U3 statistic takes the value of zero, whereas if the Guttman vector is perfectly inverse, this value becomes one. In other words, increasing U3 values provide stronger indicators of a person's misfit. Later, Emons (2008) generalized the U3 statistics (Van Der Flier, 1982) to polytomous items (denoted as U3-poly). U3-poly values can also vary between 0 and 1.

In contrast to the U3 statistic, the increase in value indicates that the level of fit is also increasing. Therefore, increasing U3-poly values provide stronger indications of person fit. Among the IRT-based person-fit statistics, the Lz value comes to the fore regarding its applicability to polytomous items. The Lz statistics is a standardized likelihood-based statistic used to determine the likelihood of an item response pattern in the context of the selected IRT model. The low Lz values indicate a stronger misfit (Drasgow et al., 1985).

Although many studies on the person fit have been conducted to date, including those using simulative data (Karabatsos, 2003; Tendeiro & Meijer, 2014), recent studies are also increasingly using real data (Engelhard, 2009; Conrad, 2010). Considering that studies on understanding the effects of person fit on practical test results provide valuable information for researchers and test developers in education and psychology, this study aimed to investigate the effect of aberrant responses on item response theory (IRT) based model estimates. More specifically, this study aims to show how IRT-based model fit statistics, item fit statistics, item discrimination values, the amount of information provided by items, the total amount of information provided by the scale, and empirical reliability levels changed across different levels of the anxiety trait when misfit persons were removed from the dataset. This study is essential because item response theory-based estimates have been becoming common among researchers, and the results of this study should help researchers better identify the effect of aberrant responses and shed light on which person fit index should be preferred while researchers estimate model parameters.

METHOD/MATERIALS

The purpose of this study is to show how misfit persons alter IRT-based model estimates. Since the study aims to reveal the existing situation without manipulating the conditions or showing the relationships between different variables, it carries the specifications of a descriptive research design (Karasar, 2005).

Participants and procedure

A large representative data was collected in this study. The sample group consists of 1104 university students. The participants were selected from 8 different universities. Seven of these universities are public, and one is private, and all universities are located in two different major cities. Of the participants, 808 (74%) are females, and the remaining 284 (26%) are males. The distribution of participants based on the faculty in which they were enrolled as follows: Faculty of Education 557 (51%), Faculty of Business and Administrative Sciences 149 (13.7%), Faculty of Dentistry 118 (10.8%), Faculty of Science and Literature 108 (9.8%), Faculty of Fine Arts 106 (9.7%) and Faculty of Health Sciences 54 (4.9%). Data was collected through an online platform due to the 2020 pandemic outbreak. Convenience sampling was used due to the difficulty of using systematic sampling techniques in online data collection. Participants were informed of the purpose of the study before they began participating and were informed that participation in the study was voluntary. Written consent was obtained before they began answering the questions.

Measures

Generalized Anxiety Disorder-7 Scale (GAD-7) is a seven-item self-report measurement instrument developed by Spitzer et al. (Spitzer et al., 2006). It was developed to evaluate the General Anxiety Disorder based on DSM-IV (American Psychiatric Association, 2000) criteria. It is scored on four-point Likert scale (0 = none, 1 = many days, 2 = more than half the days, 3 = almost every day).

GAD-7 questions ask respondents to assess their experiences in the last 2 weeks. The cut-off points of 5, 10, and 15 correspond to mild, moderate and severe anxiety, respectively. Patients with a total score of 10 or above are recommended to be closely screened further for possible diagnosis of GAD. The adaptation study of the scale to Turkish was conducted in 2013 by Konkan et al. (Konkan et al., 2013). In the adaptation study, the Cronbach's alpha value for GAD-7 total score was 0.852. In addition, validity-related evidences were provided and it was concluded that GAD-7 is a reliable and valid assessment tool (($\chi2$=14.48, $p>0.05$, $\chi2$/df=1.03, CFI=0.99, TLI=0,99, GFI=0.96, RMSEA=0.02 and AGFI=0.93).

Statistical analysis

Firstly, analyses were carried out using a mirt package (Chalmers, 2012) to examine which polytomous IRT model fits the data better. Graded Response Model (GRM) (Samejima, 1970) and Generalized Partial Credit Model (GPCM: Muraki, 1992) were compared with likelihood-based ratio test. The results suggested that GAD-7 data showed a significantly better fit with the GRM model ($p<0.01$), and further analysis were decided to be continued with the GRM. This preference is also in line with previous IRT-based analyses of GAD-7 (Jordan et al., 2017). In the next stage, simulative datasets were generated 100 times based on the estimated person (traditionally denoted as θ) and item parameters obtained after fitting the GRM model, and further analyses were carried out by using those datasets to increase the generalizability of the results. The expected a-posteriori (EAP) (Bock & Aitkin, 1981) estimation method was used to estimate θ parameters. This method is based on the Bayesian statistical approach. Among other alternatives, the EAP was selected because it has no inherent problems like non-convergence and dependence on starting value. In addition, estimation is not affected by the existence of the highest and the lowest possible scores in the dataset.

Firstly, it was examined whether GAD -7 data met the assumptions of IRT analysis. There are two basic assumptions for parametric IRT models: unidimensionality and local independence (Hambleton, 1991). Unidimensionality implies that only one latent feature underlies a group of items. The existence of an underlying dominant factor is sufficient to meet this assumption. The ratio of the first two eigenvalues obtained from the exploratory factor analysis (EFA) is the first criteria for evaluating the unidimensionality of GAD -7 data. Morizot et al. (2009) stated that having a ratio of 3 and above is sufficient evidence for unidimensionality. The acceptable level of fit by confirmatory factor analysis (CFA) model testing unidimensional structure is another criterion used. Model fit was assessed with $\chi2$/df, Comparative Fit Index (CFI), Tucker-Lewis Index (TLI), Goodness-of-fit index (GFI), Normed Fit Index (NFI), Root Mean Square Error of Approximation (RMSEA), Standardized Root Mean Square Residual (SRMR). These fit indices are interpreted as suggested by Kline (2015). Explanatory factor analysis was performed with IBM SPSS Statistics for Windows (Version 21), while the lavaan package (Rosseel, 2012), which is available in the R software environment (R Core Team, 2020), was used for the CFA.

On the other hand, local independence implies that the probability of the response of a person taking the test is not to be affected by the responses given to other items in the test. Q3 statistics proposed by Yen (1984) were used to evaluate the local independence assumption. Q3 values are calculated between each item pair, and values lower than 0.2 imply no local dependence among the corresponding item pairs, while the values in the range of 0.2-0.3 are acceptable even if they need closer attention (Christensen, 2017).

At the last stage, the person fit statistics were calculated. PerFit package (Tendeiro et al., 2016) was used for this procedure. As cited in the previous section, four different statistics can be used with polytomous items: Lz-poly, G-poly, Gnormed-poly, and U3-poly. Among them, the Lz-poly statistic is parametric (IRT model-based), and the rest of the statistics are nonparametric (based on Guttman errors). When specifying misfit people, the decision was made based on the empirical cut-off value specified by bootstrapping procedures at a specified p-value (the p-value for the current study was set at a conventional 0.05 significance level). After removing misfit persons from simulated datasets, misfit person-free datasets were obtained. For the rest of this article, simulated datasets will be referred to as complete datasets, while the dataset that responses of misfit persons were removed will be called misfit-free datasets to facilitate the readers' understanding. For both datasets, the GRM model was fit, and the model fit statistics, item fit statistics, item and test information levels, and predicted discrimination (as denoted by a) parameter values were compared. Model fit of IRT models was evaluated by log-likelihood statistics, G2, AIC, and BIC values. Higher log-likelihood values and lower G2, AIC and BIC values imply better model fit. In addition, item fit was evaluated with S-χ2 statistics similar to χ2 statistics, and lower values indicate a better fit of items to the given IRT model tool.

FINDINGS

Please The results section consists of three parts. In the first part, it was examined whether IRT assumptions are met in order to see the applicability of the IRT model. Later, GAD-7 data was fit to the GRM model and main findings were shortly interpreted. For the rest of the analysis, the statistics of Lz-poly, U3-poly, G-poly, and Gnormed-poly were calculated and after removing response vectors belong to misfit persons, misfit-free datasets were obtained. Those misfit-free datasets were fit to the GRM model again and the results were compared.

Checking the assumptions of IRT

The statistical results presented in the findings section represent the arithmetic mean of corresponding statistics obtained from 100 different simulated datasets. To begin, principal components analysis was used to investigate the unidimensionality of simulated datasets. The analyses' findings provided sufficient evidence for the factorability of GAD-7 data: the average KMO value was obtained as 0.92 and in none of the values obtained from simulated datasets fell below .90; the average χ2 value for Bartlett's Test of Sphericity is 3391, and with 21 degrees of freedom, all the tests were statistically significant. Furthermore, the average anti-image correlations range from 0.89 to 0.95, and the average communality values never fall below 0.50. After the analyses of factorability were conducted, the unidimensionality of the simulated datasets was investigated. As outlined in the previous section, both the EFA and the CFA were conducted to investigate the unidimensionality assumption. Based on the EFA, one factor with an Eigenvalue greater than one was extracted. The average variance value explained by this first factor was found as %55.6 of the overall variance, and the average first-to-second eigenvalue ratio was found to be 6.42, and the ratios of each simulated dataset never fell below 4. According to Gorsuch (2003), if the first to second-factor ratio is greater than 3, the scale may be considered unidimensional. These findings demonstrated the simulated datasets' unidimensionality. A looser way to look at the unidimensionality is the inspect the amount of the first variance explained by the first factor. Even this value does not provide evidence for strict unidimensionality; it enables us to see whether or not there is an underlying dominant factor for the responses provided. Drasgow and Hulin (1990) pointed out that in the context of the IRT analysis, the existence of a dominant factor is sufficient to fulfill the assumption of one-dimensionality. The variance explained by the first factor is generally viewed as an index to assess the existence of the dominant factor, even in multidimensional datasets (i.e., Miguel, 2013). Reckase (1989) found that stable estimates can be obtained when the first factor explains at least %20 of the total variance. The average value of %55.6 is well above the %20 threshold and supports the existence of an underlying dominant factor. Unidimensionality was further explored with the CFA. As the average values of the fit indices were investigated, it was seen that, the average values of them also supported the unidimensionality [χ2/df = 1.29, TLI = 0.997, GFI = 0.968, NFI = 0.977, RMSEA = 0.071, SRMR = 0.026]. All in all, the results of both the EFA and the CFA show that the assumption of unidimensionality was met.

Local independence assumption of dataset items was scrutinized via investigating Yen's Q3 statistics. As stated in the previous section, the Q3 values less than 0.2 are regarded as local independence for a given item pair, while a value of 0.3 or less needs closer inspection according to Christensen's (2017) guidelines. The results are shown in Table 1 below. In the Table, upper diagonal values represent the average Q3 statistics obtained from 100 simulated datasets, while the lower diagonal values represent the average standard errors of corresponding Q3 statistics. As the Table was investigated, it can be seen that none of the Q3 statistics for item pairs exceeded the 0.3 thresholds. It would be more desired to observe values lower than 0.2 to be sure for local independence, taking 0.3 value as a criterion; none of the Q3 values seemed to violate the local independence assumption. Because inspecting the possible violations and their underlying reasons is beyond the aim of the current study, an eventual conclusion was drawn as no violation of the local independence assumption.

To summarize, both the unidimensionality and the local independence assumptions of the IRT models were satisfied by different analyses. As a next step, the GRM model fits each simulated dataset.

Table 1. The average Q3 statistics among GAD-7 item pairs

	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7
Item 1		-0.21	-0.16	-0.13	-0.10	-0.10	-0.14
Item 2	0.03		-0.25	-0.21	-0.15	-0.16	-0.23
Item 3	0.03	0.03		-0.16	-0.12	-0.12	-0.17
Item 4	0.03	0.03	0.03		-0.10	-0.10	-0.13
Item 5	0.03	0.03	0.03	0.03		-0.07	-0.10
Item 6	0.03	0.03	0.03	0.03	0.03		-0.10
Item 7	0.03	0.03	0.03	0.03	0.03	0.03	

After checking the IRT assumptions, the GRM model was fit to the datasets separately. The average estimated item parameters and the information values of the analyses were shown in Table 2 below. The results indicated that, the second item had higher discrimination parameter value while the fifth and the sixth items have the lowest. Also, as it can be inferred from the discrimination values, the second item provided the highest information and contributed to the accuracy of the measurement while the fifth and the sixth items have lowest information values and provide the lowest accuracy to the measurement of simulated datasets.

Table 2. The average item parameters and item information values of items.

	a1	d1	d2	d3	Information
Item 1	2.17	-0.39	1.02	1.82	5.05
Item 2	3.45	-0.45	0.85	1.68	9.34
Item 3	2.50	-0.86	0.53	1.47	6.28
Item 4	2.06	-0.57	0.80	1.78	4.80
Item 5	1.59	-0.24	1.40	2.61	3.68
Item 6	1.55	-1.25	0.39	1.53	3.49
Item 7	2.30	-0.09	1.04	1.75	5.04

Note: *a1* denoted to discrimination parameter while d1-d3 denotes to successive item difficulty parameters for each category thresholds.

Based on the GRM analyses, obtained reliability plot of datasets was provided in Figure 1. As deduced from the figure, it could be stated that the average reliability of datasets never falls below 0.80 across the ability levels of -2 to 2. The curve has no clear peak, and keep going almost horizontally in this range. The figure suggest that average reliability value of datasets can provide accurate results at that range of the ability spectrum. On the other hand, as expected, the average reliability level of datasets fell dramatically at both extreme θ levels. In addition, even the test information plot of the datasets was not provided in the present study, considering that the amount of information a test yields is related to the reliability, the similar conclusions can be drawn for the information level of the datasets across the ability levels.

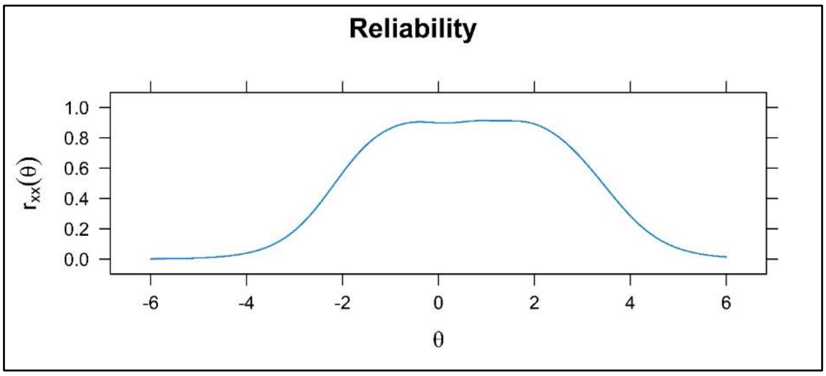


Figure 1. Empirical reliability plot of GAD-7 scale

Comparison of the GRM results for complete datasets and misfit person-free datasets.

Effect of misfit persons on the GRM model results were investigated by comparing the results for complete datasets and misfit-free datasets. The misfit-free datasets were obtained by simply removing misfit persons based on either *Lz-poly*, *U3-poly*, *G-poly* and *Gnormed-poly* person-fit statistics. This process was repeated for each simulated dataset. In the first stage, the simulated complete datasets used to estimate *Lz-poly*, *U3-poly*, *G-poly* and *Gnormed-poly* person fit statistics. After estimating person fit statistics, those cases that were flagged as misfit person were removed based each person fit statistics separately and four different misfit-free datasets were obtained. The model fit values of estimated the GRM model were presented in Table 3.

Table 3. The average model fit values of complete datasets and mean Δ after case removal.

	Lz-poly		U3-poly		G-poly		Gnormed poly	
	Mean	Mean Δ	Mean	Mean Δ	Mean	Mean Δ	Mean	Mean Δ
Loglikelihood	-7180	-649	-7033	-796	-6744	-1086	-6964	-866
G^2	2014	394	2188	220	1818	590	2123	285
AIC	14417	1299	14123	1593	13543	2172	13984	1732
BIC	14555	1300	14260	1595	13681	2175	14121	1734

The values in the table represents the arithmetic average model fit values of 100 simulated datasets. The change of model fit values when misfit persons were removed were represented with difference (Δ) scores. These difference scores were obtained by simply subtracting model fit value of misfit-free dataset from the corresponding fit value of the complete dataset. Positive Δ scores for G^2, AIC and BIC values and negative scores for the log-likelihood implies model improvement after removal of misfit person. The results showed that highest decline for log-likelihood scores were observed after removing the misfit persons *G-poly* method. Similarly, the highest increase of scores for G^2, AIC and BIC were also observed for *G-poly* method. This results suggest that, misfit person removal based on *G-poly* values contribute better to the model fit results compared to other person fit statistics.

As similar to investigation on model fit results, average item fit S-χ^2 values and mean Δ item fit S-χ^2 values for each person fit statistics were calculated and shown in Table 4.

Table 4. The average item fit values of complete datasets and mean Δ after case removal.

	Lz-poly		U3poly		Gpoly		Gnormed	
	S- χ^2	S- $\chi^2\Delta$	S- χ^2	S- $\chi^2\Delta$	S- χ^2	S- $\chi^2\Delta$	S- χ^2	S- $\chi^2\Delta$
Item 1	32.31	-2.26	33.13	-1.44	31.04	-3.53	32.88	-1.69
Item 2	22.32	-1.92	23.74	-0.51	22.37	-1.88	23.28	-0.97
Item 3	27.51	-1.15	28.40	-0.25	27.17	-1.48	28.06	-0.59
Item 4	32.82	-2.09	33.27	-1.65	31.59	-3.33	32.86	-2.06
Item 5	33.27	-1.52	37.26	2.47	31.01	-3.78	36.14	1.35
Item 6	33.55	-1.07	37.81	3.19	33.62	-1.00	37.36	2.74
Item 7	31.36	-0.83	31.91	-0.28	31.07	-1.12	31.85	-0.34

As previously stated, S-χ^2 values could be interpreted as higher values imply worse fit of the item to the given model (Orlando & Thissen, 2000). Hence, the negative values of Δ scores in the table imply model fit improvement after removal of misfit persons while the positive values imply model deterioration.

The results revealed that misfit person removal based on *G-poly* statistics contribute the most to the item fit while the worst performance was observed for *U3-poly* statistic. The only exception of this result was observed for the second item where the use of *Lz-poly* statistics contributed the most. Interestingly, this item was also found as the one with highest information value. This result can lead us to the conclusion that if an item has relatively modest level of fit to the given IRT model, *G-poly* may not be superior over the other alternative person-fit statistics. Another interesting finding showed that misfit person removal based on *U3-poly* and *Gnormed-poly* statistics worsen the fit of item 5 and item 6 because, *S-χ^2* show rise as misfit persons were removed. These two items were also being found to be the worst fitting items based on the GRM analysis with complete datasets. This result implies that when item fit poorly to the given IRT model, person removal based on *U3-poly* and *Gnormed-poly* statistics further deteriorate the fit of item and should not be preferred by the researchers.

The average item discrimination parameter values obtained with complete datasets and the average scores obtained after item removal based on each person fit statistics were presented in Table 5 below. The values in the table were calculated by taking arithmetic average of the corresponding values from each simulated dataset. The values for each person-fit statistics were interpreted based on their comparison with the values obtained for the complete datasets. Contrary to previous tables, Δ scores were not used in Table 5.

Table 5. The average discrimination parameter values of complete datasets and for misfit-free datasets.

	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7
A parameter							
Complete	2.17	3.45	2.50	2.06	1.59	1.55	2.30
Lz-poly	2.37	3.86	2.72	2.21	1.70	1.64	2.50
U3-poly	2.24	3.36	2.40	2.14	1.67	1.60	2.36
G-poly	2.41	3.63	2.70	2.29	1.85	1.81	2.55
Gnormed-poly	2.26	3.34	2.43	2.11	1.72	1.63	2.40

The results revealed that person removal based on *G-poly* statistics provided the highest increase in item discrimination values for the last four items while the highest increases were observed for the first three items by *Lz-poly* statistics. The contribution of *G-poly* and *Gnormed-poly* on item discrimination values never became the highest or the second highest for any items. Beyond that, for some conditions, the use of them even decrease discrimination values. This result imply that person removal based on *G-poly* and *Lz-poly* statistics contribute relatively more to the accuracy of measurement for each items and the researcher need to be cautious when using *G-poly* and *Gnormed-poly.*

The average item and test information values and Δ values were presented in Table 6 below.

Table 6. The average item information values of complete datasets and mean Δ after case removal.

	Lz-poly		U3-poly		Gpoly		Gnormed-poly	
	Mean	MeanΔ	Mean	MeanΔ	Mean	MeanΔ	Mean	MeanΔ
Item 1	5.61	0.59	5.29	0.27	5.84	0.82	5.38	0.36
Item 2	10.60	1.35	9.06	-0.19	9.95	0.70	9.06	-0.20
Item 3	6.97	0.72	6.06	-0.19	6.95	0.70	6.16	-0.09
Item 4	5.28	0.49	5.12	0.32	5.58	0.78	5.04	0.24
Item 5	4.00	0.37	4.01	0.38	4.52	0.89	4.17	0.54
Item 6	3.75	0.27	3.71	0.23	4.25	0.76	3.78	0.30
Item 7	5.66	0.58	5.34	0.26	5.92	0.84	5.48	0.40
Total Info	41.87	4.36	38.59	1.07	43.02	5.50	39.06	1.55

The results showed that The highest increases in total test information were observed for *Lz-poly* and *G-poly* statistics. Hence, it can be concluded that case removal based on *Lz-poly* and *G-poly* statistics improve the test information better relative to the other two person fit statistics. On the other hand, the use of *U3-poly* and *Gnormed-poly* statistics even deteriorate total information values for some items.

The table 7 below show the amount of average change on test information across ability spectrum (ability spectrum was taken between -6 and +6 θ levels) based on case removal by different person fit statistics.

Table 7. The average change of test information change across ability spectrum for different person-fir statistics

	Lz		U3		G		G normed	
θ interval	Mean	MeanΔ	Mean	MeanΔ	Mean	MeanΔ	Mean	MeanΔ
$\lvert -6.-5 \rvert$	0.00	0.001	0.01	-0.001	0.00	0.003	0.00	0.000
$\lvert -5.-4 \rvert$	0.02	0.005	0.03	-0.004	0.01	0.012	0.03	-0.001
$\lvert -4.-3 \rvert$	0.11	0.019	0.15	-0.026	0.08	0.043	0.14	-0.014
$\lvert -3.-2 \rvert$	0.61	0.050	0.80	-0.136	0.56	0.105	0.76	-0.099
$\lvert -2.-1 \rvert$	3.48	-0.154	3.69	-0.373	3.41	-0.093	3.64	-0.314
$\lvert -1.0 \rvert$	9.70	-1.254	8.46	-0.012	9.74	-1.296	8.52	-0.076
$\lvert 0.1 \rvert$	10.59	-1.284	9.23	0.072	10.70	-1.391	9.32	-0.012
$\lvert 1.2 \rvert$	11.55	-1.622	10.00	-0.067	11.67	-1.745	10.11	-0.179
$\lvert 2.3 \rvert$	4.65	-0.206	4.81	-0.364	5.50	-1.050	5.07	-0.627
$\lvert 3.4 \rvert$	0.94	0.050	1.11	-0.124	1.11	-0.118	1.18	-0.185
$\lvert 4.5 \rvert$	0.18	0.025	0.24	-0.034	0.20	0.013	0.25	-0.037
$\lvert 5.6 \rvert$	0.03	0.008	0.05	-0.005	0.03	0.011	0.05	-0.003

The results showed that case removal based on *Lz-poly* and *G-poly* statistics improved the amount of information at central θ levels while decreased the amount of information at extreme θ levels. In practical conditions, increase of information at specific ability range is generally a desired feature because most of the measurement tools aims precision for specific target group. In this sense, *Lz-poly* and *G-poly* may be more practical to obtained more desired information curve depending on the aim of measurement. On the other hand, case removal based on *U3-poly* and *Gnormed-poly* statistics improve the amount of information throughout the whole spectrum but this improvement is relatively less compared to *Lz-poly* and *G-poly* statistics.

DISCUSSION

This study contributed to the existing literature in various ways. First of all, as a result of the fitting GAD-7 derived data with the GRM, the second item of GAD-7 provided the most information while the fifth and sixth items provided the least amount of information. These findings also coincide with the previous findings (Jordan et al., 2017). Another finding is that this study showed that GAD-7 could provide highly reliable results for a relatively wide range of θ spectrum.

Removing the data belonging to misfit persons from the complete datasets improved model fit indices for all four-person fit statistics. The most effective one was G-poly, while data removal based on Lz-poly statistics was less effective and improved the model fit relatively more minor. In parallel with this finding, it was observed that item fit statistics improved with the removal of misfit individuals from the dataset for G-poly and Lz-poly item removal.

As emphasized in the introduction section, simulative data are generally used in the studies conducted for person fit analysis literature (Karabatsos, 2003; Tendeiro & Meijer, 2014), and none of these studies suggested G-poly person fit statistics as a practical approach. One reason for this contrast may be the usage of simulated datasets based on parameters obtained from real data in the current study. Tendeiro and Meijer (2014) stated that data simulation creates an unfair advantage in the effectiveness of parametric methods and that parametric statistics may yield worse results than non-parametric ones in studies performed with real data. The current study partly supported this view because Lz-poly statistic, as the only parametric technique used in this study, was not found to be the most effective method. On the other hand, in the context of this study, G-poly statistics provided the best results.

In a study examining the effect on model fit for the CFA (Conijn et al., 2014), removing misfit individuals from the dataset improved the model fit. Hence, aberrant responses affect validity negatively. Similarly, in another study (Meijer & Nering, 1997), removing misfit persons from the dataset has positively affected the model fit. Even though the current study was conducted in an IRT context, the observed improvement in model fit indices is compatible with the results of these studies conducted in the context of classical test theory. If evaluated from this point of view, it could be suggested that removing misfit persons improve model fit results regardless of the measurement paradigm adopted by the researcher.

In addition, another remarkable point is that observed model fit improvements are observed for log-likelihood, G2, AIC and BIC indices. On the other hand, in a recent study (Liu et al., 2019), when the effect of the misfit individuals on CFA models' fit indices was examined, it was observed that RMSEA and SRMR fit indices were insensitive to the elimination of misfit individuals and no consistent results were obtained for all of the fit indices. Contrary to this previous finding, improvement in all indices were observed in the current study.

In addition, in the same study, Liu and his colleagues stated that the model fit of CFA can remain acceptable when the proportion of misfit individuals is not less than %30. For none of the simulative data in this current study, the percentage of misfit individuals did not reach %20, which is far less than the %30 rates. For this reason, it has been observed that the IRT model can provide satisfactory results in the analyses performed even with complete datasets containing misfit response vectors.

According to another finding, item discrimination levels and the amount of item information (and test information) values also increased due to removing misfit persons from the dataset. In this context, the most influential person fit statistics were found as Lz-poly and G-poly. Finally, when data was removed based on Lz-poly and G-poly fit statistics, the improvement in the amount of information was observed at central θ levels, while the amount of information for extreme θ levels was decreased. On the other hand, when the U3-poly and Gnormed-poly person fits statistics were considered for data removal, a more consistent but relatively more minor amount of improvement was observed for the entire spectrum of θ.

CONCLUSION AND RECOMMENDATIONS

All in all, this study is essential in terms of showing the effects of misfit persons on IRT-based model estimation results. It provides valuable information for practitioners to choose which person fits statistics to prefer when using GAD-7 with university students. Likewise, the results showed that removing misfit persons improve validity-related findings and, accordingly, proper test interpretations become more possible. On the other hand, despite the valuable contribution, this study also has some limitations. In addition, it differs from many studies by using real data. Primarily, online data was used in this study. The test conditions were not checked accordingly since the data was not collected under controlled conditions. In addition, the data used in this study were collected from university students whose clinical history is unknown and, possibly, the majority of them do not have a psychiatric disorder. It is suggested to examine the generalizability of these findings to the clinical sample groups. Finally, the simulated datasets were derived from an instrument measuring anxiety. Hence, the findings need to be interpreted in this context. Future studies may replicate this study with the instruments measuring different constructs.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, author-ship, and/or publication of this article.

Statements of publication ethics

I/We hereby declare that the study has not unethical issues and that research and publication ethics have been observed carefully.

Researchers’ contribution rate

The study was conducted and reported by the researcher.

Ethics Committee Approval Information

As All procedures performed in this study involving human participants were in accordance with the ethical standards of the institutional research committee at Marmara University. It has the Ethics Committee Certificate with the Decision of Marmara University Institute of Education Sciences Publication Ethics Committee Dated 04.11.2020 and Numbered 2000310207.

REFERENCES

American Educational Research Association, American Psychological Association, Joint Committee on Standards for Educational, Psychological Testing (US), & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. American Educational Research Association.

American Psychiatric Association (2000). *Diagnostic and statistical manual of mental disorders (4th ed., Text Revision)*. Washington, DC.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 443–459. https://doi.org/10.1007/BF02293801

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1-29. doi: https://doi.org/10.18637/jss.v048.i06

Christensen, K. B., Makransky, G., & Horton, M. (2017). Critical values for Yen’s Q 3: Identification of local dependence in the Rasch model using residual correlations. *Applied psychological measurement*, *41*(3), 178-194. doi: https://doi.org/10.1177/0146621616677520

Conijn, J. M., Emons, W. H., & Sijtsma, K. (2014). Statistic lz-based person-fit methods for noncognitive multiscale measures. *Applied Psychological Measurement*, *38*(2), 122-136. doi: https://doi.org/10.1177/0146621613497568

Conrad, K. J., Bezruczko, N., Chan, Y. F., Riley, B., Diamond, G., & Dennis, M. L. (2010). Screening for atypical suicide risk with person fit statistics among people presenting to alcohol and other drug treatment. *Drug and Alcohol Dependence*, *106*(2-3), 92-100. doi: https://doi.org/10.1016/j.drugalcdep.2009.07.023

Drasgow, F., & Hulin, C. L. (1990). *Item response theory*. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (p. 577–636). Consulting Psychologists Press.

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*, *38*(1), 67-86. doi: https://doi.org/10.1111/j.2044-8317.1985.tb00817.x

Emons, W. H. (2008). Nonparametric person-fit analysis of polytomous item scores. *Applied Psychological Measurement*, *32*(3), 224-247. doi: https://doi.org/10.1177/0146621607302479

Engelhard Jr, G. (2009). Using item response theory and model—data fit to conceptualize differential item and person functioning for students with disabilities. *Educational and Psychological Measurement*, *69*(4), 585-602. 10.1177/0013164408323240

Gorsuch, R. L. (2003). *Factor analysis*. In J. A. Schinka & W. F. Velicer (Eds.), *Handbook of psychology: Research methods in psychology*, Vol. 2 (p. 143–164). John Wiley & Sons Inc.

Guttman, L. (1944). A basis for scaling qualitative data. *American sociological review*, *9*(2), 139-150. doi: https://doi.org/10.2307/2086306

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage publications.

Jordan, P., Shedden-Mora, M. C., & Löwe, B. (2017). Psychometric analysis of the Generalized Anxiety Disorder scale (GAD-7) in primary care using modern item response theory. *PloS one*, *12*(8), e0182162.

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education*, *16*(4), 277-298. doi: https://doi.org/10.1207/S15324818AME1604_2

Karasar, N. (2005). Bilimsel araştırma yöntemi. Nobel Yayın Dağıtım

Kline, R. B. (2015). *Principles and practice of structural equation modeling*. Guilford publications.

Konkan, R., ŞENORMANCIŞenormancı, Ö., Güçlü, O., Aydin, E., & Sungur, M. Z. (2013). Yaygın Anksiyete Bozukluğu-7 (YAB-7) Testi Türkçe Uyarlaması, Geçerlik ve Güvenirliği. *Archives of Neuropsychiatry/Noropsikiatri Arsivi*, *50*(1), 53-59. doi: https://doi.org/10.4274/npa.y6308

Liu, T., Sun, Y., Li, Z., & Xin, T. (2019). The impact of aberrant response on reliability and validity. *Measurement: Interdisciplinary Research and Perspectives*, *17*(3), 133-142. doi: https://doi.org/10.1080/15366367.2019.1584848

Meijer, R. R. (1996). Person-fit research: An introduction. *Applied Measurement in Education*, *9*(1), 3-8. https://doi.org/10.1207/s15324818ame0901_2

Meijer, R. R., & Nering, M. L. (1997). Trait level estimation for nonfitting response vectors. *Applied Psychological Measurement*, *21*(4), 321-336. doi: https://doi.org/10.1177/01466216970214003

Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied psychological measurement*, *25*(2), 107-135. doi: https://doi.org/10.1177/01466210122031957

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American psychologist*, *50*(9), 741. doi: https://doi.org/10.1037/0003-066X.50.9.741

Miguel, J. P., Silva, J. T., & Prieto, G. (2013). Career decision self-efficacy scale—short form: a Rasch analysis of the Portuguese version. Journal of Vocational Behavior, *82*(2), 116-123. https://doi.org/10.1016/j.jvb.2012.12.001

Molenaar, I. W. (1997). Nonparametric Models for Polytomous Responses. In W.J. van der Linden & R.K. Hambleton (Eds.), *Handbook of modern item response theory*, 369-380. Springer.

Morizot J., Ainsworth A.T., & Krueger S.P. (2009). Toward modern psychometrics: Application of item response theory models in personality research: In Robins R.W., Fraley R.C., Krueger RF (Eds.). *Handbook of Research Methods in Personality Psychology*. Guilford Press.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *ETS Research Report Series*, *1992*(1), i-30. doi: https://doi.org/10.1177/014662169201600206

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, *24*(1), 50–64. https://doi.org/10.1177/01466216000241003

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of educational statistics*, *4*(3), 207-230. https://doi.org/10.3102/10769986004003207

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling and more. Version 0.5–12 (BETA). *Journal of statistical software*, *48*(2), 1-36. doi: https://doi.org/10.18637/jss.v048.i02

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. Psychometrika monograph supplement. *Psychometrika*, *34*: 1-97. doi: https://doi.org/10.1002/j.2333-8504.1968.tb00153.x

Schmitt, N., Chan, D., Sacco, J. M., McFarland, L. A., & Jennings, D. (1999). Correlates of person fit and effect of person fit on test validity. *Applied Psychological Measurement*, *23*(1), 41-53. Doi: Https://doi.org/10.1177/01466219922031176

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory (Vol. 5)*. Sage publications.

Spitzer, R. L., Kroenke, K., Williams, J. B., & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: the GAD-7. *Archives of internal medicine*, *166*(10), 1092-1097. doi: https://doi.org/10.1001/archinte.166.10.1092

Tendeiro, J. N., & Meijer, R. R. (2014). Detection of invalid test scores: The usefulness of simple nonparametric statistics. *Journal of Educational Measurement*, *51*(3), 239-259. doi: https://doi.org/10.1111/jedm.12046

Tendeiro, J. N., Meijer, R. R., & Niessen, A. S. M. (2016). PerFit: An R package for person-fit analysis in IRT. *Journal of Statistical Software*, *74*(5), 1-27. doi: https://doi.org/10.18637/jss.v074.i05

Van Der Flier, H. (1982). Deviant response patterns and comparability of test scores. *Journal of Cross-Cultural Psychology*, *13*(3), 267-298. doi: https://doi.org/10.1177/0022002182013003001

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, *8*(2), 125-145. doi: https://doi.org/doi.org/10.1177/014662168400800201