Simulation of comparing the sensitivity between response model and response time model detecting aberrant behavior

Danang Kamal Musthafa
Supranto
UIN Syarif Hidayatullah Jakarta

https://doi.org/10.37517/978-1-74286-697-0-14

Danang is a graduate in Psychometrics from the UIN Syarif Hidayatullah Jakarta. He is an experienced data analyst with deep interest in psychological assessment. He excels in R Program and other statistical software such as SPSS, JASP, and LISREL. He currently works as Research Assistant at the Australian Council for Educational Research (ACER) Indonesia.

Abstract

This study aims to determine whether there is a difference in detection rates between response model and response time model to detect aberrant behavior. Also, to determine examinee's ability estimation accuracy of each model along with checking the strengths and weakness when there is an aberrant behavior in testing data especially CBT. This research is a simulation study where concentrate on test-length, sample-size, and aberrant level with 50 replications. Analyzing parameter recovery from replicated data to check the strengths and weaknesses of each model. Further, comparing the lz person-fit that using response's data and response time's data to see which one the most sensitive is from both models. Moreover, the estimation of examinee's ability also compared to see the accuracy of each model. The result shows that increment of aberrant level would make item parameter estimation more bias for both models. Detection rates using response time (or response time model) indicates more sensitivity than detection rates using response (or response model). Both models also showed there was an increment estimation for examinee who doing aberrant behavior, so becomes bias and invalid if used as decision-making.

Introduction

Theoretically, when a person cheats to get correct answers during exam is called aberrant behavior. The presence of aberrant behavior caused by several things such as anxiety during the test, boredom, out of time to answer all items, low motivation, guessing, random responding, cultural bias, discrepancy of response intentionally, and misunderstand of test instruction (Birenbaum, 1985; Seo & Weiss, 2013). When this behavior occurs, there's a great consequence for test developer or test taker itself (Drasgow et al., 1987; Fox & Marianti, 2017; Karabatsos, 2003; Man et al., 2018). For test developers, aberrant behavior will reduce the accuracy of the test. Mousavi and Cui (2020) showed that test information function and the accuracy of item parameter estimation have decreased because of aberrant behavior. A study (Liu et al., 2019) that was conducting a simulation under Multidimensional Item Response Theory (IRT) also showed that aberrant behavior have impacts on model fit, such as the construct validity and reliability of the test. Therefore, methods to detect aberrant behavior are urgently required.

Detection of aberrant behavior normally conducted in IRT approach using person-fit statistics (Lord, 1980; Man et al., 2018). The goal is identifying an examinee who has a response pattern differ to expected response from a model, hence it can distinguish an examinee who classified as aberrant and non-aberrant is (Karabatsos, 2003; Reise, 1990). However, detection of aberrant behavior using person-fit statistics is analyzing unexpected response vector, so it known well as analysis of person-misfit (Meijer & Sijtsma, 1995; Reise, 1990). That's it, a person-fit that analyzing a response vector only shows a person which his/her response deviates but unable to know what type of aberrant behavior s(he) did during a test. In order to detect aberrant behavior more accurately and a transition paper-based test (PBT) into computer-based test (CBT) in current era, which technology grows rapidly (Fox, 2012, p. 227; Lee & Chen, 2011), since 1983, response time modelling as extension of IRT model has been developed as well as detection of aberrant behavior using response time (RT) data.

RT has been studied in IRT approach as an extension of IRT which accounts speed and response accuracy in a model (Thissen, 1983). Since the presence of response time modelling, aberrant behavior is expected to be more accurately detected with the availability of RT. Qian et al. (2016) whom analyzed the residual of log-response time illustrated there was an indication item pre-knowledge. The study statistically showed that aberrant behavior was significantly more in 2012 than 2010 which assumed as item pre-knowledge. Another study under lognormal response time modeling (Marianti et al., 2014) which conducted using simulation and real data also showed that person-fit was detecting 20%. When time discrimination accounted, resulting an increment around 34%. Thus, RT will help to detect aberrant behavior during a test.

Although several studies showed an increment on detection rates of aberrant behavior, cheating behavior is still the most difficult behavior to detected (Karabatsos, 2003). As a result, the score of examinee obtained from cheating would be questionable its validity since incongruent with a measurement model and becomes unfair if used as decision-making (de la Torre & Deng, 2008; Reise & Due, 1991). Therefore, this study aims to compare detection rates between response model and response time model as well as seeking the strengths and weaknesses of each model when examinees have aberrant behavior during a test.

Aberrant Behavior

When an examinee gives responses that differ from a model, there will be a discrepancy between observed response and expected response from the model. This discrepancy is considered as misfitting response, aberrant behavior, aberrant response, unexpected response, improbability response, anomalous behavior, etc. However, from several names , the most common name used is aberrant behavior (Meijer, 2003; Meijer & Sijtsma, 2001; Molenaar & Hoijtink, 1990; Mousavi & Cui, 2020; Wright & Stone, 1979, p. 66).

Imagine a person with a high ability taking a test with 30 items and the items on the test have variation of difficulty level, from the easiest to the hardest. Assume those items are sorted from the easiest to hardest, the person mentioned previously, unfortunately incorrectly answers five easy items but correctly answers the rest items. It is impossible if a person with a high ability giving incorrect answer on easy item, and this circumstance known as aberrant (Meijer & Sijtsma, 1995). Hence, the consequence is the score will be spurious (spuriously high or spuriously low) (Karabatsos, 2003). There are several aberrant behavior types (Karabatsos, 2003; Lee & Chen, 2011; Meijer, 1996; van der Linden & Guo, 2008; Wang et al., 2018):

1. *Sleeping behavior*
   When a test begins from easy to hard, a person that did not check again his/her answer on easy items and gives incorrect answer unintentionally. Therefore, the proportion of correct response on easy items is smaller than on medium and hard items.
2. *Cheating*
   Cheating is a behavior that unfairly gets correct answer on an item that s(he) actually unable to answer correctly. If a person has low ability and cheats (such as copying answer for his/her neighbor) so s(he) will correctly answer on item that hard for him/her.
3. *Careless responding*
   An examinee's behavior occurs when incorrectly answers an item intentionally and s(he) knows the correct answer.
4. *Creative responding*
   When an examinee incorrectly answers on easy items because s(he) interprets those items uniquely and resulting creative way to answer it. This behavior frequently appears when a competent person perhaps finding items that too easy to him/her, also assume that those items are too easy to be answered (or too simple to be true).
5. *Lucky guessing*
   An examinee luckily and correctly answers on items where s(he) doesn't know the answer or unable to gives the correct answer. In general, this behavior is identical giving a quick answer but has low accuracy.
6. *Plodding*
   This behavior can be exemplified in Guttman model where items have sorted from easy to hard. There's a chance that an examiner will slowly and methodically answer the items. Also, rejecting to answer next item until s(he) can solve an item after find the correct answer.
7. *Alignment errors*
   Discrepancy of response pattern because of mistake giving correct answer. The assumption is paper-based test (PBT), while the answer sheet and the questioned sheet are separated. There's a chance that an examiner forgets s(he) skipped a certain item and hasn't solved it yet. Unfortunately, s(he) gives answer on the skipped item. As a result, several items will incorrectly answer.
8. *Random responding*
   There will be a situation when a test is administered, and an examinee randomly selects the answer on each item especially test in multiple-choice.
9. *Deficiency of ability*
   This behavior can be exemplified when a test consists of two sub-ability such as sub-ability A dan sub-ability B. If the easy items measure A and hard item measure B, a person that has knowledge of A would have correct answer's proportion higher than correct answer's proportion on B.
10. *Memorization*
    This behavior doesn't occur directly during a test but would happen when a test is repeated so s(he) memorizes the items that have already been given to him/her.
11. *Item Pre-knowledge*
    When there's an unusual combination of answering items and response time. The combination means an examinee correctly answers an item in a relatively quick whereas the item has small probability of correct answer (hard item).
12. *Pacing*

A situation where an examinee only focuses on items that s(he) can answer and hopes to maximize the total score. This behavior often leaves certain item that hasn't answer yet resulting empty response or omit.

Person-fit

Person-fit is a statistical method that is used in IRT approach identifying to what extent the discrepancy between examinee's observed response and expected response from a chosen model. Person-fit statistics can be distinguished into two groups (Karabatsos, 2003; Meijer & Sijtsma, 1995, 2001), such as parametric and non-parametric where we can see on Table 1.

Table 1. 36 Person-fit Statistics

Non-Parametric Person-Fit Statistics (11)	Parametric Person-Fit Statistics (25)
G (Guttman, 1944, 1950)	*U* (Wright & Stone, 1979)
*G** (van der Flier, 1977)	*ZU* (Wright, 1980)
r_{pbis} (Donlon & Fischer, 1968)	*lnU* (Wright & Stone, 1979)
C (Sato, 1975)	*W* (Wright, 1980)
MCI (Harnisch & Linn, 1981)	*ZW* (Wright, 1980)
U3 (van der Flier, 1980)	*lnW* (Wright & Stone, 1979)
ZU3 (van der Flier, 1982)	ECI1, ECI2, ECI3, ECI4, ECI5, ECI6, ECIlz, ECIl2z, ECI4z, ECIl6z (Tatsuoka, 1984)
H^T (Sijtsma, 1986; Sijtsma & Mejer, 1992)	*l* (Levine & Rubin, 1979)
A, D, E_i (Kane & Brennan, 1980)	l_z (Drasgow, Levine, & Williams, 1985)
	M (Molenaar & Hojtink, 1990)
	M(*p-value*) (Bedrick, 1997)
	Item-Grouping Person-Fit Statistics
	D(θ) (Trabin & Weiss, 1983)
	l_{zm} (Drasgow, Levine, & Mclaughlin, 1991)
	UB (Smith, 1986)
	ZUB (Smith, 1986)
	lnUB (Wright & Stone, 1979)

Sources: Karabatsos (2003)

The most obvious difference between parametric and non-parametric approach is in the measurement scale while a parametric leads to measurement on interval or ratio, whereas a non-parametric leads to measurement on ordinal scale (Meijer & Sijtsma, 1995).

Method

This study used simulation where data was generated under IRT approach which would be compared in several conditions. The IRT model selected for response model and response time model was 2PL which consists item difficulty dan item discrimination. Condition in this study two IRT approach (response model and response time model) with aberrant level of N (5%, 10%, and 20% from sample-size), two different test-lengths (20 items and 40 items), also two different sample-sizes (N = 500 and N = 2.000). Therefore, the simulation study has a design 2 x 3 x 2 x 2 = 24 model data with 50 replications as well.

Manipulation of aberrant behavior conducted by taking random sample of ability $\theta < 0$ based on previous study (Maeda & Zhang, 2020) and considering aberrant level. The randomly sampled person considered a person who was doing aberrant behavior during a test. Next, items which have item difficulty ≥ 1.0 would be chosen and five items randomly sampled as item with aberrant behavior. The response from the generated model changes into a correct response on five selected items. The assumption is a person with low ability on five hard items

selected previously would answer them correctly. So, the probability of correct answer would become 100% which an indication of cheating. To make it more realistic, following on previous study (Man et al., 2018) that a certain value of time for selected items was added using uniform distribution around 30 to 120 seconds. Value that generated by uniform distribution added into response time of aberrant person on selected items. This step is assumed that aberrant persons need more addition time to copy answers from his/her neighbor (Man et al., 2018).

In order to generate data and estimate the model, this study used R Program version 4.1.2 (Team, 2021) with several packages such as "mirt" (Chalmers, 2012) for estimating response model, "PerFit" (Tendeiro, 2021) for analysis of person-fit, and "LNIRT" (Fox et al., 2021) for estimating response time model. The estimation of models was using Bayesian approach with MCMC algorithm via Gibbs Sampler for response time model. Two chains of 10.000 iterations were run and the burn-in cycle of each iteration was 2.500. For response model, "mirt" package provides MHRM as Bayesian approach. Also, two chains of 10.000 iterations and 2.500 burn-in periods were run. But, when the maximum change of iteration reaches < 0.0001, the iteration was terminated which indicated as the model converged. All processing was done in R Program.

Severals methods of generating data like standard normal distribution, lognormal distribution, uniform distribution, and so on, were used in this study. Item discrimination generated using lognormal distribution with 0.0 of meanlog and 0.3 of standard deviation, whereas item difficulty generated using standard normal distributiton. All item parameters generated at two test-length (20 and 40 items). Multivariate normal distribution with 0.0 of mean dan 1.0 of variance was used for ability parameter

Using l_z statistics (Drasgow et al., 1985) for person fit in response model, while response time model was person-fit proposed by Marianti et al. (2014) The cut-off point for response model was $l_z < -1.645$ (one-tailed, $\alpha = 0.05$) based on previous studies (Reise & Due, 1991; Seo & Weiss, 2013). Meanwhile in the response time model, cut-off point with significant level at 5% ($\alpha = 0.05$) was used. Also, Type I and Type II error rates were analyzed, where Type I error refers to what extent the proportion of non-aberrant persons detected as aberrant (False Positive), whereas Type II error refers to what extent the proportion of aberrant persons detected as non-aberrant (False Negative) (Cizek & Wollack, 2017, p. 12; Maeda & Zhang, 2020). Thus, the detection rates is an index that calculated by 1 – Type II error (Man et al., 2018).

Nevertheless, evaluation criteria that is used to compare true parameter dan estimated parameter known as parameter recovery. This step also wants to identify the extent of the difference between generated data and estimated data. Several indices were used such as bias dan mean absolute difference (MAD) (Bulut, 2015; Feinberg & Rubright, 2016; Mousavi & Cui, 2020) shown on equation (1) dan (2).

$$Bias = \frac{\sum_{i=1}^{n}\left(\hat{\theta}_i - \theta_{True}\right)}{n}, \text{dan} \quad (1)$$

$$MAD = \frac{\sum_{i=1}^{n}|\hat{\theta}_i - \theta_{True}|}{n} \quad (2)$$

Where $\hat{\theta}_i$ is estimated parameter's value, θ_{True} is generated parameter's value, whereas *n* is the number of replications. Bias and MAD can be used for each parameter by means change θ to α for item discrimination, β for item difficulty, and so on. Mousavi and Cui (2020) suggested that range of -0.20 to 0.20 of recommended bias, whereas MAD is no more than 0.6. Each model was identified by both indices where value of *bias* and MAD that close to 0.0 which considered as better estimate.

Result and Discussion

Descriptive statistics of generated data

Descriptive statistics conducted to check whether generated data was according to simulation method. One of replicated data was chosen as representation of all replications. The result of generated data will be shown on Table 2 down below.

Table 2. Descriptive statistics of generated data

K	AL	Item Discrimination				Item Difficulty				Ability			
		Mean	SD	Min	Max	Mean	SD	Min	Max	Mean	SD	Min	Max
20	5	1.02	0.19	0.68	1.37	0.00	1.03	-1.84	1.72	0.00	1.00	-3.06	3.55
	10	1.02	0.20	0.44	1.35	0.00	1.10	-2.84	1.62	0.00	1.00	-3.49	2.73
	20	1.03	0.23	0.60	1.43	0.00	1.03	-1.47	1.47	0.00	1.00	-2.52	2.62
40	5	1.03	0.25	0.57	1.58	0.00	0.98	-1.77	2.68	0.00	1.00	-4.02	2.88
	10	1.05	0.29	0.24	1.83	0.00	0.78	-1.78	1.72	0.00	1.00	-3.35	2.83
	20	1.03	0.24	0.36	1.46	0.00	0.77	-1.70	1.57	0.00	1.00	-2.97	2.76

Note. K = test-length, AL = aberrant-level (%), Mean = average value, SD = standard deviation.
Sources. Personal Data (2022).

As we can see, true parameter of item discrimination which generated by lognormal distribution, has mean around 1.0 along with 0.2 of standard deviation. Highest value of item discrimination is 1.83, whereas the lowest was 0.24 at 40 items with aberrant level of 10%. At the moment, item difficulty that generated by standard normal distribution, has 0.0 of mean for all conditions along with around 1.0 of standard deviation. The lowest value of tem difficulty was -2.84 whereas the highest value was 2.68. Meanwhile for ability parameter that generated by multivariate normal distribution, shows 0.0 of mean with 1.0 of standard deviation. The lowest value was -4.02 and the highest value was 3.55.

Analysis of Parameter Recovery

Parameter Recovery on Sample-size 500 (N = 500)

Data which had been simulated and manipulated on each condition will be compare among estimated parameter and true parameter as evaluation on response model and response time model. The analysis will be used to diagnose the impact of aberrant on each model. To clarify the result, it will be separated into two analyses based on sample-size (N = 500 and 2.000). The following is the result of parameter recovery of each model on sample-size 500 respondents which has shown on Table 3.

Table 3. Parameter Recovery pada Sample-size 500

K	AL (%)	N	Response Model				Response Time Model			
			Bias		MAD		Bias		MAD	
			a	b	a	b	a	b	a	b
20	5	500	-0.062	0.055	0.131	0.389	0.032	-0.106	0.180	0.175
	10	500	-0.093	-0.049	0.162	0.405	0.105	-0.166	0.338	0.230
	20	500	-0.130	-0.163	0.211	0.332	0.247	-0.229	0.611	0.293
40	5	500	0.018	0.089	0.122	0.342	0.027	-0.071	0.141	0.121
	10	500	-0.025	0.084	0.142	0.342	0.078	-0.102	0.233	0.154
	20	500	-0.015	-0.049	0.161	0.275	0.198	-0.134	0.408	0.190

Note. K = test-length, AL = aberrant level (%), a = item discrimination, b = item difficulty, MAD = mean absolute difference, bold = exceed criteria index. Sources: Personal Data (2022).

According to Table 3, it can be seen that response model has fulfilled the criteria index. However, along aberrant level increased, the estimation of item discrimination and item difficulty have underestimated. It means that the higher aberrant's presence on a certain testing data, the lower item parameter estimation will be. When item discrimination decrease, the item's discriminating power to distinct low and high ability person will be lose . Whereas, the lower item difficulty estimation, the items will represent easier than they should.

In response time model, it can be seen that only at aberrant level of 20% and test-length of 20 items hasn't met evaluation criteria. In contrast with response model where item parameters decreased along with aberrant level raised, item discrimination on response time model has overestimated. Though item discrimination raised, the estimation result doesn't reflect what they should be. On item difficulty, it was consistent with response model as the estimation became underestimated and indicated that items became easier than they had to be.

Parameter Recovery on Sample-size 2000 (N = 2000)

Table 4. Parameter Recovery on Sample-size 2.000

K	AL (%)	N	Response Model				Response Time Model			
			Bias		MAD		Bias		MAD	
			a	b	a	b	a	b	a	b
20	5	2000	0.173	-0.026	0.220	0.342	0.035	-0.114	0.172	0.145
	10	2000	0.127	-0.030	0.271	0.342	0.117	-0.175	0.355	0.203
	20	2000	0.084	-0.196	0.350	0.275	0.504	-0.259	0.913	0.284
40	5	2000	0.173	0.005	0.232	0.314	0.027	-0.071	0.133	0.095
	10	2000	0.152	-0.020	0.268	0.295	0.096	-0.106	0.258	0.130
	20	2000	0.168	-0.163	0.310	0.246	0.211	-0.139	0.429	0.166

Note. K = test-length, AL = aberrant percentage (%), a = item discrimination, b = item difficulty, MAD = mean absolute difference, bold = exceed criteria index. Sources: Personal Data (2022)

On Table 4, response model has fulfilled the criteria indices (bias and MAD) on all conditions. However, item discrimination estimation has the maximum evaluation criteria. Along with the increment of aberrant level, thus the item discrimination dan item difficulty also decreased on 20 items or 40 items. The result is congruent to parameter recovery on sample-size 500 that has presented on Table 3.

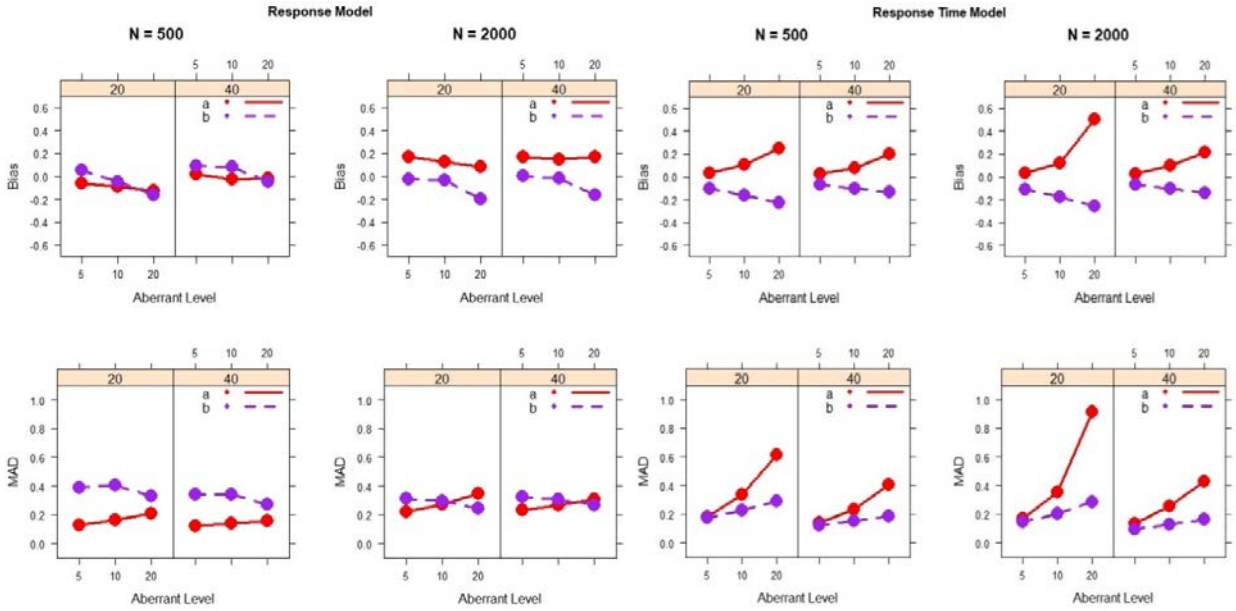


Figure 1. Interaction Plot of Parameter Recovery between Response Model and Response Time Model

Sources: Personal Data (2022)

On response time model, when aberrant level reaches 20%, item discrimination hasn't fulfilled the criteria index like bias < 0.2 and MAD < 0.6 on model at 20 items as well as at 40 items. But at 40 items, item discrimination only exceeded bias index. The result was also consistent to sample-size 500 on Table 3 where item discrimination suffers overestimation while item difficulty suffers underestimation.

To make it easier to understand Table 3 and Table 4, interaction plot of aberrant level with criteria indices of each model will be presented on Figure 1. As we can see the higher aberrant level on response model, the more underestimates would be on both item parameters. Whereas on response time model, it clearly has shown the higher aberrant level affects underestimation on item difficulty but overestimation on item discrimination.

Analysis of Detection Rates

The analysis of detection rates represents the sensitivity of each model detecting aberrant which replicated 50 times. The result of comparison of sensitivity between response model dan response time model to detect aberrant behavior shown on Table 5.

Table 5. Comparison of Detection Rates between Response Model and Response Time Model

AL (%)	N	Detection Rates											
		Response Model						Response Time Model					
		K = 20			K = 40			K = 20			K = 40		
		D	Type I	Type II	D	Type I	Type II	D	Type I	Type II	D	Type I	Type II
5	500	0.997	0.007	0.003	0.988	0.011	0.012	1.000	0.008	0.000	1.000	0.011	0.000
10	500	0.979	0.010	0.021	0.936	0.017	0.064	0.976	0.010	0.024	0.972	0.014	0.028
20	500	0.836	0.020	0.164	0.813	0.021	0.187	0.878	0.013	0.122	0.868	0.017	0.132
5	2000	1.000	0.008	0.000	0.999	0.012	0.001	1.000	0.008	0.000	1.000	0.011	0.000
10	2000	0.985	0.012	0.015	0.957	0.017	0.043	0.978	0.010	0.022	0.979	0.013	0.021
20	2000	0.860	0.020	0.140	0.833	0.021	0.167	0.887	0.015	0.113	0.876	0.017	0.124

Note. AL = aberrant level (%), N = sample-size, K = test-length, D = detection rates, Type I = non-aberrant detected as aberrant, Type II = aberrant detected as non-aberrant.
Sources: Personal Data (2022).

According to Table 5, it has shown that detection rates of both model decrease along with aberrant level increases. It means that the higher aberrant level, the more difficult aberrant behavior would be detected. Further, the detection rates on both models are vary where response time model is more sensitive to detect aberrant behavior rather than response model. The highest difference on both models is 5.5% where we can see on sample-size 500 at 40 items. Also, it can be seen that the higher aberrant level, the higher type i error on both models. This is an indication when the more aberrant persons present in a testing data, not only detection rates would be low but the chance of misclassifying non-aberrant persons as aberrant will increase as well. In simply words, it will be more difficult to decide who the aberrant persons are, and vice versa.

Analysis of Distance between Theta and Item Difficulty

The addition analysis in this study is analysis of distance between *ability* (θ) and item difficulty. The comparison between unmanipulated model and estimated model from manipulated data containing aberrant. Data was randomly chosen, and 3rd condition (N = 500, test-length = 20 items, and aberrant level = 20%) was selected from last replication which can be seen on Table 6.

Table 6. Distance between Ability and item Difficulty

Model	θ - b					
	AI		Non-AI		All Item	
	Mean	SD	Mean	SD	Mean	SD
True	-1.468	1.000	-1.441	1.000	0.000	1.027
RM	-1.088	1.088	-1.976	1.088	0.188	1.272
RTM	-0.319	0.615	-1.467	0.615	0.230	0.809

Note. AI = aberrant item, Non-AI = non-aberrant item, True = true parameter, RM = response model, RTM = response time model.
Sources: Personal Data (2022).

According to Table 6, we can see there was a shift between ability and item difficulty. Before the data was manipulated, the distance between ability and item difficulty was -1.468 (b = 1.468). It means when the value from the distance accounts in IRT model, selected samples have probability of correct answer under 50% on selected items. However, when the response of selected samples has been estimated on both models, it can be seen that the distance between ability and item difficulty became closer to zero. On response model, the distance was -1.088 whereas response time model's distance was -0.319 for all sample's mean. It means, aberrant behavior would affect the distance between ability and item difficulty. In other words, it becomes hard to detect by person-fit statistics.

Analysis of Ability Estimation's Accuracy

The accuracy of ability estimation was also analyzed by separating the aberrant and all samples (total) using criteria indices (bias and MAD). In this step, each model was compared to see which model has smallest bias and MAD as indicator of better accuracy. Also, the result will be separated into two tables which sample-size 500 and sample-size 2.000.

Accuracy Comparison on Sample-size 500 (N = 500)

Table 7. Ability Estimation on Sample-size 500

K	AL	N	Response Model				Response Times Model			
			Aberrant		Total		Aberrant		Total	
			Bias	MAD	Bias	MAD	Bias	MAD	Bias	MAD
20	5	500	0.807	0.815	0.059	0.355	0.660	0.665	0.000	0.301
	10	500	0.652	0.671	0.033	0.366	0.521	0.531	0.000	0.335
	20	500	0.491	0.533	0.063	0.373	0.414	0.437	0.000	0.392
40	5	500	0.250	0.323	0.070	0.335	0.338	0.357	0.000	0.211
	10	500	0.167	0.292	0.079	0.346	0.281	0.306	0.000	0.227
	20	500	0.087	0.259	0.056	0.310	0.266	0.295	0.000	0.275

Note. K = test-length, AL = aberrant level (%), Aberrant = aberrant samples, Total = all samples, N = sample-size, MAD = mean absolute difference. Sources: Personal Data (2022).

Based on Table 7, for aberrant sample at 20 items, it showed that ability estimation of response time model is lower than response model. It can be interpreted as the estimation of aberrant person's ability on response model became more spuriously high (SH) than response time model when the sample-size was small. Moreover, it also can be seen for all samples (total) ability's estimation of response time model had surprisingly small bias under three digits round which represented to zero. This phenomenon indicated that the response time model had excessively smaller difference between true ability parameter and estimated ability parameter than response model. Nonetheless, there was a different ability estimation when aberrant level reached 20% where MAD of response model smaller than response time model. It means, when data contains so many aberrant persons, did not affect so much to all samples (total) which contains non-aberrant persons.

Furthermore, at 40 items, it was supported with previous result. On this condition, for aberrant samples, response model showed better accuracy of ability estimation than response time model. It was proven by the smaller criteria indices as better accuracy. It means, response model will have better accuracy of ability estimation for aberrant persons in a longer test. However, for all samples (total), response time model showed better ability estimation than response model based on both indices which gives smaller value.

Accuracy Comparison on Sample-size 2000 (N = 2000)

Table 8. Ability Estimation on Sample-size 2.000

K	AL	N	Response Model				Response Times Model			
			Aberrant		Total		Aberrant		Total	
			Bias	MAD	Bias	MAD	Bias	MAD	Bias	MAD
20	5	2000	0.728	0.735	-0.010	0.325	0.646	0.650	0.000	0.301
	10	2000	0.511	0.540	-0.009	0.331	0.479	0.494	0.000	0.333
	20	2000	0.310	0.393	-0.022	0.332	0.445	0.463	0.000	0.447
40	5	2000	0.187	0.273	-0.013	0.289	0.350	0.362	0.000	0.210
	10	2000	0.053	0.236	-0.036	0.285	0.268	0.297	0.000	0.233
	20	2000	-0.002	0.227	-0.042	0.266	0.266	0.296	0.000	0.277

Note. K = test-length, AL = aberrant level (%), Aberrant = aberrant samples, Total = all samples, N = sample-size, MAD = mean absolute difference. Sources: Personal Data (2022).

Considering Table 8, at 20 items on aberrant samples, the response model has higher value than response time model on shorter test. Yet, on condition where aberrant level reached 20%, response model literally has smaller value than response time model. It can be interpreted as a response model had better ability estimation than response time model on aberrant level 20% when the sample-size is big enough even with shorter test. The result was

also consistent for all samples (total), where response time model had better estimation when aberrant level 5% to 10% only. Even though bias showed small value, MAD on aberrant level 20% showed response model smaller value which indicated as better accuracy.

Thereafter, at 40 items as longer test, response model had better estimation which smaller bias and MAD than response time model on aberrant samples. It means that when the test was longer and sample-size was big enough, the response model did not give enlargement estimation than response time model. However, it would be a different scenario when referring on ability estimation of all samples (total) where response time model had smaller value. Even though there was escalation of ability estimation on aberrant persons, still non-aberrant could be estimated approaching true ability parameter which means representing the true ability of examinee. At 20 % of aberrant level, there was slightly difference between response model and response time model where response model had smaller value of MAD. The interaction plot of accuracy is presented in Figure 2 down below for convenient.

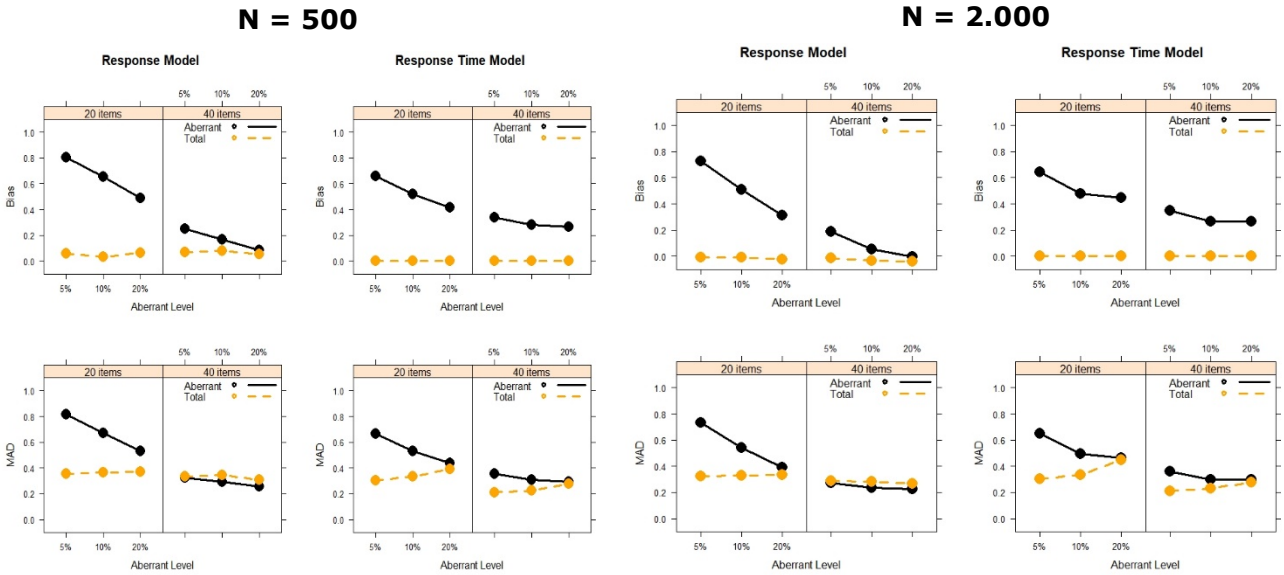


Figure 2. Interaction Plot of Accuracy between Response Model and Response Time Model

Sources: Personal Data (2022)

Conclusion

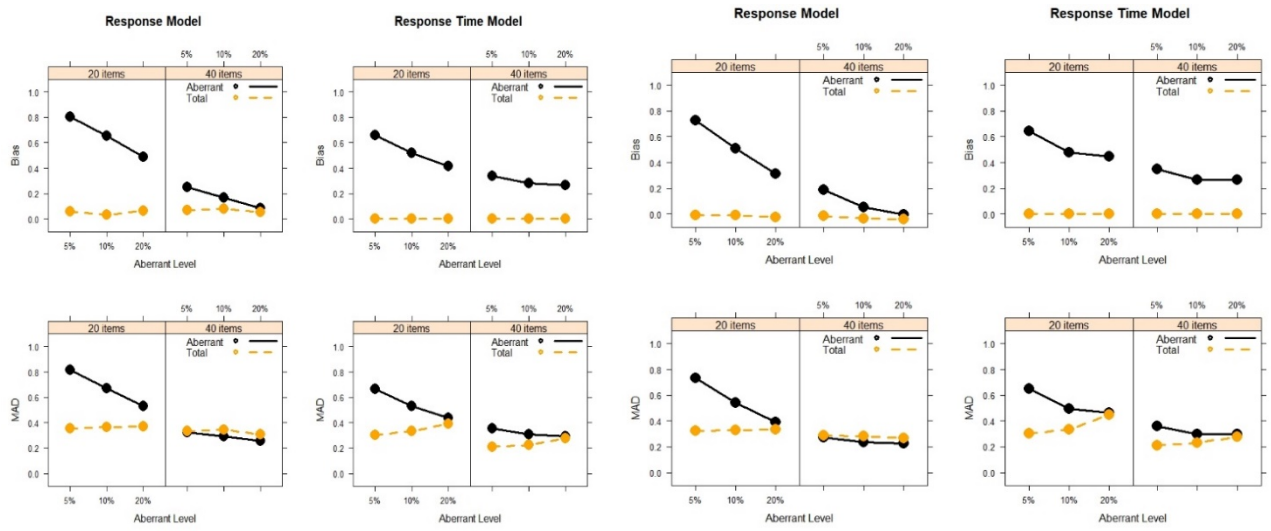
The aim of this study was to determine whether there is a difference of sensitivity detection rates between response model and response time model. Turned out, 9 out of 12 showed that response time model was more sensitive to detect aberrant behavior. It means, counting response time (RT) in response time model, resulting the detection rates became more sensitive than response model. However, the highest difference between both models was only 5.5%. According to the analysis of distance between ability and item difficulty, the smaller the difference ($|\theta - b|$), the more difficult to detect aberrant behavior especially if the value is less than two ($|\theta - b| < 2$) (Smith, 1986; Wright & Stone, 1979, p. 75).

In terms of the item parameter estimation, the finding support Maeda and Zhang (2020) where aberrant will directly affect item parameter estimation. As seen that when aberrant level reached 20%, many items would be estimated easier than they should be. The finding also showed that the ability estimation of both models suffered by the occurrence of aberrant behavior. As a result, the accuracy would be low and the aberrant persons have a spuriously high ability, which means would be questionable if it used as a decision-making in terms of testing (de la Torre & Deng, 2008; Molenaar & Hoijtink, 1990; Reise & Due, 1991).

In general, the finding of this study only as an illustration how the sensitivity of both models handling data containing aberrant behavior. Also, it would be unwise if the person-fit statistics were used as a decision-making of someone who is cheating or guesses. Indeed, a person-fit statistics will show us someone has a discrepancy between his/her observed response (or response time) patterns and expected response patterns (or response time) on a certain IRT model. If the person-fit statistics shows us that someone has a significant, it doesn't mean s(he) doing aberrant behavior during a test. Clearly, we need more evidence such as audit of the test security, directly monitoring or via cctv, proctoring, etc. Because, person-fit statistics is only a tool that help us to find aberrant behavior which sometimes can be statistically mistaken and never be sufficient as primary evidence of someone doing cheating or guessing, or whatsoever on a test.

Limitation and Suggestion

Among the limitations on current study are first, the fact that manipulation on this simulation study would never reflect the real situation of aberrant behavior during a real test. Second, the number of items that assumed aberrant behavior would appear were fixed as five items. It means that a short or longer test would have different proportion of aberrant (5 of 20 items or 5 of 40 items). Third, the addition time of aberrant behavior was also fixed for all aberrant persons. In reality, the addition time could be varied depending on a person, which sometimes could be faster or slower than in this simulation. Last, the aberrant level in this study was proportion of sample (% of N), in fact that a real test would be a different situation.



Further studies, especially a simulation study that related to response time model, may increase the variation of test-length such as range from 10 items to 80 items. Also, may varying other aberrant behavior such as plodding, random response, pacing, etc. Moreover, the aberrant level may be constructed by gradation which refers on how many items of each aberrant person has aberrant behavior on items depending on his/her ability. The variation of manipulating response time (RT) also may be randomly generated to make it more realistic. An analysis of ROC curve (Receiver Operating Characteristic) also may be conducted as well to determine which model has a better classification of aberrant. Other suggestions may apply to other IRT models (i.e., Rasch Model, 1PL, 3PL, etc.) that can be adjusted following research assumption. Among other things, many other person-fit statistics, which have a powerful detection rate beside l_z, may be utilized as well. Therefore, the detection rates of person-fit statistics could be considered.

References

Birenbaum, M. (1985). Comparing the effectiveness of several IRT based appropriateness measures in detecting unusual response patterns. *Educational and Psychological Measurement, 45*(3), 523–534. https://doi.org/10.1177/001316448504500309

Bulut, O. (2015). Applying Item Response Theory Models to entrance examination for graduate studies: Practical issues and insights. *Journal of Measurement and Evaluation in Education and Psychology, 6*(2), 313–330. https://doi.org/10.21031/epod.17523

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

Cizek, G. J., & Wollack, J. A. (2017). *Handbook of quantitative methods for detecting cheating on tests*. Routledge.

de la Torre, J., & Deng, W. (2008). Improving person-fit assessment by correcting the ability estimate and its reference distribution. *Journal of Educational Measurement, 45*(2), 159–177. https://doi.org/10.1111/j.1745-3984.2008.00058.x

Drasgow, F., Levine, M. V., & McLaughlin, M. E. (1987). Practical with Optimal and Indices Appropriateness. Applied Psychological Measurement, 11(1), 59–79.

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. https://doi.org/10.1111/j.2044-8317.1985.tb00817.x

Feinberg, R. A., & Rubright, J. D. (2016). Conducting simulation studies in psychometrics. *Educational Measurement: Issues and Practice, 35*(2), 36–49. https://doi.org/10.1111/emip.12111

Fox, J.-P. (2012). *Bayesian Item Response Modeling: Theory and Applications*. Springer.

Fox, J.-P., Klotzke, K., & Entink, R. K. (2021). Package ‘ LNIRT .’ 1–18. https://doi.org/10.18637/jss.v020.i07>.License

Fox, J.-P., & Marianti, S. (2017). Person-Fit statistics for joint models for accuracy and speed. *Journal of Educational Measurement, 54*(2), 243–262. https://doi.org/10.1111/jedm.12143

Karabatsos, G. (2003). Comparing the Aberrant Response Detection Performance of Thirty-Six Person-Fit Statistics. *Applied Measurement in Education, 16*(4), 277–298. https://doi.org/10.1207/S15324818AME1604_2

Lee, Y.-H., & Chen, H. (2011). A review of recent response-time analyses in educational testing. *Psychological Test and Assessment Modeling, 53*(3), 359–379. http://www.psychologie-aktuell.com/fileadmin/download/ptam/3-2011_20110927/06_Lee.pdf

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. https://doi.org/10.1080/15366367.2019.1584848

Lord, F. M. (1980). *Application of Item Response Theory to Practical Testing Problems*. Lawrence Erlbaum Associates.

Maeda, H., & Zhang, B. (2020). Bayesian Extension of Biweight and Huber Weight for Robust Ability Estimation. *Journal of Educational Measurement, 57*(1), 51–70. https://doi.org/10.1111/jedm.12240

Man, K., Harring, J. R., Ouyang, Y., & Thomas, S. L. (2018). Response Time Based Nonparametric Kullback-Leibler Divergence Measure for Detecting Aberrant Test-Taking Behavior. *International Journal of Testing, 18*(2), 155–177. https://doi.org/10.1080/15305058.2018.1429446

Marianti, S., Fox, J.-P., Avetisyan, M., Veldkamp, B. P., & Tijmstra, J. (2014). Testing for aberrant behavior in response time modeling. *Journal of Educational and Behavioral Statistics, 39*(6), 426–451. https://doi.org/10.3102/1076998614559412

Meijer, R. R. (1996). Person-Fit research : An iIntroduction. *Applied Measurement in Education, 9*(1), 3–8. https://doi.org/10.1207/s15324818ame0901

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. https://doi.org/10.1037/1082-989X.8.1.72

Meijer, R. R., & Sijtsma, K. (1995). Detection of aberrant item score patterns: A Review of recent developments. *Applied Measurement in Education, 8*(3), 261–272. https://doi.org/10.1207/s15324818ame0803_5

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

Molenaar, I. W., & Hoijtink, H. (1990). The many null distributions of person fit indices. *Psychometrika, 55*(1), 75–106. https://doi.org/10.1007/BF02294745

Mousavi, A., & Cui, Y. (2020). The effect of person misfit on item parameter estimation and classification accuracy: A simulation study. *Education Sciences, 10*(11), 1–15. https://doi.org/10.3390/educsci10110324

Qian, H., Staniewska, D., Reckase, M., & Woo, A. (2016). Using response time to detect item preknowledge in computer-based licensure examinations. *Educational Measurement: Issues and Practice, 35*(1), 38–47. https://doi.org/10.1111/emip.12102

Reise, S. P. (1990). A Comparison of Item- and Person-Fit methods of assessing model-data fit in IRT. *Applied Psychological Measurement, 14*(2), 127–137. https://doi.org/10.1177/014662169001400202

Reise, S. P., & Due, A. M. (1991). The Influence of test characteristics on the detection of aberrant response patterns. *Applied Psychological Measurement, 15*(3), 217–226. https://doi.org/10.1177/014662169101500301

Seo, D. G., & Weiss, D. J. (2013). lz Person-Fit Index to identify misfit students with achievement test data. *Educational and Psychological Measurement, 73*(6), 994–1016. https://doi.org/10.1177/0013164413497015

Smith, R. M. (1986). Person Fit in The Rasch Model. *Educational and Psychological Measurement, 46*(2), 359–372. https://doi.org/10.1177/001316448604600210

Team, R. C. (2021). R: A language and environment for statistical computing. https://www.r-project.org/index.html

Tendeiro, J. N. (2021). Package "PerFit." 1–56.

Thissen, D. (1983). Timed testing: An approach using Item Response Theory. In *New Horizons in Testing*. Academic Press. https://doi.org/10.1016/b978-0-12-742780-5.50019-6

van der Linden, W. J., & Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika, 73*(3), 365–384. https://doi.org/10.1007/s11336-007-9046-8

Wang, C., Xu, G., Shang, Z., & Kuncel, N. (2018). Detecting aberrant behavior and item preknowledge: A comparison of mixture modeling method and residual method. *Journal of Educational and Behavioral Statistics, 43*(4), 469–501. https://doi.org/10.3102/1076998618767123

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Mesa Press. https://research.acer.edu.au/measurement/1