2021 International Conference on Innovation in Open & Distance Learning (2021 INNODEL)
ISSN 2963-2870
Vol 2 2022

# DETECTING DATA IRREGULARITY BY CONSIDERING RESPONSES AND RESPONSE TIME

**Heru Widiatmo, Ph.D.**

Measurement Consultant at Universitas Terbuka (Indonesia)

***ABSTRACT***

*During a computerized test administration, either online or offline examination, the time that an examinee spends on an item can be easily recorded. This response time information, combined with item responses, could provide more information to detect data irregularities than responses alone. When some examinees answer multiple-choice test items much faster than other examinees, this could be an indication of a data irregularity. It could be occurred for many reasons, including pre-knowledge of the items and rapid guessing if running out of time at the end of a test. These aberrant behaviors, which cannot be detected from paper-based tests, could threaten test security of computerized tests, and invalidate the integrity of test results. Therefore, efforts should be made to detect data irregularities and further investigations may be needed to ensure the test results are as reliable, fair, and valid as possible. In addition, by taking care of data irregularities some researchers have shown better measures of ability (Bolsinova, De Boeck, & Tijmstra, 2017; De Boeck, & Minjeong, 2019; Marianti, Fox, Avetisyan & Veldkamp, 2014; Widiatmo & Wright, 2015).*

*Several methods can be used for excluding irregularities based on response times and responses (Ratcliff, 2003). One possible method is to apply a threshold method Wise and Kong (2015). This method is called Response Time Effort (RTE) that is the proportion of the items on which each examinee spent sufficient time. The other method is using a statistical model that could detect data irregularities (e.g., Anders, Alario, & Van Maanen, 2016; van der Linden, 2006; van der Maas, Molenaar, Maris, Kievlt, & Borsboom, 2011). Among them, van der Linden's (2006) developed a lognormal model to examine the relationship between item responses and latencies. This method was called the "effective response time" (ERT) in Meijer and Sotaridona (2006). ERT is defined as the time required for an examinee to answer an item correctly, and a chi-square distribution is used to check if the value is beyond a certain confidence level for given examinee ability and item parameters.*

*The purpose of this study is to investigate whether the RTE method and/or the ERT model can produce "cleaner" data than the current data cleaning method employed. There are three procedures of the data cleaning proposed in this study. The first is only using the RTE method, the second is only the ERT method, and the third is using those two methods together. For the third option, after excluding examinees using the first method, the remaining data are examined using the second method to investigate whether any examinees are needed to be excluded further.*

2021 International Conference on Innovation in Open & Distance Learning (2021 INNODEL)
ISSN 2963-2870
Vol 2 2022

*Three sets of data from the three procedures are calibrated into the 3-PL IRT that is the current calibration model used. The results will be compared among the procedures and with the current calibration procedure. How many items are fit to the 3-PL IRT is the criterion measured. It is expected that the more items are fit the model for a given procedure, the more preferable the procedure is.*

***Keywords**: Aberrant Behaviors, Computerized Test Administration, Item Response Theory, Response Time*

## 1. OVERVIEW

The time that an examinee spends on an item can be easily recorded for computer administered tests. This response time information, combined with item responses, could providemore information to detect data irregularities than responses alone. When some examinees answertest items much faster than other examinees, this could be an indication of a data irregularity that can occur for many reasons including pre-knowledge of the items and rapid guessing if running out of time at the end of a test. Variation in response times can also be important to consider. Theseaberrant behaviors, which cannot be detected from paper-based tests, could threaten test security of computerized tests and invalidate the integrity of test results. Therefore, efforts should be madeto detect data irregularities and further investigations may be needed to ensure the test results are as reliable, fair, and valid as possible. In addition, by taking care of data irregularities some researchers have shown better measures of ability ((Bolsinova, De Boeck, & Tijmstra, 2017; De Boeck, & Minjeong, 2019; Marianti, Fox, Avetisyan & Veldkamp, 2014; Widiatmo & Wright, 2015).

Several methods can be used for excluding irregularities based on response times and responses (Ratcliff, 1993). A threshold method was proposed by Wise and Kong (2005). They defined a response time boundary ($T_i$) between rapid-guessing and solution behaviors ($SB_{ji}$) for given examinee *j* and item *i* for a response time $RT_{ji}$. Mathematically, it is written as follows.

$$SB_{ji} = \begin{cases} 1 \; if \; RT_{ji} \geq T_i \\ 0 \; otherwise \end{cases} \qquad (1)$$

Ti is based on item length and whether or not the item uses a figure, an illustration, or reading material. The longer and the more complicated the item is, the higher $T_i$ is.

In this study, however, due to test security concerns the author cannot examine the features of the test items to define $T_i$. The author used 10-seconds as a threshold, given the time required to answer a multiple choice (MC) item of a Mathematics test should be above 10 seconds (Bridgeman, Laitusis, & Cline, 2007).

2021 International Conference on Innovation in Open & Distance Learning (2021 INNODEL)
ISSN 2963-2870
Vol 2 2022

Using the Wise and Kong (2005) method, an index over all items for examinee j, called Response Time Effort (RTE), is computed by:

$$RTE = \frac{\sum^{k} SB_{ji}}{k}$$

j

(2) where k is the total number of items. RTE is the proportion of the items on which examinee j spentsufficient time. If RTEi is close to 1, it indicates strong examinee effort to the test. If it is close to 0, then the examinee appears not to have performed solution behaviors for many items.

Several statistical models that incorporate item responses and response times might detectdata irregularities (e.g., Anders, Alario, & Van Maanen, 2016; van der Linden, 2016; van der Maas,Molenaar, Maris, Kievlt, & Borsboom, 2011). Among them, van der Linden's (2006) developed a lognormal model to examine the relationship between item responses and latencies. This method was called the “effective response time” (ERT) in Meijer and Sotaridona (2006). ERT is defined as the time required for an examinee to answer an item correctly, and a chi-square distribution is used to check if the value is beyond a certain confidence level for given examinee ability and itemparameters. The details of the method are discussed as follows.

## 2. METHOD

As discussed in van der Linden and van Krimpen-Stoop (2003), a loglinear model can be used to model response time and the model is written mathematically as follows:

$$\ln T_{ij} = \mu + \delta_i + r_j + \epsilon_{ij} \quad (3)$$

with $\epsilon_{ij} \sim N(0, \sigma^2)$ (4)

where, $lnT_{ij}$ is the natural logarithm of the time taken by examinee j to response item i, $\mu$ is a parameter indicating the general response time level for the examinee population on the item pool,

$\delta_i$ is the response time parameter required by item i, $r_j$ is the slowness parameter of examinee j, and $\epsilon_{ij}$ is a normally distributed interaction term between item i and examinee j with mean 0 and variance $\sigma^2$. Accordingly, the model distribution is $\ln T_{ij} \sim(\mu + r_i + \delta_j, \sigma^2)$. Therefore, the parameters of equation 3 can be estimated as follows:

$$\mu \equiv (\ln T_{ij}) \quad (5)$$

$$\delta_i \equiv (\ln T_{ij}) - \mu \quad (6)$$

$$r_j \equiv (\ln T_{ij}) - \mu \quad (7)$$

$$\sigma^2 \equiv (\ln T_{ij} - r_i - \delta_j)^2 \quad (8)$$

2021 International Conference on Innovation in Open & Distance Learning (2021 INNODEL)
ISSN 2963-2870
Vol 2 2022

Moreover, Meijer and Sotaridona (2006) proposed a regression approach for modifying equation 3 as ERT that is the time required by examinee j with ability $\theta_j$ to answer correctly item

i. The ERT can be applied to an examinee if he/she answers correctly for a given item i and if his/her expected probability of answering that item is greater than the guessing parameter ( ). The rationale for these two requirements is discussed in more detail in their paper.

The ERT for each item i and for examinee j is modeled by regression ln $T_{ij}$ on $\theta_j$ and $r_j$ as follows:

$\ln T_{ij} = \beta + \beta_1\theta_j + \beta_2 r_j + \epsilon_j$, (9) where $\ln T_{ij}$, $\theta_j$ *and* $r_j$ are known, the $\beta$'s are the regression coefficients, and $\epsilon_j$is an error term assumed to be normally distributed with mean 0 and variance $\sigma^2$. Then, the expected *ERT* for item
*i* and examinee *j* follows: $\hat{\ln T}_{iy} = E(\beta + \beta_1\theta_j + \beta_2 r_j + \epsilon) = \hat{\beta}_i + \hat{\beta}_1\theta_j + \hat{\beta}_2 r_j$ (10)

If *c* is an examinee suspected of having irregularity responses, then the response time ofexaminee *c* to item *i* can be evaluated against his/her expected response time as a standard normal

$Z_{ic} = \frac{\ln T_{ic} - \ln \hat{T}_{iy}}{\sigma_i}$, (11)

where $\sigma_i^2 = \frac{\sum_j (\ln T_{ij} - \ln T_{iy})^2}{(J_i - 1)}$ (12)

with $J_i$ being the number of examinees who answer item *i* correctly. Therefore, $Z_{ic}^2$ is distributedas chi-square with one degree of freedom. The sum of the $Z^2$ across all items answered correctly

## 3. PURPOSE AND METHOD

The purpose of this study is to investigate whether the RTE method and/or the ERT method can screen examinees who had irregular response time so that "clean" data can be obtained. There are three different methods/procedures of data cleaning proposed in this study. The first uses the RTE method only, the second uses the ERT method only, and the third uses both the RTE and ERT methods. For the third method, after excluding examinees using the RTE method, the remaining data are examined using the ERT method to investigate whether further examinees need to be excluded. For the RTE method, three levels of RTE for inclusion are used: ≥90%, ≥ 80%, and

≥50%. For example, if it is ≥ 90%, then examinees that spent above the response time boundary (10 seconds) on 90% or more of the total items were included. On the ERT method, three significance levels (α) are used: 0.01, 0.05, and 0.10. In total, fifteen methods (3 RTE, 3 ERT, and 9 RTE-ERT) were compared and investigated to find an optimal method for excluding data irregularities and create a clean data.

The proposed methods are used for excluding data irregularities for a multiple-choice mathematics test. The original data and each "cleaned" dataset resulting from the proposed methods were calibrated using a 3-PL IRT calibration and compared. The

2021 International Conference on Innovation in Open & Distance Learning (2021 INNODEL)
ISSN 2963-2870
Vol 2 2022

number of score misfits, absolute standardized residual, the number of item-fits to the model, and chi-square goodness of fit were used to compare the methods.

The importance of this study is to estimate the relative effectiveness of each method to clean data irregularities. Also, this study is important because no previous study has compared and combined these two methods.

## 4. DATA

The study used data from a computerized mathematics test in an operational testing program for high school students. The test is a timed test, so that examinees had to finish the test within the allocated time. The test consists of 50 MC items with five options. Item responses along with their response times from 1,137 examinees were available for this study. The 3-PL model with fixed c parameter at 0.15 was implemented to compute the item parameters and the estimated abilities.

**Table 1.**
*Raw and Response Time Statistics*

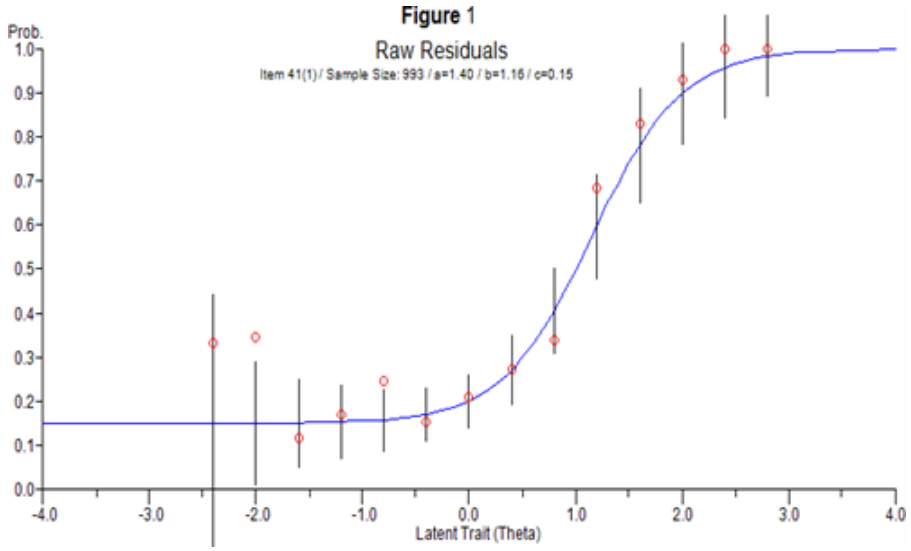
| | Raw Score | Response Time (in seconds) |
|---|---|---|
| Average | 25.00 | 2788 |
| SD | 9.89 | 300 |
| Min | 4 | 722 |
| Max | 50 | 3330 |
| SK | 0.33 | -1.98 |
| Kur | 2.42 | 8.72 |
| N | 1137 | 1137 |

Table 1 shows the summary statistics of the data. Given the average raw score is 25 out of 50, the test could be considered as a difficult test. On average, examinees spent 2,789 seconds for this 50-item MC test which is about 56 seconds per item. The minimum and maximum times needed by the examinees to finish the test are, respectively, about 12 and 56 minutes.

## 5. CRITERIA MEASURES

The original data that consists of 1,137 examinees were used as a benchmark data. The data resulted from the fifteen methods were calibrated into the 3-PL model and compared to eachother and to the results of benchmark in terms of IRT model fits to find the optimal method for excluding data irregularities. For each item on each method, the model fit was measured by the number of score misfits, absolute standardized residual, and chi-square goodness-of-fit. A plot ofestimated abilities resulted from each method is presented for an overall comparison.

2021 International Conference on Innovation in Open & Distance Learning (2021 INNODEL)
ISSN 2963-2870
Vol 2 2022




As an illustration of how criterion measures were conducted in this study, ResidPlot-2 (Liang, Han & Hambleton, 2009) was used to present Figure 1. The figure shows a plot of item residuals for measuring the IRT model fit on one of the test items. The red dots are the observed proportion of examinees within the intervals, and the blue line is the expected proportion along the theta-axis with its two standard errors represented by the error bars. Among the fourteen dots,two dots are beyond their error bars and it indicates that the model is misfit at those two score intervals for the item. The number of score misfits along with absolute standardized error and thechi-square criteria were used for measuring the model fit, and they are described more detail as follows.

## Number of Score Misfit

Each item was measured by how many misfit scores. The bigger the number the more misfits theitem has. Figure 1 indicates that the item has two misfit scores.

## Absolute Standardized Residual

Absolute standardized residual was calculated by standardizing the residual as

$$SR_j = \frac{|O_j - E_j|}{\sqrt{\frac{E(1-E_j)}{N_j}}}$$

where $O_j$ and $E_j$ are respectively the observed and expected proportions of correct answers for examinees in score interval $j$. $N_j$ is the number of examinees in that score interval. Graphically, Figure 1 shows that $O_j$ is the red dot and $E_j$ is the blue line. The overall standardized residual of each item was computed by averaging the values from all score intervals.

## Chi-square goodness-of-fit

For each item $i$, the goodness-of-fit was computed as follows:

$$N\ (O \quad -\ )^2$$

$k$

2021 International Conference on Innovation in Open & Distance Learning (2021 INNODEL)
ISSN 2963-2870
Vol 2 2022

$$X^2 = \sum_{j} \quad ij \quad ij$$

where k is the number of the score intervals. This mathematical equation follows a chi-square distribution with degree of freedom equals the number of the score intervals minus the number ofitem parameters being estimated. Thus, a chi-square distribution with the significant level α = 0.05 was used to test whether or not the observed data fit the model.

Estimated Abilities

The estimated item parameters resulted from each proposed method were used to estimate all examinees' ability based on his/her score responses. Given there were 1,137 examinees, then therewere 1,137 estimated abilities for each method. The differences between these estimated abilitiesand those of the original data were computed and compared.

## 6. RESULTS

**Table 2.**
*Raw amd RTs Statistics*

| Method | Type | Mean | SD | Min | Max | N (Exclusion) |
|---|---|---|---|---|---|---|
| Benchmark | Raw | 25.00 | 9.894 | 4 | 50 | |
| | RT | 2789 | 301 | 722 | 3330 | 1137 (0%) |
| ERT01 | Raw | 26.00 | 9.675 | 4 | 50 | |
| | RT | 2810 | 265 | 1243 | 3330 | 993 (13%) |
| ERT05 | Raw | 26.72 | 9.624 | 4 | 50 | |
| | RT | 2814 | 258 | 1243 | 3330 | 944 (17%) |
| ERT10 | Raw | 27.09 | 9.538 | 4 | 50 | |
| | RT | 2812 | 257 | 1243 | 3330 | 904 (20%) |
| RTE90 | Raw | 28.36 | 9.551 | 5 | 50 | |
| | RT | 2807 | 237 | 1473 | 3179 | 731 (36%) |
| RTE80 | Raw | 27.04 | 9.636 | 5 | 50 | |
| | RT | 2800 | 260 | 1473 | 3206 | 912 (20%) |
| RTE50 | Raw | 25.29 | 9.811 | 5 | 50 | |
| | RT | 2794 | 286 | 1243 | 3330 | 1111 (2%) |
| RTE90_ERT05 | Raw | 29.30 | 9.219 | 6 | 50 | |
| | RT | 2808 | 235 | 1473 | 3179 | 633 (44%) |
| RTE90_ERT10 | Raw | 29.53 | 9.228 | 6 | 50 | |
| | RT | 2811 | 229 | 1726 | 3179 | 607 (47%) |
| RTE80_ERT05 | Raw | 28.38 | 9.287 | 6 | 50 | |
| | RT | 2812 | 239 | 1573 | 3179 | 767 (33%) |

For the following sections, the proposed methods are coded as ERT and/or RTE with one or two index(s). For examples, ERT01 is abbreviations for Effort Response Time method with 1% of the significant level, RTE90 is for Response Time Effort with 90% of the inclusion level, and RTE90_ERT01 indicates that the two methods were used together with RTE as the first method implemented followed by ERT with their indices.

Table 2 shows the descriptive statistics of raw scores and RTs of the observed data after excluding data irregularities based on the considered method. The percentage exclusion in the last column is the percentage of the examinees that were excluded

2021 International Conference on Innovation in Open & Distance Learning (2021 INNODEL)
ISSN 2963-2870
Vol 2 2022

from the original data for the given method. Due to limited space, not all results of RTE_ERT methods are presented in this paper, except for the results from three interesting methods: RTE90_ERT05, RTE90_ERT10, and RTE80_ERT05.

Compared to the benchmark, ERT01 increases the averages of raw scores and RTs by about one raw score and 21 seconds, respectively. The increases also occurred for the other two ERT methods. Within ERT methods, as the significant level decreases from 1% to 10%, the exclusion number increases from 13% to 20%, and the average of the raw scores also increases from 26.00 to 27.09. In the case of RTs, it increases from ERT01 to ERT05, but it decreases from ERT05 to ERT10.

Unlike on the ERT methods, the index on the RTE methods represents the inclusion level as the larger the number the more restriction the method is. Therefore, as the inclusion level decreases from 90% to 50%, the percentage exclusion decreases from 36% to 2%, and the averages of the raw scores and RTs decreases from 28.36 to 25.29 and 2807 to 2794, respectively.

Furthermore, the number of examinees could be further excluded if both the RTE and ERT methods are used together. Given the largest reductions of examinees for RTE and ERT methods are respectively on RTE90 and ERT10, then it is expected that RTE90_ERT10 should produce the largest examinee exclusion. The table shows that this method can reduce the number of examinees to about half of the benchmark/original data, from 1,137 to 607 with the raw score average increasing to about 4.5 (29.53 – 25.00) score points and the RTs increasing to 22 seconds.

Those results show that ERT and/or RTE can be used to exclude examinees that were indicated to have response irregularities such as the averages of raw scores and/or response times of the observed data could be increased.

**Table 3.**

*Number of Item Misfit*

| Score Misfit | Bench. | ERT01 | ERT05 | ERT10 | RTE90 | RTE80 | RTE50 | RTE90_ ERT05 | RTE90_ ERT10 | RTE80_ ERT05 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 19 | 24 | 28 | 27 | 29 | 25 | 21 | 28 | 28 | 29 |
| 1 | 20 | 13 | 16 | 16 | 13 | 13 | 18 | 15 | 15 | 13 |
| 2 | 7 | 8 | 1 | 2 | 4 | 7 | 8 | 5 | 7 | 6 |
| 3 | 2 | 2 | 3 | 4 | 4 | 3 | 1 | 1 | 0 | 0 |
| 4 | 1 | 2 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 2 |
| 5 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 |
| N | 1137 | 993 | 944 | 904 | 731 | 912 | 1111 | 633 | 607 | 767 |

Table 3 shows the total number of items based on the score misfit criterion. The first column of the table is the number of score misfits from the 14 score intervals and the other columns are the number of items. For example, on the benchmark from the 50 items there are 19 items without any score misfit, 20 items with one score misfit, seven items with two score-misfits, and so on.

The table indicates that all proposed methods have more items that are fit to the 3-PL model than the benchmark. For an example, using RTE50 that is the least restricted model in which only 26 (1,137 – 1,111) examinees were excluded, the method increases the number of the item fits from 19 to 21. Within the method types,

2021 International Conference on Innovation in Open & Distance Learning (2021 INNODEL)
ISSN 2963-2870
Vol 2 2022

ERT05, RTE90, and RTE90_ERT10 are respectively the best for ERT, RTE, and ERT_RTE methods.

**Table 4.**

*Descriptive Statistics for Standardized Residual*

| Method | Bench. | ERT01 | ERT05 | ERT10 | RTE90 | RTE80 | RTE50 | RTE90_ ERT05 | RTE90_ ERT10 | RTE80_ ERT05 |
|---|---|---|---|---|---|---|---|---|---|---|
| Average | 0.819 | 0.808 | 0.763 | 0.773 | 0.752 | 0.794 | 0.814 | 0.748 | 0.725 | 0.742 |
| SD | 0.237 | 0.228 | 0.216 | 0.187 | 0.196 | 0.247 | 0.246 | 0.193 | 0.170 | 0.213 |
| Min | 0.410 | 0.447 | 0.391 | 0.440 | 0.456 | 0.422 | 0.377 | 0.438 | 0.459 | 0.401 |
| Max | 1.485 | 1.353 | 1.372 | 1.269 | 1.438 | 1.482 | 1.489 | 1.223 | 1.156 | 1.388 |

Table 4 is the descriptive statistics for the standardized residual. The averages of standardized residuals of all methods are smaller than that of the benchmark. Within the method types, the results are consistent with those of Table 3 which indicate that ERT05, RTE90, and RTE90_ERT10 are respectively the best for ERT, RTE, and ERT_RTE methods.

**Table 5.**

*Misfit Items Based on Chi-square Goodness of Fit*

| Method | Bench. | ERT01 | ERT05 | ERT10 | RTE90 | RTE80 | RTE50 | RTE90_ ERT05 | RTE90_ ERT10 | RTE80_ ERT05 |
|---|---|---|---|---|---|---|---|---|---|---|
| Item Number | 5, 8, 9, 10, 13, 16, 23, 27, 29, 37, 41, 44, 45, 46, 48 | 5, 8, 9, 10, 16, 23, 29, 30, 37, 41, 44, 48 | 5, 8, 9, 16, 23, 29, 37, 44 | 5, 8, 9, 16, 21, 23, 27, 29, 37, 44 | 5, 10, 15, 23, , 24, 27, 29, 37, 44, 48 | 1, 5, 8, 9, 10, 11, 16, 23, 24, 27, 37, 41, 44, 48 | 5, 8, 9, 10, 16, 23, 27, 28, 29, 37, 41, 42, 44, 46, 48 | 5, 10, 23, 29, 37, 42, 44 | 1, 5, 9, 10, 23, 28, 37, 44 | 5, 9, 10, 13, 23, 27, 29, 30, 37, 44 |
| Total | 15 | 12 | 7 | 10 | 10 | 14 | 15 | 7 | 8 | 10 |
| N | 1137 | 993 | 944 | 904 | 731 | 912 | 1111 | 633 | 607 | 767 |

Table 5 shows the number of misfit items based on the chi-square goodness of fit. Out of the 50 items administered, as shown from the benchmark method, 15 items did not fit into the 3- PL model. Interestingly, five of the 15 items came from the last ten items of the test, and those items are 41, 44, 45, 46, and 48. In other words, one-third of the misfit items came from one-fifth of the last items. This indicates that data irregularities might occur at the end of the test when a significant number of students answered the last items much faster than they should have.

Among the ERT methods, ERT05 is the best method by reducing more than half of the misfit items from 15 to 7. Particularly, only one from the last ten items of the test was identified as a misfit item. Compared to ERT methods, RTE method might not be as optimal in terms of reducing the number of the misfit items. The number of fit items was increased when both method types were implemented together, as can be seen from RTE_ERT model, but it could not be reduced to less than 7 items. Among the RTE_ERT methods, RTE90_ERT05 is favorable.

Finally, to investigate overall relative effectiveness of each proposed method for data cleaning, the difference between the estimated abilities resulting from each method

2021 International Conference on Innovation in Open & Distance Learning (2021 INNODEL)
ISSN 2963-2870
Vol 2 2022

and those from the benchmark was computed and compared. The difference was computed by subtracting the theta result of the benchmark data from the theta result of each method for a given examinee. Therefore, if it is a positive number, the method produced a larger estimation than the benchmark. Alternatively, if it is a negative number, the benchmark produced a larger estimation.

To enhance of the clarity of the results, only three methods are presented in this paper. The tree methods (ERT05, RTE90, and RTE90_ERT05) represent the favorable methods within their method types.

**Figure 2.**

*Differences of Estimated Ability*

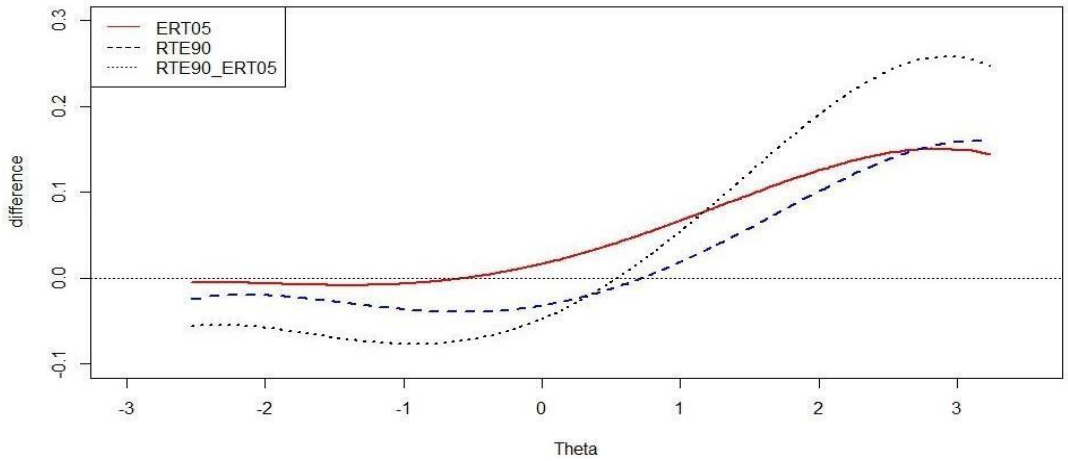



Figure 2 reveals that using different methods for data cleaning would produce different estimated abilities along the theta-axis. The values of RTE90_ERT05 are always lower or higher than those of the other two methods, except for the theta range of about 0.5 to 1.2, and the values of ERT05 are almost always higher than those of RTE90. However, given all proposed methods produced a cleaner data set than the original data by excluding examinees' responses due to their response irregularities, the three methods would produce better measures of ability than the benchmark. Particularly, using ERT05, the estimated ability would be greater than it should be for the theta range of greater than about -0.7, but it might not produce different results than the benchmark for the ranges of less than -0.7.

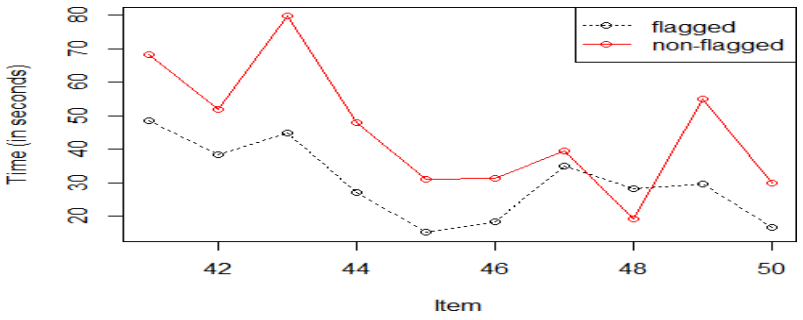
## 7. DISCUSSION AND CONCLUSIONS

This study shows that using ERT and/or RTE for excluding data irregularities would produce a "clean" data. ERT05, for example, detected 193 examinees that were indicative to have response irregularities. By excluding those examinees, the method can reduce the number of misfit from 15 items in a 50-item test to 7 (see Table 5). These data irregularities could be examined by revealing their RTs as shown in Figure 3.

2021 International Conference on Innovation in Open & Distance Learning (2021 INNODEL)
ISSN 2963-2870
Vol 2 2022

**Figure 3.**

*Average Response Times*




The figure is the averages of observed RTs for a “clean” data (non-flagged examinees) and those examinees flagged on the last ten items of the test. On average, the RTs of the non-flagged examinees are longer by about 10 seconds or more than those of the flagged examinees except forthe items 47 and 48. Giventhe average RT per item is 56 seconds (see Table 1), these results mightindicate that test speediness impacted more on the flagged examinees than the non-flagged examinees. Therefore, by excluding the flagged examinees, *ERT05* can reduce the number of misfit items.

In addition to reducing the number of misfit items, the use of the methods that incorporate RTs and responses together could produce better measures of ability as shown in Figure 2. Prior studies (e.g., Marianti, Fox, Avetisyan & Veldkamp, 2014; Meijer & Sotaridona, 2006; van der Linden, 2006; Wise & Kong, 2005) found similar finding. Particularly, this paper supports the finding of Widiatmo and Wright (2015) that uncovered that the use of responses and RTs into a model might be beneficial more for able students than less able students. As shown in the figure, for each model the absolute differences are higher for higher abilities than lower abilities.

Comparing ERT and RTE methods in terms of criteria measured, ERT worked better. It can be seen in the comparison of ERT10 to RTE80 in which those two methods employed approximately the same number of examinees. Table 3 shows that the number of items having zero score misfit for ERT10 is greater than that of the counterpart (27 vs 25). Tables 4 shows the standardized residual of ERT10 is smaller than that of RTE80 (0.773 vs 0.794), and finally from Table 5 ERT10 produced smaller number of misfit items compared to the other method (10 vs 14). The use of the ERT and RTE methods together might work better than only using of ERT. However, given the improvement is small, using both methods together might not be worthwhile.

For example, Table 5 shows that using the two methods might not reduce the number of misfit items to less than 7 items that can be achieved using ERT alone.

Although the present results are generally encouraging for implementing an ERT method for excluding data irregularities, a number of considerations should be addressed. First, this study used a timed test that might be considered as a difficult test, so that many examinees might not have enough time to finish the test. Therefore, the results might be different if the test is not a timed test.

2021 International Conference on Innovation in Open & Distance Learning (2021 INNODEL)
ISSN 2963-2870
Vol 2 2022

Second, the item parameter and ability estimations were based on the 3-PL model with fixed c parameter at 0.15. Different results might be obtained if a different IRT model was implemented. Third, the original data were based on a relatively small sample size (1,137 examinees). By increasing the sample size, the benefits of data cleaning might be enhanced. Fourth, the results of RTE methods were less encouraging than those of ERT methods. However, the results of RTE methods might be improved if the threshold used to define the response time boundary was defined based on the features of the test items as suggested by Wise and Kong (2005).

## 8. REFERENCES

Anders, R., Alario, F. X., and Van Maanen, L. (2016). The shifted Wald distribution for response time data analysis. Psychol. Methods 21, 309–327. doi: 10.1037/met0000066

Bolsinova, M., De Boeck, P., and Tijmstra, J. (2017). Modelling conditional dependencebetween response time and accuracy. Psychometrika 82, 1126–1148. doi: 10.1007/s11336-016-9537-6

De Boeck, P., & Minjeong J. (2019). An Overview of Models for Response Times and Processes in Cognitive Tests. Journal Frontiers in Psychology. (www.frontiersin.org/article/10.3389/fpsyg.2019.00102).

Bridgeman, B., Laitusis, C. C., & Cline, F. (2007). Time requirements for the different itemtypes proposed for use in the revised SAT®. Research Report No. 2007-3.

Ratcliff, R. (1993). Methods for dealing with reaction time outliers. Psychological Bulletin, 114,510-532.

Liang, T., Han, K.T., & Hambleton, R. K. (2009). Computer Software for IRT Graphical Residual Analysis, Version 2.0. Center for Educational Assessment Research Report No.688. Amherst, MA: University of Massachusetts, Center for Educational Assessment.

Marianti, S., Fox, J. P., Avetisyan, M., & Veldkemp, B. P. (2014). Testing for aberrant behaviorin response time modeling. Journal of Educational and Behavioral Statistics. 19, 426- 451.

Meijer, R. R., & Sotaridona, L. S. (2006). Detection of advance item knowledge using responsetimes in computer adaptive testing. Computerized Testing Report 03-03. Princeton, NJ: Law School Admission Council.

van der Linden, W. J. (2006). A lognormal model for response times on test items. Journal ofEducational and Behavioral Statistics, 31, 181-204.

Van der Linden, W.J., & van Krimpen-Stoop, E.M.L.A. (2003). Using response times to detect aberrant response patterns in computerized adaptive testing, Psychometrika, 68, 351-265.

2021 International Conference on Innovation in Open & Distance Learning (2021 INNODEL)
ISSN 2963-2870
Vol 2 2022

Van der Maas, H. L., Molenaar, D., Maris, G., Kievit, R. A., & Borsboom, D. (2011). Cognitivepsychology meets psychometric theory: On the relation between process models for decision making and latent variable models for individual differences. Psychological Review, 118, 339-356.

Wise, S. T., & Kong, X. (2005). Response time effort: A new measure of examinee motivation incomputer-based tests. Applied Measurement in Education, 18, 163-183.

Widiatmo, H., & Wright, D. (2015). Comparing Two Item Response Models That Incorporate Response Times. Paper presented at the meeting of the National Council on Measurementin Chicago, Illinois.