ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2019; 10(4); 377-393

Comparison of Person-Fit Statistics for Polytomous Items in Different Test Conditions *

Asiye ŞENGÜL AVŞAR **

Abstract
The validity of individual test scores is an important issue that needs to be studied in psychological and educational assessment. An important factor affecting the validity of individual test scores is aberrant item response behavior. Aberrant item scores may increase/decrease the individuals' scores and as a result individuals' ability can be estimated above/below their true ability. Person-fit statistics (PFS) are useful tools to detect aberrant behavior. There are a great number of parametric and nonparametric PFS in the literature. The general purpose of the study is to examine the effectiveness of the parametric and nonparametric PFS in data sets which consist of polytomous items. This study is fundamental research aimed at determining the effectiveness of PFS using simulated data sets. According to the results, as expected, as the Type I error rates (significance alpha level) increased, detection rates (power) increased. In general, it is seen that as the number of misfitting item score vector and number of items increased, detection rates increased. Generally, nonparametric PFS (N-PFS) (especially G^P) detected more aberrant individuals than parametric PFS (P-PFS) l_z^p. However, in some tests' conditions l_z^p detected more aberrant individuals than N-PFS for longer tests. The results indicate that N-PFS outperformed P-PFS in most of the test conditions.

Key Words: Polytomous items, aberrant item response, person-fit statistics.

INTRODUCTION

It is known that psychological and educational tests are important in making decisions about individuals and identifying their learning problems, developmental problems, and psychological disturbances. It is clear that test users will focus on individual scores, especially in psychological diagnoses and treatments (Emons, 2003, 2009). Therefore, the validity of individual test scores is an important issue that needs to be studied in psychological and educational assessment.

An important factor that affects the validity of individual scores is aberrant item response behavior. For example, an individual may give incorrect answers to easy items in an exam because of being anxious during a test. This situation can lead to the person's ability estimated below her/his true ability. Another example is a situation that low-skilled individuals copy correct answers from highly skilled individuals sitting around them. This situation can lead the person's ability estimated above her/his true ability. Not taking the test seriously, lacking motivation, concentration problems in cognitive tests, giving fake responses in personality tests also form the basis for aberrant item responses. Thus, the validity of individuals' ability estimates can be negatively affected (Emons, 2003, 2008; Sijtsma & Molenaar, 2002).

Aberrant item scores may increase/decrease the individuals' scores and as a result individuals' estimated ability will be above/below their true ability. According to this, the ability of cheaters and lucky guessers are estimated spuriously high, while the abilities of examinees who are confused at the beginning of test, who never reach to items towards the end, who have language deficiencies are estimated lower than their actual ability levels (Meijer, 1996). Moreover, sometimes random guessers or examinees who respond without an idea about the item content, creatives (examinees who interpret items in a creative way) and examinees (misalign their answer sheets) also have aberrant item scores

* This study is a part of 2219 Tubitak Project which was directed by supervisor Dr. W. H. M. Emons.
** Assist. Prof., Recep Tayyip Erdoğan University, Faculty of Education, Rize-Turkey, asiye.sengul@erdogan.edu.tr, ORCID ID: 0000-0001-5522-2514

To cite this article:
Şengül-Avşar, A. (2019). Comparison of person-fit statistics for polytomous items in different test conditions. *Journal of Measurement and Evaluation in Education and Psychology, 10*(4), 377-393. doi: 10.21031/epod.525647
Received: 11.02.2019
Accepted: 24.08.2019

and the abilities of the individuals may be estimated lower or higher than their real ability levels (Meijer, 1996). In all these cases, it is clear that individuals are not evaluated correctly. Therefore, in order to be able to make right decisions according to the test results, it is important to evaluate the validity of individual item-score patterns, which raise concerns about validity.

The purpose of person-fit analysis is to determine the fit of individual response patterns with the postulated model and to identify aberrant-misfitting individual item-score vectors (Meijer & Sijtsma, 2001). To accomplish this goal, person-fit statistics (PFS) are used. PFS reveal atypical test performance with the response patterns that the individuals gave to the test items (Emons, 2008; Meijer & Sijtsma, 2001). PFS play an important role in reaching more valid results since it prevents important decisions about the individual from possibly invalid test results (Emons, 2008). Also, person-fit analysis is a valuable method for validity, which is one of the important psychometric properties of measurement tools.

Many PFS have been developed in the literature. Examples of these statistics include caution indices, norm-conformity indices, and appropriateness measurement (Drasgow, Levine & McLaughlin, 1987; Embretson & Reise, 2000; Levine & Drasgow, 1983; Tatsuoka, 1984; Tatsuoka & Tatsuoka, 1982; as cited in Emons, 2003). PFS are generally divided into parametric and nonparametric statistics (Karabatsos, 2003; Mousavi, Tendeiro, & Younesi, 2016). Parametric PFS (P-PFS) are based on parametric item response theory (PIRT), while nonparametric PFS (N-PFS) are based on group statistics (i.e., item means) or nonparametric item response theory (NIRT) (Karabatsos, 2003). Table 1 shows examples of PFS according to the item type (Tendeiro, 2016).

Table 1. Parametric and Nonparametric PFS According to Item Type

P-PFS	**Explanation**	**Item Type**
l_z	The standardized log-likelihood of the response vector	Dichotomous
l^*_z	Developed l_z (to overcome l_z limitation)	Dichotomous
l_z^p	Natural extension of l_z to polytomously scores	Polytomous
N-PFS	**Explanation**	**Item Type**
r_{pbis}	Personal biserial statistic	Dichotomous
C	The caution statistic	Dichotomous
G	Number of Guttman errors	Dichotomous
G_N	Normalized version of G	Dichotomous
A, D, E	Agreement, disagreement, and dependability statistics	Dichotomous
$U3, ZU3$	van der Flier's $U3$ and $ZU3$	Dichotomous
C	Caution statistic	Dichotomous
C^*	Modified caution statistic	Dichotomous
NCI	$NCI = 1 - 2G_{N(normed)}$	Dichotomous
H^T	Sijtsma's H^T person-fit statistic	Dichotomous
G^p	Number of Guttman errors for polytomous items ($Gpoly$)	Polytomous
G_N^p	Normalized version of $Gpoly$	Polytomous
$U3^p$	Generalization of $U3$ person-fit statistic for polytomous items ($U3\ poly$)	Polytomous

In the literature, log likelihood based l_z statistic is the most frequently studied for binary items (Rupp, 2013). It is expressed that the most frequently used P-PFS for polytomous items is l_z^p; whereas popular N-PFS include G^p, G_N^p, and $U3^p$ (Emons, 2008; Rupp, 2013; Syu, 2013).

Statistic l_z^p is the extended version of l_z for polytomous items developed by Drasgow, Levine, and Williams (1985). Statistic l_z^p is assumed to be standard normally distributed under the null model of no aberrance, where large negative values (say less than -1.645) of l_z^p suggest aberrant response behavior (Meijer, 2003). One of the N-PFS is Guttman errors (G). Statistic G is the number of item pairs for which the respondent passed/answered the difficult item but failed the easy items for dichotomous items. As for polytomous items, G is also based on item pairs. In particular, a Guttman error occurs when a respondent passed difficult steps on one item and fails easy steps on another item (Meijer, 1996, 2003). Emons (2008) proposed a normed version which takes into account the maximum of the G^p based on the sum score of the test. Both G^p's and G_N^p's minimum value is zero, which means no Guttman error, in other words, no misfit was observed. The maximum value of G^p

depends on the total score, while the maximum value of G_N^p is one and means extreme misfit (Emons, 2008). Another N-PFS is $U3^p$ (Emons, 2008), which is the extended version of *U3*. Minimum value of $U3^p$ is zero indicating no misfit, a maximum value of $U3^p$ is one indicating extreme misfit (Emons, 2008).

N-PFS have few advantages over P-PFS. N-PFS methods only require the fit of a nonparametric model and do not require fit of more restrictive parametric models (Emons, 2003). In particular, for N-PFS it is sufficient that the data set fits the Mokken Homogeneity Model (MHM). This model assumes unidimensionality, local independence, and monotonicity (i.e., nondecreasing item characteristic curves). Therefore, these assumptions should be examined before using N-PFS (Emons, 2008).

Person-fit analysis which is emphasized as an important issue in education and psychology has been successfully applied especially in achievement tests and cognitive tests (Meijer & Sijtsma, 2001). Educational studies (examining inconsistencies in curriculum, Harnisch & Linn, 1981), cognitive psychology studies (determining of learning strategies, Tatsuoka & Tatsuoka, 1982), intercultural comparison (comparing and evaluating test scores of groups from different languages, van der Flier, 1982), personality measurement studies (identification of fake answers in the measurement tools developed for the purpose of measuring personality, Dodeen & Darabi, 2009; Ferrando, 2004, 2009, 2012; Reise & Waller, 1993; Woods, Oltmanns, & Turkheimer, 2008; Zickar & Drasgow, 1996), studies on work and organization psychology (identification of individuals with unexpected item vector score in a chosen test, Meijer, 1998), evaluating attitudes (Curtis, 2004), and research on health outputs (Custers, Hoijtink, van der Net & Hel, 2000; Tang et al., 2010) can be presented as examples (as cited in Emons, 2003; Rupp, 2013). Psychological evaluations (Conijn, Emons, De Jong & Sijtsma, 2015; Meijer, Egberink, Emons & Sijtsma, 2008) also can be presented as for PFS studies.

In addition to these studies, a literature review shows that researchers developed new PFS and tested PFS in different test conditions (Emons, 2008; Glass & Dagohoy, 2007; Karabatsos, 2003; Twiste 2011; van der Flier, 1982), determined aberrant behavior via real data test applications (Egberink, 2010; Emmen, 2011; Meijer, 2003; Spoden, 2014), tested which PFS perform best detecting aberrancy (Emons, 2008; Karabatsos, 2003; Syu, 2013; Voncken, 2014). As indicated in the literature review conducted by Rupp (2013), person-fit analyses are researched via both simulated and real data sets. However, the review also shows that the person-fit analyses are studied often for binary items, and only little for polytomous items. Hence, the literature review shows paucity in research on polytomous PFS and need for more studies on the effectiveness of polytomous PFS in various simulated test conditions, especially under small samples and skew distributions of test.

Purpose of the Study

The general purpose of the study is to examine the effectiveness of parametric and nonparametric PFS in data sets which consist of polytomous items. The following questions are addressed, which are in line with the overall objective that is determined:

1. How does the proportion of detected individuals with aberrant item scores vary across test conditions such as sample size, distribution of ability, test length, and proportion of aberrancy which depends on manipulation of items and persons?
2. Which PFS performs best in different test conditions?

METHOD

This study includes a fundamental research aimed at determining the effectiveness of PFS using simulated data sets.

Data Simulation

In this study, data were simulated under Samejima's Graded Response Model (GRM), which is a suitable model for items with ordered answer categories. This model is defined by three basic assumptions, including unidimensionality, local independence, and monotonicity between latent trait and item responses (Hambleton, van der Linden & Wells, 2011; Meijer & Tendeiro, 2018).

To formally define the model, the following notation will be used. Let J be the number of items indexed by j. Each item is assumed to have (M+1) ordered answer categories. Let X_j be the random variable with realizations xj (0, …, M). The core of GRM is the item-step response functions (ISRF), which are defined as:

$$P_{jx_j}(\theta)=P\left(X_j\geq x_j\middle|\theta\right)=\frac{e^{\alpha_j(\theta-\delta_{jx_j})}}{1+e^{\alpha_j(\theta-\delta_{jx_j})}};\ x_j=(1,\ 2,\ \ldots,\ M) \quad (1)$$

In equation 1, θ is person ability, α_j is the item-slope parameter, and δ_{jxj} (1, …, M) is the location parameter. This means that each item is modeled by one common discrimination parameter and M location parameters. The location parameters δ_{jxj} shows where on the ability scale the probability of score x_j (1, …, M) or higher is equal to .50. Because item-step response functions are defined by two parameters, the model is a generalized two parametric logistic model (Embretson & Reise, 2000; Hambleton et al., 2011).

R software was employed to generate simulated data. By using the "catIRT" package (Nydick, 2015) in the R software, data sets that fit for the GRM are produced. Regardless of NIRT analysis (especially for N-PFS), the main reason data are generated based on GRM is that GRM is a special form of the MHM, and data that fit to GRM also fit to the MHM (Emons, 2008; Sijtsma, Emons, Bouwmeester, Nyklícek & Roorda, 2008). In addition, the "fungible" package (Waller & Jones, 2016) was used to generate skewed ability distributions. To compute l_z^p, one needs estimates of θ, which can be obtained using weighted maximum likelihood estimation method (WML) (Wang, 2001; Warm, 1989). Dedicated algorithms in R programming language were used for WML estimation. Accompanying R code was obtained from Emons and are available upon request.

Design factors

In this study, simulations were done as follows:

1. Data were generated under the null model according to GRM using the test conditions envisaged.
2. According to the aim of the research, data were manipulated to mimic aberrant response behavior.
3. Extreme scores when respondents choose the same extreme response options were excluded from the analyses (e.g., strongly agree or strongly disagree) for all items. That is because Emons (2008) emphasized, extreme scores do not provide adequate information for person-fit analyses.
4. Abilities were estimated using WML estimation. While estimating the abilities, true item parameters for generating the data were used.
5. PFS were computed to detect aberrancy in different conditions with "perfit package" developed by Tendeiro (2016) in R.

Test conditions are the independent variables of the study. Test conditions included different levels of sample size (100, 250, 500, and 1,000), different shapes for the distribution of person ability (normal, positively skewed, and negatively skewed), different levels of test length (J = 10 and J = 30 items), and two levels of aberrancy (low and high). For low level of aberrancy, 20% of respondents showed aberrant response behavior on half of the items; and for high level of aberrancy, 30% of respondents showed aberrant response behavior on all items.

Table 2 shows the descriptive statistics of the simulated ability distribution. For all ability distributions, mean approximately equals zero and standard deviation equals one. Inspection of skewness coefficients shows that under the normal distribution, these coefficients were very close to zero, between of 0.54 to 0.61 for positively skewed distribution, and between of -0.58 to -0.55 for negatively skewed distribution.

Table 2. Descriptive Statistics for Ability Distributions

	Mean	**Sd**	**Median**	**Mad**	**Min.**	**Max.**	**Range**	**Skewness**	**Kurtosis**	**Se**
Normal										
100	-0.03	0.87	-0.11	0.84	-2.15	2.07	4.22	0.17	-0.10	0.09
250	-0.01	0.94	-0.07	0.94	-2.99	2.13	5.12	0.01	-0.32	0.06
500	-0.02	0.95	-0.03	0.90	-2.99	2.67	5.65	-0.03	0.02	0.04
1,000	-0.03	0.96	-0.04	0.89	-3.05	3.11	6.15	0.02	0.10	0.03
Positively Skewed										
100	0.00	1.00	-0.10	0.99	-1.81	2.91	4.72	0.54	0.06	0.10
250	0.00	1.00	-0.11	1.00	-1.90	3.41	5.31	0.58	0.19	0.06
500	0.00	1.00	-0.10	1.00	-1.94	3.7	5.64	0.59	0.24	0.04
1,000	0.00	1.00	-0.11	1.00	-1.97	4.04	6.01	0.61	0.31	0.03
Negatively Skewed										
100	0.00	1.00	0.10	0.99	-2.89	1.81	4.70	-0.55	0.01	0.10
250	0.00	1.00	0.10	1.00	-3.34	1.91	5.25	-0.55	0.12	0.06
500	0.00	1.00	0.11	1.00	-3.64	1.95	5.59	-0.57	0.18	0.04
1,000	0.00	1.00	0.11	1.00	-3.96	1.98	5.94	-0.58	0.24	0.03

Sd: Standard deviation, Mad: Median absolute deviation, Min: Minimum, Max: Maximum, Se: Standard error of mean

To generate item responses under the GRM, the *a* parameters were chosen between 1.50 and 2.00 and *b* parameters were, consistent with the literature, drawn from the uniform distribution in between -2.00 and 1.50 (Bahry, 2012; Cohen, Kim, & Baker, 1993; DeMars, 2002; Jiang, Wang & Weiss, 2016; Syu, 2013). Table 3 shows the item parameters for the 10 items and 30 items test.

Table 3. Item Parameters

	Item	***a***	***b1***	***b2***	***b3***	***b4***	**Item**	***a***	***b1***	***b2***	***b3***	***b4***
	1	1.96	-1.40	-0.79	0.51	1.51	6	1.71	-1.01	0.33	1.49	2.65
	2	1.73	-1.80	-0.66	0.63	1.39	7	1.67	-1.18	-0.24	0.37	0.99
J=10	3	1.96	-1.03	-0.02	0.83	1.82	8	1.88	-1.75	-0.28	0.37	1.38
	4	1.63	-1.35	-0.14	0.42	1.03	9	1.92	-1.31	-0.67	0.76	1.56
	5	1.67	-1.63	-0.27	0.80	1.81	10	1.51	-1.17	0.11	1.08	2.34
	Item	***a***	***b1***	***b2***	***b3***	***b4***	**Item**	***a***	***b1***	***b2***	***b3***	***b4***
	1	1.81	-1.40	-0.40	0.42	1.82	16	1.53	-1.16	-0.23	0.93	1.95
	2	1.65	-1.80	-1.05	0.45	0.96	17	1.61	-1.55	-0.72	0.04	1.49
	3	1.67	-1.03	-0.04	0.96	1.59	18	1.78	-1.04	0.22	0.95	2.36
	4	1.56	-1.35	-0.73	0.49	1.08	19	1.95	-1.86	-0.51	0.08	1.24
	5	1.64	-1.63	-0.62	0.81	2.25	20	1.82	-1.22	-0.71	0.53	1.35
	6	1.55	-1.01	0.15	1.59	2.23	21	1.53	-1.20	-0.03	1.11	1.80
	7	1.55	-1.18	-0.56	0.71	1.97	22	1.67	-1.21	0.01	1.40	2.78
J=30	8	1.63	-1.75	-0.73	0.10	0.88	23	1.52	-1.64	-0.37	0.89	1.63
	9	1.53	-1.31	-0.51	0.82	2.15	24	1.75	-1.94	-0.50	0.83	1.47
	10	1.80	-1.17	0.09	1.50	2.16	25	1.55	-1.43	-0.69	0.81	2.01
	11	1.56	-1.90	-0.48	0.70	1.95	26	1.71	-1.34	0.07	1.48	2.68
	12	1.75	-1.35	-0.40	0.78	2.14	27	1.65	-1.89	-0.77	-0.10	1.27
	13	1.68	-1.49	-0.07	0.83	2.18	28	1.93	-1.85	-0.58	0.78	1.84
	14	1.89	-1.29	-0.53	0.65	1.25	29	1.76	-1.07	0.25	1.11	2.07
	15	1.85	-1.14	-0.29	1.06	1.96	30	1.83	-1.52	-0.75	0.55	1.57

Baker (2001) suggested the following guidelines for interpreting *a* coefficients: 0 *none*, 0.01-0.34 *very low*, 0.35-0.64 *low*, 0.65-1.34 *moderate*, 1.35-1.69 *high*, > 1.70 *very high*, and ∞ (+ infinity) *perfect*. Hence, the tests in this study consisted of relatively high discriminating items, but these values are

unrealistic in practice. Previous studies convincingly showed that the power of PFS relates to the items' discrimination power (Emons, 2008; Meijer, Molenaar, & Sijtsma, 1994; Meijer & Sijtsma, 2001). Higher discrimination power may produce a higher detection rate (Emons, 2008).

There are many kinds of aberrant behavior that may affect test results. One of them is *careless and inattention*. In some test applications, individuals answer items randomly because they are careless, or a random pattern emerges due to misreading or not reading the questions, or due to alignments errors (Emons, 2008). Randomness-like response behaviors from important types of aberrant behavior (Conijn et al. 2015) and will be the subject of this study. To accomplish this goal, aberrant item response vectors were created by simulating random scores from the uniform distribution similar to Emons's (2008) study.

The selected test conditions are based on the literature (Lee, 2007; Lee, Wollack & Douglas, 2009; Liang, Wells & Hambleton, 2014; Ramsay, 1991; Syu, 2013). In particular, variation in the shape of ability distribution, small sample sizes and short tests are often seen in classroom measurement applications. One condition nevertheless consisted of a large sample size (1,000). This condition was chosen to see how PFS function in large samples and can be seen as a benchmark for the other results.

Data were generated using a fully factorial design including 4 (sample size) × 3 (ability distribution) × 2 (test length) × 2 (aberrancy levels) = 48 conditions. In total 100 replications were obtained for each test condition, thus in total 4800 data sets were simulated.

Data Analysis

Empirical Type I error rates and detection rates (power) are the dependent variables of the study. For each PFS (l_z^p, $U3^p$, G_N^p and G^p), the empirical Type I error rates and detection rates were evaluated at four the theoretical Type I error rates (nominal significance levels) ($\alpha = .01$, $\alpha = .05$, $\alpha = .10$ and $\alpha = .20$). Empirical Type I error rate is the observed proportion of non-aberrant persons identified as aberrant. Also, the detection rate is the proportion of aberrant persons correctly identified as aberrant (Voncken, 2014).

The theoretical Type I error rates which were chose in the study determined from the literature view results. It is stated in the literature that large alpha levels (e.g., .05, .10 and .20) are preferable because PFS have relatively low power detect aberrancy for small test lengths and low alpha levels (Emons, 2008; Emons, Glas, Meijer & Sijtsma, 2003; Meijer, 2003; Spoden, 2014; Voncken, 2014).

To decide whether a pattern shows significant misfit, one needs to have critical values. Certain rules are followed in the calculation of critical values for the PFS. In particular, the critical values for parametric l_z^p is determined, as in Voncken's (2014) study, to be -2.32, -1.645, -1.28, and -0.84. These are critical values from the standard normal distribution for alphas of .01, .05, .10 and .20 (one-tailed tests). Because N-PFS lack theoretical distributions, the critical values have to be determined differently. This study uses critical values of N-PFS that were determined automatically by perfit package in a pilot study. These cut-off values were fixed for every simulation and replication. Researchers are strongly recommended to fix the cut-off score with the command *set.seed ()* before identifying individuals with aberrant item patterns according to the cut-off score in the relevant package (Meijer, Niessen & Tendeiro, 2016; Tendeiro, 2016). Otherwise, different critical values with small differences are reached in each calculation.

RESULTS

There are two levels of aberrancy in this study. PFS analysis results are given in Table 4 to Table 9. Table 4 gives the findings for normally distributed ability for 10 items.

Table 4. Detection Rates for Normal Distributed Sample for 10 Items with Low and High Aberrancy Level

PFS	Low Aberrancy								High Aberrancy							
	Nominal Significance Levels and Detection Rates								Nominal Significance Levels and Detection Rates							
	.01	D.R.	.05	D.R.	.10	D.R.	.20	D.R.	.01	D.R.	.05	D.R.	.10	D.R.	.20	D.R.
	N = 100															
l_z^p	.03	**.05**	.03	.10	.04	.10	.08	.35	.00	.10	.00	.30	.00	.43	.03	.60
$U3^p$	.01	**.05**	.04	.10	.04	.30	.21	.70	.00	.10	.01	**.40**	.01	**.57**	.07	.67
G_N^p	.01	**.05**	.03	.10	.05	.30	.18	.65	.00	.13	.00	**.40**	.01	.53	.07	.67
G^p	.01	**.05**	.03	**.15**	.08	**.35**	.16	**.75**	.00	**.17**	.00	.37	.01	.50	.07	**.77**
	N = 250															
l_z^p	.00	**.18**	.02	.32	.02	.40	.07	.48	.00	**.17**	.01	.33	.01	.44	.01	.67
$U3^p$	.01	.04	.03	.42	.06	.52	.16	.64	.01	.11	.01	.33	.03	.49	.05	.71
G_N^p	.01	.08	.03	.42	.08	**.56**	.16	.66	.01	.13	.01	.35	.02	.52	.05	.72
G^p	.00	**.18**	.03	**.48**	.05	.52	.12	**.70**	.00	.13	.00	**.37**	.02	**.55**	.04	**.77**
	N =5 00															
l_z^p	.00	.11	.03	.20	.04	.30	.11	.42	.00	.15	.00	.34	.01	.47	.02	.63
$U3^p$	.02	.04	.06	.27	.08	.40	.17	.60	.01	.12	.03	.38	.04	.54	.09	**.75**
G_N^p	.02	.11	.06	.28	.08	.43	.14	.58	.01	.12	.03	.35	.03	.52	.07	.72
G^p	.01	**.14**	.04	**.34**	.06	**.49**	.14	**.69**	.00	**.17**	.01	**.41**	.02	**.59**	.07	**.75**
	N = 1 000															
l_z^p	.01	.09	.02	.18	.04	.30	.09	.40	.00	.12	.00	.33	.01	.44	.02	.62
$U3^p$	.01	.08	.05	.23	.09	.34	.14	.52	.01	.12	.02	.35	.04	.49	.08	.65
G_N^p	.02	.11	.05	.25	.09	.35	.15	.56	.01	.11	.03	.35	.04	.49	.07	.63
G^p	.01	**.15**	.03	**.28**	.07	**.45**	.13	**.61**	.00	**.14**	.00	**.37**	.02	**.52**	.06	**.71**

Note. The bolded detection rates denote the conditions in which PFS perform best. D.R.: Detection rates. N: Sample size

Inspection of Table 4 shows that as sample size increased, the detection rate increased in many test conditions. Almost all conditions, detection rates increased with increasing aberrancy levels. In general, G^p showed best performance to detect aberrancy. In addition to these findings, it is found that nonparametric $U3^p$ and G_N^p statistics are very close to each other. When empirical Type I error rates are examined, it is seen that these values exceed their nominal levels especially for low aberrancy level at $\alpha = .01$ and $\alpha = .05$. Also, empirical Type I error rates are smaller than their nominal levels in all conditions for high aberrancy level except for $\alpha = .01$. It can be seen that as increased of aberrancy, empirical Type I error rates decreased.

Table 5 gives the findings for positively skewed ability distribution for 10 items. Table 5 shows empirical Type I error rates and detection rates for PFS for positive distributed ability, for different sample sizes and low and high aberrancy levels. As expected, it is seen that as the Type I error rates increased, the detection rate increased. It is seen that as sample size increased, the detection rate increased in many test conditions for high aberrancy level. Almost all conditions detection rates increased according to the aberrancy level. In general, G^p showed best performance to detect aberrancy. In addition to these findings, it is found that nonparametric $U3^p$ and G_N^p statistics are very close to each other. When empirical Type I error rates are examined, it is seen that these values are smaller than their nominal levels both low and high aberrancy except for $\alpha = .01$. Empirical Type I error rates are equal to or smaller than their nominal level for $\alpha = .01$. It can be seen that as increased of aberrancy, empirical Type I error rates decreased.

Table 6 gives the findings for negatively skewed distribution for 10 items. Table 6 shows the detection rates for negatively distributed ability, for different sample sizes and low and high aberrancy. It is seen that as the nominal significance level increased, the detection rates increased almost all test conditions. In general, as sample size increased, the detection rates increased. However, detection rates of l_z^p decreased dramatically for large sample in low aberrancy level when $\alpha = .05$. Detection rates increased according to the aberrancy level in all test conditions. In general, G^p showed best performance to detect aberrancy. In addition to these findings, it is found that nonparametric $U3^p$ and G_N^p statistics are very close to each other. When empirical Type I error rates are examined, in general, these values are smaller than their nominal levels both low and high aberrancy except for $\alpha = .01$. Also, empirical Type

I error rates are equal to or smaller than their nominal α = .01. It can be seen that as increased of aberrancy, empirical Type I error rates decreased.

Table 5. Detection Rates for Positively Skewed Distributed Sample for 10 Items with Low and High Aberrancy Level

PFS	**Low Aberrancy**								**High Aberrancy**							
	Nominal Significance Levels and Detection Rates								**Nominal Significance Levels and Detection Rates**							
	.01	**D.R.**	**.05**	**D.R.**	**.10**	**D.R.**	**.20**	**D.R.**	**.01**	**D.R.**	**.05**	**D.R.**	**.10**	**D.R.**	**.20**	**D.R.**
	N = 100															
l_z^p	.00	.07	.01	.19	.03	.29	.07	.42	.00	.11	.00	.28	.01	.41	.03	.57
$U3^p$	.01	.07	.04	.24	.08	.38	.16	.59	.00	.09	.02	.30	.04	.46	.09	.66
G_N^p	.01	.08	.03	.26	.07	.41	.15	.60	.00	.10	.02	.30	.03	.47	.08	.67
G^p	.00	**.12**	.02	**.31**	.06	**.46**	.14	**.64**	.00	**.12**	.01	**.34**	.02	**.53**	.06	**.71**
	N = 250															
l_z^p	.00	.07	.01	.20	.03	.30	.07	.45	.00	**.14**	.00	.31	.01	.43	.02	.60
$U3^p$	.01	.07	.04	.28	.08	.43	.16	.61	.00	.11	.02	.33	.04	.50	.08	.69
G_N^p	.01	.09	.04	.30	.07	.45	.16	.62	.00	.11	.02	.33	.03	.50	.08	.70
G^p	.00	**.14**	.02	**.35**	.06	**.49**	.14	**.66**	.00	**.14**	.00	**.39**	.01	**.54**	.05	**.73**
	N = 500															
l_z^p	.00	.07	.01	.20	.03	.30	.07	.44	.00	.14	.00	.32	.01	.45	.02	.61
$U3^p$	.01	.08	.04	.28	.08	.42	.16	.61	.01	.12	.02	.35	.03	.51	.08	.70
G_N^p	.01	.10	.04	.30	.08	.45	.16	.62	.00	.12	.02	.35	.03	.51	.08	.69
G^p	.00	**.14**	.03	**.34**	.06	**.49**	.14	**.66**	.00	**.15**	.00	**.39**	.01	**.54**	.05	**.73**
	N = 1 000															
l_z^p	.00	.08	.01	.20	.03	.30	.07	.45	.00	.14	.00	.33	.01	.45	.02	.61
$U3^p$	.01	.08	.04	.29	.08	.44	.17	.61	.01	.13	.02	.36	.04	.52	.09	.71
G_N^p	.01	.11	.04	.31	.08	.46	.16	.63	.01	.13	.02	.36	.03	.52	.08	.71
G^p	.00	**.15**	.03	**.36**	.06	**.49**	.14	**.66**	.00	**.17**	.01	**.40**	.02	**.56**	.05	**.74**

Note. The bolded detection rates denote the conditions in which PFS perform best. D.R.: Detection rates. N: Sample size

Table 6. Detection Rates for Negatively Skewed Distributed Sample for 10 Items with Low and High Aberrancy Level

PFS	**Low Aberrancy**								**High Aberrancy**							
	Nominal Significance Levels and Detection Rates								**Nominal Significance Levels and Detection Rates**							
	.01	**D.R.**	**.05**	**D.R.**	**.10**	**D.R.**	**.20**	**D.R.**	**.01**	**D.R.**	**.05**	**D.R.**	**.10**	**D.R.**	**.20**	**D.R.**
	N = 100															
l_z^p	.00	.07	.01	.20	.03	.29	.07	.45	.00	.12	.00	.28	.01	.41	.02	.58
$U3^p$	.01	.07	.04	.24	.08	.40	.16	.56	.01	.09	.02	.30	.04	.48	.09	.67
G_N^p	.01	.08	.04	.26	.07	.42	.15	.58	.00	.09	.02	.31	.04	.47	.08	.67
G^p	.00	**.13**	.02	**.33**	.05	**.46**	.13	**.64**	.00	**.13**	.01	**.36**	.02	**.52**	.06	**.72**
	N = 250															
l_z^p	.00	.07	.01	.20	.03	.30	.07	.45	.00	.14	.00	.31	.01	.44	.02	.60
$U3^p$	.01	.07	.04	.28	.08	.43	.16	.61	.01	.10	.02	.33	.04	.50	.08	.70
G_N^p	.01	.10	.04	.30	.07	.44	.16	.62	.01	.11	.02	.33	.03	.50	.08	.70
G^p	.00	**.15**	.03	**.34**	.06	**.48**	.14	**.66**	.00	**.15**	.01	**.38**	.02	**.55**	.05	**.73**
	N = 500															
l_z^p	.00	.08	.01	.20	.03	.30	.07	.44	.00	.14	.00	.32	.01	.45	.02	.61
$U3^p$	.01	.08	.05	.27	.08	.42	.17	.60	.01	.12	.02	.36	.04	.52	.08	.70
G_N^p	.01	.10	.04	.30	.08	.44	.17	.62	.01	.12	.02	.36	.04	.52	.08	.70
G^p	.01	**.14**	.03	**.34**	.06	**.48**	.14	**.65**	.00	**.16**	.01	**.40**	.02	**.55**	.06	**.73**
	N = 1 000															
l_z^p	.00	.08	.00	.08	.03	.30	.07	.44	.00	.14	.00	.33	.01	.45	.02	.61
$U3^p$	.01	.07	.05	.29	.09	.43	.17	.61	.01	.12	.02	.37	.04	.53	.09	.71
G_N^p	.01	.10	.04	.31	.08	.45	.17	.62	.01	.13	.02	.36	.04	.52	.08	.71
G^p	.00	**.15**	.03	**.35**	.06	**.49**	.14	**.65**	.00	**.17**	.01	**.40**	.02	**.56**	.06	**.74**

Note. The bolded detection rates denote the conditions in which PFS perform best. D.R.: Detection rates. N: Sample size

Table 7 gives the findings for normally distributed ability for 30 items. Table 7 shows the detection rates for normally distributed ability, for different sample sizes and aberrancy levels. As expected, it

is seen that as the nominal significance levels increased, the detection rates increased as well. There is no specific trend regarding the effect of sample size on the detection rates. However, when all test conditions are examined, the highest detection rates were observed in the largest sample. For l_z^p, detection rates increased with increasing aberrancy levels at all nominal significance levels. In general, G^p showed best performance to detect aberrancy in low aberrancy level, while l_z^p showed best performance to detect aberrancy in high aberrancy level. In addition to these findings, it is found that nonparametric $U3^p$ and G_N^p statistics were very close to each other. When empirical Type I error rates are examined, it is seen that these values never exceed their nominal levels in all test conditions. Empirical Type I error rates are smaller than or equal to their nominal $\alpha = .01$ for low aberrancy. Also, all empirical Type I error rates are smaller than their nominal levels for high aberrancy. It can be seen that as increased of aberrancy, empirical Type I error rates decreased.

Table 7. Detection Rates for Normal Distributed Sample for 30 Items with Low and High Aberrancy Level

PFS	**Low Aberrancy**								**High Aberrancy**							
	Nominal Significance Levels and Detection Rates								**Nominal Significance Levels and Detection Rates**							
	.01	**D.R.**	**.05**	**D.R.**	**.10**	**D.R.**	**.20**	**D.R.**	**.01**	**D.R.**	**.05**	**D.R.**	**.10**	**D.R.**	**.20**	**D.R.**
	N = 100															
l_z^p	.00	**.25**	.03	**.45**	.05	.55	.11	.75	.00	**.53**	.00	**.77**	.03	**.83**	.04	**.93**
$U3^p$	.00	.15	.04	.40	.05	**.70**	.10	**.80**	.00	.07	.00	.40	.00	.70	.04	.87
G_N^p	.00	.15	.04	.35	.05	**.70**	.11	.75	.00	.07	.00	.33	.00	.70	.04	.87
G^p	.00	**.25**	.00	.40	.05	.65	.06	**.80**	.00	.07	.00	.27	.00	.67	.00	.90
	N = 250															
l_z^p	.00	**.26**	.02	**.46**	.05	.58	.08	.68	.00	**.56**	.00	**.75**	.00	**.85**	.00	.92
$U3^p$	.00	.18	.02	.36	.05	.48	.10	.76	.00	.16	.00	.56	.00	.76	.03	**.95**
G_N^p	.00	.18	.01	.36	.04	.48	.11	.74	.00	.12	.00	.51	.00	.77	.03	.92
G^p	.00	.20	.01	.44	.01	**.62**	.07	**.84**	.00	.15	.00	.52	.00	.75	.01	.93
	N = 500															
l_z^p	.01	.19	.02	.44	.03	.55	.07	.70	.00	**.55**	.00	**.77**	.00	**.85**	.01	**.94**
$U3^p$	.01	.16	.02	.47	.06	.57	.10	.77	.00	.07	.00	.50	.01	.69	.02	.90
G_N^p	.01	.16	.02	.48	.06	.60	.12	.75	.00	.07	.01	.46	.01	.69	.02	.87
G^p	.00	**.26**	.01	**.49**	.03	**.65**	.09	**.85**	.00	.13	.00	.51	.00	.76	.01	.91
	N = 1 000															
l_z^p	.00	.28	.01	.50	.02	.64	.05	.76	.00	**.61**	.00	**.78**	.00	**.87**	.00	**.95**
$U3^p$	.01	.23	.02	.49	.04	.64	.09	.82	.00	.42	.00	.63	.01	.75	.01	.91
G_N^p	.01	.30	.02	.50	.04	.65	.10	.83	.00	.42	.00	.62	.01	.75	.01	.92
G^p	.00	**.31**	.01	**.59**	.02	**.74**	.06	**.88**	.00	.41	.00	.63	.00	.77	.00	.92

Note. The bolded detection rates denote the conditions in which PFS perform best. D.R.: Detection rates. N: Sample size

Table 8 gives the findings for positively skewed ability distribution for 30 items. Table 8 shows the detection rates for PFS for positively skewed distributed ability for different sample sizes, low and high aberrancy. In general, detection rates increased with increasing aberrancy levels. However, for N-PFS results show higher detection rates for low aberrancy level than for high aberrancy level. This result is seen in test conditions which are consist for sample size 100 and at $\alpha = .01$ and $\alpha = .05$ nominal levels, for sample size 250 at $\alpha = .01$ nominal level. Statistic G^p showed best performance to detect aberrancy at low aberrancy levels except for sample size 100 at $\alpha = .01$ and $\alpha = .05$ nominal levels, and for sample size 250 at $\alpha = .01$ nominal level. It is seen that l_z^p showed best performance to detect aberrancy for all sample sizes and all Type I error rates in high aberrancy level. In addition to these findings, it is found that detection rates for nonparametric $U3^p$ and G_N^p statistics were very close to each other. When empirical Type I error rates are examined, it is seen that these values were not exceed their nominal levels in most of test conditions. Only for $U3^p$, empirical Type I error rate was equal to its $\alpha = .01$ nominal level for large sample and low aberrancy. Also, it is found that all empirical Type I error rates are smaller than their nominal levels for high aberrancy.

Table 8. Detection Rates for Positively Skewed Distributed Data for 30 Items with Low and High Aberrancy Level

PFS	Low Aberrancy								High Aberrancy							
	Nominal Significance Levels and Detection Rates								Nominal Significance Levels and Detection Rates							
	.01	**D.R.**	**.05**	**D.R.**	**.10**	**D.R.**	**.20**	**D.R.**	**.01**	**D.R.**	**.05**	**D.R.**	**.10**	**D.R.**	**.20**	**D.R.**
	N = 100															
l_z^p	.00	**.27**	.01	**.49**	.02	.62	.06	.74	.00	**.51**	.00	**.74**	.00	**.84**	.01	**.91**
$U3^p$	.00	.12	.01	.38	.03	.59	.08	.78	.00	.11	.00	.38	.00	.60	.01	.86
G_N^p	.00	.12	.01	.39	.03	.58	.08	.78	.00	.10	.00	.36	.00	.60	.01	.86
G^p	.00	.15	.00	.44	.01	**.64**	.06	**.84**	.00	.11	.00	.37	.00	.61	.00	.87
	N = 250															
l_z^p	.00	**.29**	.01	.49	.02	.62	.05	.76	.00	**.57**	.00	**.79**	.00	**.87**	.00	**.94**
$U3^p$	.00	.19	.02	.47	.04	.65	.09	.82	.00	.19	.00	.51	.00	.72	.01	.89
G_N^p	.00	.20	.01	.47	.03	.64	.09	.82	.00	.18	.00	.50	.00	.71	.01	.89
G^p	.00	.23	.00	**.53**	.02	**.70**	.06	**.87**	.00	.20	.00	.52	.00	.72	.00	.91
	N = 500															
l_z^p	.00	.28	.01	.50	.02	.62	.06	.75	.00	**.59**	.00	**.80**	.00	**.88**	.00	**.94**
$U3^p$	.00	.23	.02	.52	.04	.67	.10	.82	.00	.28	.00	.60	.00	.78	.02	.91
G_N^p	.00	.25	.02	.52	.04	.66	.09	.81	.00	.27	.00	.59	.00	.77	.02	.91
G^p	.00	**.30**	.01	**.58**	.02	**.73**	.07	**.87**	.00	.28	.00	.60	.00	.78	.00	.92
	N = 1,000															
l_z^p	.00	.29	.01	.50	.02	.61	.05	.76	.00	**.60**	.00	**.81**	.00	**.89**	.00	**.95**
$U3^p$	.01	.27	.02	.55	.04	.68	.10	.82	.00	.31	.00	.64	.01	.80	.02	.92
G_N^p	.00	.29	.02	.55	.04	.68	.10	.82	.00	.30	.00	.62	.01	.78	.02	.92
G^p	.00	**.34**	.01	**.60**	.02	**.74**	.07	**.87**	.00	.32	.00	.63	.00	.80	.00	.93

Note. The bolded detection rates denote the conditions in which PFS perform best. D.R.: Detection rates. N: Sample size

Table 9 gives the findings for negatively skewed distribution for 30 items. Table 9 shows the detection rates for PFS for negatively skewed distributed ability, for different sample sizes and for low and high aberrancy levels.

Table 9. Detection Rates for Negatively Skewed Distributed Data for 30 Items with Low and High Aberrancy Level

PFS	Low Aberrancy								High Aberrancy							
	Nominal Significance Levels and Detection Rates								Nominal Significance Levels and Detection Rates							
	.01	**D.R.**	**.05**	**D.R.**	**.10**	**D.R.**	**.20**	**D.R.**	**.01**	**D.R.**	**.05**	**D.R.**	**.10**	**D.R.**	**.20**	**D.R.**
	N = 100															
l_z^p	.00	**.27**	.01	**.48**	.02	.60	.06	.72	.00	**.54**	.00	**.77**	.00	**.85**	.01	**.93**
$U3^p$	.00	.12	.01	.38	.03	.58	.09	.77	.00	.11	.00	.38	.01	.62	.01	.87
G_N^p	.00	.12	.01	.38	.03	.58	.08	.78	.00	.11	.00	.38	.00	.62	.01	.87
G^p	.00	.13	.00	.43	.01	**.64**	.06	**.83**	.00	.12	.00	.40	.00	.64	.00	.88
	N = 250															
l_z^p	.00	**.29**	.01	.51	.02	.63	.06	.76	.00	**.58**	.00	**.80**	.00	**.88**	.00	**.94**
$U3^p$	.01	.16	.02	.46	.04	.64	.09	.81	.00	.20	.00	.54	.01	.73	.02	.90
G_N^p	.00	.17	.02	.46	.04	.63	.09	.80	.00	.19	.00	.52	.01	.72	.02	.90
G^p	.00	.25	.01	**.54**	.02	**.70**	.06	**.86**	.00	.22	.00	.55	.00	.75	.00	.91
	N = 500															
l_z^p	.00	.29	.01	.50	.02	.62	.06	.75	.00	**.60**	.00	**.81**	.00	**.89**	.00	**.95**
$U3^p$	.01	.23	.02	.51	.04	.66	.09	.82	.00	.27	.00	.61	.01	.79	.02	.92
G_N^p	.01	.23	.02	.50	.04	.65	.10	.81	.00	.26	.01	.60	.01	.78	.02	.91
G^p	.00	**.30**	.01	**.58**	.02	**.73**	.07	**.86**	.00	.30	.00	.62	.00	.79	.00	.92
	N = 1 000															
l_z^p	.00	.29	.01	.50	.02	.62	.06	.76	.00	**.61**	.00	**.82**	.00	**.90**	.00	**.95**
$U3^p$	.01	.25	.02	.54	.05	.68	.10	.82	.00	.32	.00	.65	.01	.81	.02	.93
G_N^p	.01	.26	.02	.53	.05	.67	.10	.81	.00	.30	.01	.64	.01	.80	.02	.92
G^p	.00	**.34**	.01	**.61**	.02	**.74**	.07	**.87**	.00	.34	.00	.66	.00	.81	.00	.93

Note. The bolded detection rates denote the conditions in which PFS perform best. D.R.: Detection rates. N: Sample size

Inspection of Table 9 shows that as expected, as the nominal significance levels increased, the detection rates increased as well. It is also seen in almost all conditions of low aberrancy that as sample size increased, the detection rate increased. Although, it is seen that as sample size increased, the detection rate increased in high aberrancy level for all samples. In general, detection rates increased according to the aberrancy level except for α = .01 and α = .05 for N-PFS. Broadly speaking, across all conditions, G^p showed best performance to detect aberrancy at low aberrancy level while l_z^p showed best performance to detect aberrancy at high aberrancy level. In addition to these findings, it is found that the detection rates of nonparametric $U3^p$ and G_N^p statistics were very close to each other. When empirical Type I error rates are examined, it is seen that these values did not exceed their nominal levels in high aberrancy. However, empirical Type I error rates are smaller than or equal to their nominal α = .01 for low aberrancy. It can be seen that as increased of aberrancy, empirical Type I error rates decreased.

DISCUSSION and CONCLUSION

The general purpose of the study is to examine the effectiveness of parametric and nonparametric PFS in data sets which consist of polytomous items. According to this aim, data simulated in different test conditions and these data sets were analyzed.

The results confirmed several important effects of significance level, sample size, ability distribution, and aberrance level. As expected, the detection rates increased with increasing nominal significance levels (the theoretical Type I error rates) in all test conditions. Moreover, it is seen that detection rates increased as the number of misfitting item score vector and number of misfitting items increased. Simulation results suggest that the shape of sample distributions has little effect on the detection of aberrancy. So, it can be said that shape of ability distribution (determined in this study's test conditions) is an unimportant factor for the effectiveness of PFS.

In general, sample size affected detection rates. In most of test conditions, it is seen that as sample size increased, detection rates increased. However, this result conflicts with Syu (2013), who studied with parametric l_z^p and nonparametric G^p and $U3^p$. Syu (2013) only found small differences in the detection rates across sample sizes for specific PFS. In addition to this finding, Syu (2013) stated that findings are tentative because sample size is too small for providing sufficient calculations for PFS.

It is seen that in general, empirical Type I error rates smaller than their nominal levels (the theoretical Type I error rates). However, in all shapes of ability distributions for 10 and 30 items, empirical Type I error rates are equal to or smaller than their nominal level at α = .01. Except of this conclusion, it is seen that for normally distributed sample for 10 items, empirical Type I error rates exceed its nominal level at α = .01. In Voncken's (2014) study, detection rates were determined for binary items. In that study it is found that l_z*'s empirical Type I rate exceeds its nominal level at α = .01. Also, it is seen that as increased of aberrancy, empirical Type I error rates decreased. These findings are consistent with Voncken (2014).

To summarize, as expected, as the nominal significance level was set higher, tests were longer, and amount of the aberrant proportions increased, the detection rates increased as well. These findings are consistent with other person-fit studies (Emons, 2008; Karabatsos, 2003; Meijer & Sijtsma, 2001; Voncken, 2014).

A comparison of the effectiveness of the different PFS showed the following important trends. It is seen that detection rates were very close to each other for P-PFS and N-PFS (especially $U3^p$ and G_N^p). However, in general, G^p was the most effective in detecting aberrant individuals and even performed better than l_z^p. These results are consistent with Emons (2008) and Syu (2013). They compared same PFS as used in this study in different test conditions. Like in this study, in their studies G^p showed best performance to detect aberrancy. In Syu's (2013) study it's also stated that for small sample sizes N-PFS perform better than P-PFS.

It is found that for all test conditions detection rates were sufficiently high except at α = .01. Detection rates got their maximum value at α = .20. PFS may have very low detection rates at small significance

levels of α = .01, which questions their effectiveness at these significance levels. These findings are consistent with literature. Therefore, it is suggested that researchers should choose liberal significance levels (i.e., α = .20) to reach some power in detecting aberrancy (Emons, 2008; Meijer, 2003; Voncken, 2014).

Based on the result, the following general conclusions about the suitability of different statistics can be drawn. Results also showed that for detecting careless and inattention aberrant behavior long tests are more useful than small tests. However, long tests are not always feasible in practice. This renders PIRT models less useful in many applications because they require large sample sizes and sufficiently longer tests to obtain accurate estimates of the item parameters. NIRT models, and accompanying N-PFS do not suffer from these problems as they use observed group statistics and therefore are particularly useful in small samples and short tests (Junker & Sijtsma, 2001; Meijer, 2004; Molenaar, 2001). When PIRT and NIRT models are compared, NIRT models are less restrictive. The main difference between these models is about item characteristic curves. In PIRT model, these curves which are logistic or normal ogive are determined postulated parametric model (Lee et al., 2009; Sodano & Tracey, 2011). However, in NIRT models these curves do not require any parametric forms, especially MHM assumes only that monotony nondecreasing θ (Lee et al., 2009; Sijtsma & Molenaar, 2002). And so, it can be said that NIRT models are more flexible than PIRT models.

It must be emphasized that in practice if researchers want to study aberrant response behavior with N-PFS, researcher should investigate MHM assumptions. MHM can fit with skewed data (Şengül Avşar & Tavşancıl, 2017). MHM is an appropriate model for small samples (Junker & Sijtsma, 2001; Molenaar, 2001). These are MHM's important advantages to their parametric counterparts. Of course, if researchers want to study response aberrancy with P-PFS, they should demonstrate fit of the data with the parametric model assumptions. In general, if data do not fit PIRT models, researchers often can use NIRT models and N-PFS for detecting aberrant individuals.

An assumption was that all individuals answered all items in this study. In other words, there were no missing data in data sets. Missing data effects on PFS and missing data handling methods for best recovery PFS can be investigated. Apart from the test conditions determined in the study, the effectiveness of PFS can be determined by simulating different test conditions. Also, PFS which were used in this study can compared with real data applications.

REFERENCES

Bahry, L. M. (2012). *Polytomous item response theory parameter recovery: an investigation of nonnormal distributions and small sample size* (Master's thesis). Retrieved from ProQuest Dissertations and Theses database. (UMI No. MR90146)

Baker, F. B. (2001). *The basis of item response theory*. United State of America: Eric Clearinghouse on Assessment and Evaluation.

Cohen, A. S., Kim, S. H., & Baker, F. B. (1993). Detection of differential item functioning in the graded response model. *Applied Psychological Measurement, 17*(4), 335-350. doi: 10.1177/014662169301700402

Conijn, J. M., Emons, W. H., De Jong, K., & Sijtsma, K. (2015). Detecting and explaining aberrant responding to the outcome questionnaire-45. *Assessment, 22*(4), 513-524. doi: 10.1177/1073191114560882

DeMars, C. E. (2002, April). *Recovery of graded response and partial credit parameters in multilog and parscale*. Paper presented at the annual meeting of American Educational Research Association, Chicago.

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: 10.1111/j.2044-8317.1985.tb00817.x

Egberink, I. J. A. L. (2010). *Applications of item response theory to non-cognitive data.* Groningen: University Library Groningen.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. New Jersey, NJ: Lawrence Erlbaum Associates.

Emmen, P. (2011). *A person-fit analysis of personality data* (Master thesis). Vrije Universiteit, Amsterdam. Retrieved from https://www.innovatiefinwerk.nl/sites/innovatiefinwerk.nl/files/field/bijlage/patrick_emmen.pdf

Emons, W. H. M. (2009). Detection and diagnosis of person misfit from patterns of summed polytomous item scores. *Applied Psychological Measurement, 33*(8), 599-619. doi: 10.1177/0146621609334378

Emons, W. H. M. (2003). *Detection and diagnosis of misfitting item-score vectors*. Amsterdam: Dutch University Press.

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

Emons, W. H. M., Glas, C. A. W., Meijer, R. R., & Sijtsma, K. (2003). Person fit in order-restricted latent class models. *Applied Psychological Measurement, 27*(6), 459-478. doi: 10.1177/0146621603259270

Glass, C. A. W., & Dagohoy, A. V. T. (2007). A person-fit test for irt models for polytomous items. *Psychometrika, 72*(2), 159-180. doi: 10.1007/s11336-003-1081-5

Hambleton, R. K., van der Linden W. J., & Wells, C. S. (2011). IRT models for the analysis of polytomous scored data: Brief and selected history of model building advances. In Nering M. L., & Ostini R. (Eds.), *Handbook of polytomous item response theory models* (pp. 21-42). New York, NY: Routledge.

Jiang, S., Wang, C., & Weiss, D. J. (2016). Sample size requirements for estimation of item parameters in the multidimensional graded response model. *Frontiers In Psychology*, *7*. doi: 10.3389/fpsyg.2016.00109

Junker, B., & Sijtsma, K. (2001). Nonparametric item response theory in action: An overview of the special issue. *Applied Psychological Measurement, 25*(3), 211-220. doi: 10.1177/01466210122032028

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

Lee, Y. S. (2007). A comparison of methods for nonparametric estimation of item characteristic curves for binary items. *Applied Psychological Measurement, 31*(2), 121-134. doi: 10.1177/0146621606290248

Lee, Y. S., Wollack, J. A., & Douglas, J. (2009). On the use of nonparametric item characteristic curve estimation techniques for checking parametric model fit. *Educational and Psychological Measurement*, *69*(2), 181-197. doi: 10.1177/0013164408322026

Liang, T., Wells, C. S., & Hambleton, R. K. (2014). An assessment of nonparametric approach for evaluating the fit of item response models. *Journal of Educational Measurement*, *51*(1), 1-17. doi: 10.1111/jedm.12031

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

Meijer, R. R. (2003). Diagnosing item score patterns on a test using item response theory-based person-fit statistics. *Psychological Methods, 8*(1), 72-87. doi: 10.1037/1082-989X.8.1.72

Meijer, R. R. (2004). *Investigating the quality of items in CAT using nonparametric IRT*. (LSAC Research Report Series No. 04-05). Newton, PA: Law School Admission Council.

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

Meijer, R. R., & Tendeiro, J. N. (2018). Unidimensional item response theory. In P. Irwing, T. Booth, & D. J. Hughes (Eds.), *The wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test development* (pp. 413-443). UK: John Wiley & Sons

Meijer, R. R., Egberink, I. J., Emons, W. H. M., & Sijtsma, K. (2008). Detection and validation of unscalable item score patterns using item response theory: An illustration with harter's self-perception profile for children. *Journal of Personality Assessment, 90*(3), 227-238. doi: 10.1080/00223890701884921

Meijer, R. R., Molenaar, I. W., & Sijtsma, K. (1994). Influence of test and person characteristics on nonparametric appropriateness measurement. *Applied Psychological Measurement, 18*(2), 111-120. doi: 10.1177/014662169401800202

Meijer, R. R., Niessen, A. S. M., & Tendeiro, J. N. (2016). A practical guide to check the consistency of item response patterns in clinical research through person-fit statistics: Examples and a computer program. *Assessment, 23*(1), 52-62. doi: 10.1177/1073191115577800

Molenaar, I. W. (2001). Thirty years of nonparametric item response theory. *Applied Psychological Measurement*, *25*(3), 295-299. doi: 10.1177/01466210122032091

Mousavi, A., Tendeiro, J. N., & Younesi, J. (2016). Person fit assessment using the Perfit package in R. *The Quantitative Methods for Psychology, 12*(3), 232-242. doi: 10.20982/tqmp.12.3.p232

Nydick, S. W. (2015) *catIrt: An R package for simulating IRT-based computerized adaptive tests. R package version 0.4-2*. Retrieved from http://CRAN.R-project.org/package=catIrt

Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika, 56*(4), 611-630. Retrieved from https://link.springer.com/article/10.1007/BF02294494

Rupp, A. A. (2013). A systematic review of the methodology for person-fit research in item response theory: Lessons about generalizability of inferences from the design of simulation studies. *Psychological Test and Assessment Modeling*, *55*(1), 3-38. Retrieved from http://www.psychologie-aktuell.com/fileadmin/download/ptam/1-2013_20130326/01_Rupp.pdf

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory*. USA: Sage Publications.

Sijtsma, K., Emons, W. H., Bouwmeester, S., Nyklícek, I., & Roorda, L. D. (2008). Nonparametric irt analysis of quality of life scales and its application to the world health organization quality of life scale (whoqol-bref). *Quality Of Life Research: An International Journal Of Quality Of Life Aspects Of Treatment,Care And Rehabilitation, 17*(2), 275-290. doi: 10.1007/s11136-007-9281-6

Sodano, S. M., & Tracey, T. J. (2011). A brief inventory of interpersonal problems–circumplex using nonparametric item response theory: Introducing the iip–c–irt. *Journal of Personality Assessment, 93*(1), 62-75. doi: 10.1080/00223891.2010.528482

Spoden, C. (2014). *Person fit analysis with simulation-based methods* (Doctoral dissertation). Universitäts bibliothek Duisburg-Essen. Retrieved from https://duepublico2.uni-due.de/servlets/MCRFileNodeServlet/duepublico_derivate_00038262/DISSERTATION_Spoden.pdf

Syu, J. J. (2013). *Applying person-fit in faking detection-the simulation and practice of non parametric item response theory* (Doctoral dissertation). National Chengchi University. Retrieved from http://nccur.lib.nccu.edu.tw/bitstream/140.119/58646/1/251501.pdf

Şengül Avşar, A., & Tavşancıl, E. (2017). Examination of polytomous items' psychometric properties according to nonparametric item response theory models in different test conditions. *Educational Sciences: Theory & Practice, 17*(2). doi: 10.12738/estp.2017.2.0246

Tendeiro, J. N. (2016). *Package "PerFit"*. Retrieved from https://cran.r-project.org/web/packages/PerFit/PerFit.pdf

Twiste, L. T. (2011). *Detection of unmotivated test takers through an analysis of response patterns: beyond person-fit statistics* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses database. (UMI No. 3478798)

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

Voncken, L. (2014). *Comparison of the l_z* Person-Fit Index and ω copying-index in copying detection* (First year paper). Universiteit van Tilburg. Retrieved from http://arno.uvt.nl/show.cgi?fid=135361

Waller, G. N., & Jones, J. (2016). Package "fungible". Retrieved from https://www.rdocumentation.org/packages/fungible

Wang, S. X. (2001). *Maximum weighted likelihood estimation* (Doctoral dissertation).University of British Columbia. Retrieved from https://open.library.ubc.ca/cIRcle/collections/ubctheses/831/items/1.0090880

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*(3), 427-450. doi: 10.1007/BF02294627

Birey Uyum İstatistiklerinin Farklı Test Koşullarında Çok Kategorili Puanlanan Maddeler İçin Karşılaştırılması

Giriş

Psikolojik ölçme araçları, bireyler hakkında karar vermede ve bireylerin öğrenme problemleri, gelişimsel problemleri ve psikolojik bozukluklarının tanımlanması gibi amaçlarla kullanırlar. Özellikle psikolojik tanı ve tedavilerde bireysel test puanlarına odaklanılacağı açıktır (Emons, 2003, 2009). Bu nedenle bireysel test puanlarının geçerliği eğitimde ve psikolojik değerlendirmelerde araştırılması gereken önemli bir konudur.

Örneğin bir birey sınavda kaygılı olmasından dolayı sınavdaki kolay maddelere yanlış cevap verebilir. Bu durum kişinin yeteneğinin, gerçek yeteneğinin altında kestirilmesine neden olabilmektedir. Bir başka örnek ise düşük yetenekli bireylerin etraflarında bulunan yüksek yetenekli bireylerden kopya çekme durumlarıdır. Bu durumda bireyin yeteneği, gerçek yeteneğinin üstünde kestirilir. Motivasyon eksikliğine dayalı olarak testin ciddiye alınmaması, bilişsel testlerde konsantrasyon problemleri, kişilik testlerinde sahte yanıt verme durumları normal olmayan madde puanlarına kaynaklık etmektedir. Tüm bunların sonucunda bireylerin yeteneğiyle ilgili yapılan kestirimlerin hatalı olacağı açıktır (Emons, 2003, 2008; Sijtsma & Molenaar, 2002).

Uyumsuz madde puanları bireylerin puanlarını arttırarak bireyin yeteneğinin gerçek yeteneği üzerinde kestirilmesine neden olabileceği gibi uyumsuz madde puanları bireylerin puanlarını azaltarak bireyin yeteneğinin gerçek yeteneği altında kestirilmesine neden olabilir. Buna göre kopya çekenler ya da şans başarısı yüksek olan şanslı yanıtlayıcıların puanları yapay olarak yüksek kestirilirken, test uygulamasının başında kaygılı, testi sonuna kadar yanıtlamayan, ya da dil problemi olan bireylerin puanları gerçekte olduğundan yapay olarak düşük kestirilir (Meijer, 1996). Ayrıca bazen madde içeriği ile ilgili bilgisi olmayan, maddeleri kendilerine göre yorumlayan, yanıtlarını yanlış kodlayan (kodlama sırasında kaydırma yapan) bireyler de uyumsuz madde puan örüntülerine sahip olacaklardır. Bu bireyler için kestirilen puanlar, gerçekte olduğundan daha yüksek veya düşük olabilir (Meijer, 1996). Bütün bu durumlarda bireylerin doğru değerlendirilemeyecekleri açıktır. Bu nedenle test sonuçlarına göre bireyler hakkında doğru kararlar verebilmek için bireysel madde puan örüntülerinin geçerliğini değerlendirmek önem taşımaktadır.

Birey uyum analizlerinin amacı seçilen/önerilen ölçme modeline göre bireysel test puanlarının uyum gösterip göstermediğini belirlemek ve bireysel test puan vektörlerini tanımlamaktadır (Meijer & Sijtsma, 2001). Bu amaç için birey uyum istatistikleri (BUİ) kullanılır. BUİ'ler bireylerin test maddelerine verdikleri yanıtlardan beklenmedik test performansını ortaya çıkarır (Meijer & Sijtsma, 2001). BUİ'ler bireyler hakkında önemli kararlar vermede geçersiz puanları ortaya çıkararak daha geçerli sonuçlara ulaşılmasında önemli rol oynarlar (Emons, 2008).

BUİ'ler genellikle parametrik ve parametrik olmayan istatistikler olacak şekilde iki kategoride incelenmektedir (Karabatsos, 2003; Mousavi, Tendeiro, & Younesi, 2016). Parametrik BUİ'ler (P-BUİ) parametrik madde tepki kuramına (PMTK), parametrik olmayan BUİ'ler (PO-BUİ) parametrik olmayan madde tepki kuramına (POMTK) dayalıdır (Karabatsos, 2003). P-BUİ ve PO-BUİ arasındaki temel fark, dayandıkları madde tepki kuramıdır. POMTK modellerinin getirdiği birtakım avantajlar, PO-BUİ'lere de yansımaktadır. PO-BUİ'ler için verinin POMTK modeline uyum göstermesi gerekmektedir (Emons, 2003). Özellikle verinin POMTK modellerinden Mokken Homojenlik Modeline (MHM) uyum göstermesi, diğer bir deyişle tek boyutluluk, yerel bağımsızlık ve madde karakteristik eğrilerinin monotonluğu varsayımlarının sağlanması gerekmektedir (Emons, 2008). Literatürde çok kategorili puanlanan maddeler için en fazla kullanılan P-BUİ'nin l_z^p istatistiği, PO-BUİ'lerin G^p, G_N^p ve $U3^p$ istatistikleri olduğu ifade edilmektedir (Emons, 2008; Rupp, 2013).

Birey uyum analizleri eğitimde ve psikolojide önemli bir konu olarak ele alınmaktadır. Özellikle başarı testleri ve bilişsel testlerde başarıyla uygulanmaktadır (Meijer & Sijtsma, 2001). Eğitim çalışmalarında (örneğin müfredattaki tutarsızlıkların belirlenmesinde, Harnisch & Linn, 1981), bilişsel psikoloji çalışmalarında (öğrenme stratejilerinin belirlenmesi, Tatsuoka & Tatsuoka, 1982), kültürler arası karşılaştırmalar (farklı dil gruplarından gelen bireylerin test puanlarının değerlendirilmesi ve karşılaştırılması, van der Flier, 1982), kişilik ölçme çalışmalarında (kişilik ölçme amacıyla geliştirilen ölçme araçlarında sahte yanıtların belirlenmesi, Dodeen & Darabi, 2009; Ferrando, 2004, 2009, 2012; Reise & Waller, 1993; Woods, Oltmanns, & Turkheimer, 2008; Zickar & Drasgow, 1996), örgüt psikolojisi çalışmalarında (bireylerin seçilen test için beklenmedik madde puan vektörlerini açıklama, Meijer, 1998), tutumların değerlendirilmesi (Curtis, 2004), sağlık araştırmaları (Custers, Hoijtink, van der Net & Hel, 2000; Tang ve diğerleri, 2010) örnek olarak verilebilir (akt., Emons, 2003; Rupp, 2013). BUİ'ler psikolojik değerlendirmelerde de (Conijn, Emons, De Jong & Sijtsma, 2015; Meijer, Egberink, Emons & Sijtsma, 2008) başarıyla uygulanmaktadır.

Yapılan literatür taramasında araştırmacıların; yeni BUİ'ler geliştirdikleri ve yeni geliştirilen bu BUİ'leri çeşitli test koşullarında inceledikleri (Emons, 2008; Glass & Dagohoy 2007; Karabatsos, 2003; Twiste 2011; van der Flier, 1982), uyumsuz madde puanlarının gerçek veri setlerinde belirledikleri (Egberink, 2010; Emmen, 2011; Meijer, 2003; Spoden, 2014) ve en iyi performans gösteren BUİ'leri belirledikleri (Emons, 2008; Karabatsos, 2003; Syu, 2013; Voncken, 2014) görülmüştür. Rupp'un (2013) çalışmasında da BUİ ile ilgili literatür taranmıştır. Yapılan bu çalışmada BUİ'lerin özellikle ikili puanlanan maddelerde daha fazla çalışıldığı, çok kategorili puanlanan maddelerde yapılan çalışmaların çok sınırlı olduğu ifade edilmiştir. Bununla birlikte yapılan literatür taramasında simülatif olarak üretilen veriler üzerinde BUİ'lerin özellikle küçük örneklemler ve çarpık dağılımlar gibi çeşitli test koşullarında daha fazla araştırılması gerektiği görülmüştür.

Çalışmanın amacı

Bu çalışmanın genel amacı P-BUİ ve PO-BUİ'lerin çok kategorili puanlanan maddelerden oluşan testlerde etkililiklerinin belirlenmesidir. Belirlenen amaç doğrultusunda aşağıdaki araştırma sorularına cevap aranmıştır:

1. BUİ'lere göre belirlenen uyumsuz madde puanlarına sahip kişilerin oranı; örneklem büyüklüğü, yetenek dağılımı, test uzunluğu ve madde ve kişilerin manipülasyonuna bağlı olarak oluşturulan anormallik durumlarına göre nasıl değişmektedir?
2. Farklı test koşullarında en iyi performansı gösteren BUİ hangisidir?

Yöntem

Bu araştırma BUİ'lerin, simülatif olarak oluşturulan test koşullarında, etkililiklerinin belirlenmesinin amaçlandığı temel araştırmadır.

Veri simülasyonu

Bu araştırmada çok kategorili puanlanan maddeler Samejima'nın Dereceli Tepki Modeline (DTM) göre üretilmiştir. Bu araştırmada POMTK'ya dayalı PO-BUİ'ler araştırmasına rağmen, parametrik DTM'ye göre veri üretilmesinin nedeni DTM'ye uyumlu olan veri setinin aynı zamanda MHM'ye uyumlu olmasıdır (Emons, 2008; Sijtsma, Emons, Bouwmeester, Nyklícek & Roorda, 2008). Verilerin üretilmesinde R programı kullanılmıştır. DTM'ye uygun verilerin üretilmesinde "catIRT" paketi (Nydick, 2015), çarpık dağılımlı veri setlerinin üretilmesinde "fungible" paketi (Waller & Jones, 2016) kullanılmıştır. Bu araştırmada simülatif verilerin üretilmesinde aşağıdaki adımlar izlenmiştir:

1. Belirlenen test koşullarında DTM'ye uyumlu veri setleri üretilmiştir.
2. Araştırmanın amacı doğrultusunda, veri setleri uyumsuz madde puanı içerecek şekilde (düşük ve yüksek oranlarda) manipüle edilmiştir.
3. Manipüle edilen veri setlerinde uç değerler belirlenmiş (tüm maddelerde kesinlikle katılıyorum veya hiç katılmıyorum kategorilerini seçenler) ve analiz dışı tutulmuştur. BUİ'lerin uç değerlerde yeteri kadar bilgi vermemesi (Emons, 2008), uç değerlerin atılmasının temel nedenidir.
4. Yetenekler ağırlıklandırılmış maksimum olasılığa (weighted maximum likelihood estimation) göre kestirilmiştir. Yetenekler kestirilirken veri üretimindeki gerçek madde parametreleri kullanılmıştır.
5. Farklı test koşullarında uyumsuz madde puanlarının belirlenmesi için BUİ'ler, Tendeiro (2016) tarafından geliştirilen "perfit" paketi kullanılarak kestirilmiştir.

Bu araştırmanın bağımsız değişkenleri; dört farklı örneklem büyüklüğü (100, 250, 500 ve 1000), üç farklı örneklem dağılımı (normal dağılan, sağa çarpık dağılan ve sola çarpık dağılan), iki farklı test uzunluğu (10 maddelik ve 30 maddelik test) ve iki farklı uyumsuzluk (düşük ve yüksek düzeylerde) oranıdır. Bağımlı değişkenleri ise deneysel I. Tip Hata oranları ve bu değerler için hesaplanan güç değerleridir. Bu araştırmada dört farklı BUİ (l_z^p, $U3^p$, G_N^p ve G^p) için I. Tip Hata oranları ve güç değerleri hesaplanmıştır.

Literatürde uyumsuz madde puanlarına neden olabilecek çeşitli davranışlardan bahsedilmiştir. Bu araştırmada *dikkatsiz ve özensiz davranışlar* dikkate alınmıştır. Bazı test uygulamalarında bireyler maddeleri rastgele cevaplarlar, maddeleri yanlış okurlar, maddeleri okumazlar ya da kodlama hatası yaparlar. Bu durumlar dikkatsiz ve özensiz davranışlara örnek olarak verilebilir (Emons, 2008). Bu araştırmada, bu davranışa yönelik uyumsuz madde puan vektörleri Emons'un (2008) çalışmasında olduğu gibi tek biçimli dağılımdan yararlanarak oluşturulmuştur.

Sonuç ve Tartışma

Bu araştırmanın genel amacı, P-BUİ ve PO-BUİ'lerin etkililiklerinin çok kategorili puanlanan maddelerden oluşan test koşullarında etkililiklerinin belirlenmesidir. Araştırma sonucunda beklendiği gibi, hesaplanan BUİ'ler için, I. Tip Hata oranı arttıkça uyumsuz madde puanına sahip bireylerin belirlenme oranı artmıştır. Araştırmada oluşturulan test koşullarında madde sayısı ve uyumsuz madde puan vektörleri arttıkça uyumsuz madde puanı belirleme oranı/güç artmıştır. Simülasyon sonuçları örneklemin dağılım şeklinin uyumsuz madde puanlarını belirlemede küçük bir etkisinin olduğunu göstermiştir. Diğer bir deyişle yetenek dağılımının şekli, uyumsuz madde puanı belirlemede bu araştırmadaki test koşullarına göre önemli bir faktör değildir. Genel olarak örneklem büyüklüğü, uyumsuz madde puanı oranlarını etkilemiştir. Örneklem büyüklüğü artıkça uyumsuz madde puanlarının belirleme oranları artmıştır. Araştırmanın bu bulgusu Syu'nun (2013) bulgularıyla farklılaşmıştır. Syu (2013) çalışmasında l_z^p, G^p ve $U3^p$ istatistiklerini araştırmıştır. Syu (2013) oluşturduğu test koşullarında örneklem büyüklüğünün çok küçük farklılıklar oluşturduğunu ancak seçilen koşulların BUİ'lerle ilgili yeterli bilgi veremeyeceğini de belirtmiştir.

Özetlenecek olursa nominal I. Tip Hata oranları artıkça, uzun testler kullanıldıkça ve manipüle edilen uyumsuz madde puanlarının oranı artıkça, uyumsuz madde puanlarının belirlenmesinin oranı da artmaktadır. Bu bulgu literatürdeki diğer araştırma bulgularına paraleldir (Emons, 2008; Karabatsos, 2003; Meijer & Sijtsma, 2001; Voncken, 2014).

Araştırmada genel olarak G^p istatistiğinin en iyi performansa sahip BUİ olduğu görülmüştür. Ancak özellikle uzun testlerde parametrik l_z^p istatistiğinin daha iyi performans gösterdiği de belirtilmelidir. Kısa testlerde ve küçük örneklemlerde G^p istatistiğinin daha iyi performans göstermesi, Emons (2008) ve Syu'nun (2013) araştırma bulgularına paraleldir. Syu (2013) çalışmasında küçük örneklemlerde PO-BUİ'lerin daha iyi performans gösterdiğini belirtmiştir. Ek olarak bu araştırmada BUİ'lerin uyumsuz madde puanlarını belirleme oranları, birbirlerine yakın değerler vermiştir. PO-BUİ'lerde özellikle $U3^p$ ve G_N^p birbirine oldukça yakındır. Uyumsuz madde puanlarını belirleme oranı en fazla $\alpha = .20$ düzeyinde olmuştur. Bu durum literatüre paraleldir (Emons, 2008; Meijer, 2003; Voncken, 2014).

Araştırma sonuçlarına göre dikkatsiz ve özensiz davranışların kaynaklık ettiği uyumsuz madde puanlarının belirlenmesinde uzun testlerin tercih edilmesi önerilebilir. Ancak uzun testler pratikte her zaman çok kullanışlı değillerdir. PMTK modelleri de parametrelerin doğru kestirilmesi için büyük örnekleme duyulan ihtiyaçtan dolayı çok kullanışlı değildir. Bu durumda PMTK modellerine göre daha az sınırlayıcı olan POMTK modellerinden MHM (Junker & Sijtsma, 2001; Meijer, 2004; Molenaar, 2001) kullanılarak uyumsuz madde puan örüntüleri PO-BUİ'lerle belirlenebilir.

Bu araştırma oluşturulan test koşulları dikkate alındığında özellikle küçük örneklem büyüklüklerinde ve kısa testlerde PO-BUİ'lerin kullanılması önerilebilir. Bu araştırmada kayıp veri içeren veri setleri üretilmemiştir. Belirlenen test koşullarında kayıp verilerin BUİ'lerin performanslarını nasıl etkiledikleri araştırılabilir. Araştırmada belirlenen test koşullarının dışında, farklı test koşulları oluşturularak BUİ'lerin etkililikleri belirlenebilir. Ayrıca bu araştırmada kullanılan istatistikler, gerçek veri setlerine kullanılarak araştırmanın bulgularıyla karşılaştırılabilir.