Item Analysis of the Rasch Model Items in the Final Semester Exam Indonesian Language Lesson

Nur Azizah[1], Muchlas Suseno[2], & Bahrul Hayat[3]

[1] Educational Research and Evaluation Program, State University of Jakarta (UNJ), Jakarta, Indonesia

[2] Department of English Language and Literature, State University of Jakarta (UNJ), Jakarta, Indonesia

[3] Faculty of Psychology Syarif Hidayatullah State Islamic University Jakarta, Indonesia

Correspondence: Nur Azizah, Educational Research and Evaluation Program, State University of Jakarta (UNJ), Jakarta, Indonesia.

Received: August 2, 2021 Accepted: September 29, 2021 Online Published: October 8, 2021

doi:10.5430/wjel.v12n1p15 URL: https://doi.org/10.5430/wjel.v12n1p15

Abstract

This study aims to analyze and describe the characteristics of the items for the Final Semester Examination for Indonesian Language courses at the PKN STAN official high school. The item analysis in this study was carried out with a modern approach (item response theory) using the Rasch model with Winstep software. This research is a quantitative research with descriptive method. The data was obtained through the documentation method, namely 25 multiple choice questions with 314 respondents. The results of the analysis show that the unidimensional requirements are not met. There are several items in the test that do not meet the local independence requirements, while the monotonicity requirements have been met. The ability of the subject on this test is greater than the level of difficulty of the questions indicated. The reliability of the test is generally satisfactory with the quality of the items and the consistency of the subject's answers are quite good. Persons or subjects who do not fit are 28 people, while items that are not fit are 5 items. There are 11 items indicated bias.

Keywords: item analysis, language test, Rasch model

1. Introduction

Evaluation of learning outcomes is carried out to monitor the process, progress, achievement, and improvement of learning outcomes on an ongoing basis (Erturk & Ziblim, 2020; Sukardi, 2008). The quality of learning can be seen from the results of the evaluation carried out (Caliskan & Zhu, 2020; Mardapi, 2017). A good evaluation can be done by collecting accurate evidence related to the achievement of student learning outcomes that are useful for increasing motivation and learning achievement (Demir, 2020; Stiggins & Chappuis, 2012).

Indonesian is a field of study that is taught from the elementary level (SD) to universities in Indonesia. At the university level, Indonesian courses are taught as general subjects. As a general course, Indonesian language courses are given to students at the initial level, for example in semester I to semester 4. The material covered in Indonesian language courses at universities, among others, are linguistic rules (spelling, diction, sentences, paragraph) and academic writing skills (Aksoy-Pekacar, Kanat-Mutluoğlu, & HAKKI-ERTEN, 2020; Hassan, 2020).

One form of evaluation of Indonesian language learning outcomes in universities is the final semester exam. The final semester exam value can be used to estimate the level of the learner's ability, namely through a person's response to a number of stimuli or questions (Kasalak & Dagyar, 2020; Mardapi, 2008). The examination score is a description of the mastery of language learner competencies in taking the learning process for one semester. The test results can be used as information about the characteristics of individual abilities or groups of learners (Abulela & Harwell, 2020; Rasyid & Mansur, 2008). Therefore, quality Indonesian Language final semester exam questions are needed as a guide for assessing the accuracy of learning outcomes (Barabadi, Robatjazi, & Bayat, 2020; Deveci, 2020).

Questions are said to be of good quality if they provide appropriate information about the mastery of learning materials (Erdil-Moody & Thompson, 2020; Suprananto, 2012). To get good quality item items, educators need to analyze the item items first before assessing learning outcomes. The purpose of item analysis activities is to examine each question before use, improve the quality of test items through revision, discard ineffective questions, and find out diagnostic information about the learner's understanding of the material that has been taught (Aiken, 1985; Mameche, OMRI, & Hassine, 2020). However, the facts on the ground show that item analysis is still rarely done by educators. The reasons are, among other things, educators feel burdened in the item analysis process so they don't do it and they believe that the quality of the test questions made is good so they don't conduct further studies.

Many efforts have been made to develop valid and reliable tests, starting from improving the technique of writing test items, administering tests, to item analysis with certain measurement models. One alternative measurement model that can produce valid and reliable items is the Rasch Model. This study aims to analyze the items for the final semester examination (UAS) for Indonesian Language courses using the Rasch Model. The aim is to find out whether the items on the test have met the validity and reliability requirements based on the Rasch Model.

2. Literature Review

Analysis of Indonesian test items can be done classically and modernly. Hambleton and Swaminathan (1985) revealed several weaknesses of classical test theory, including: (1) the level of difficulty and differentiating power of the items depended on the group of participants who worked on them, (2) the use of methods and techniques for test analysis was to compare students' abilities on the division of the test. upper, middle, and lower groups, (3) the concept of score reliability is defined from the term parallel test, (4) there is no theoretical basis to determine how test takers obtain tests that are in accordance with the abilities of these participants, and (5) Standard Error of Measurement (SEM) which applies to all test takers.

Measurement experts are trying to find alternatives as an effort to overcome the weaknesses that exist in the classical theory. A valid and reliable measurement model has the following characteristics: (1) item characteristics do not depend on the group of test takers; (2) participants' ability scores depend on the test; (3) stated in item level; (4) does not require parallel tests to calculate reliability coefficients, and (5) provides an appropriate measure for each ability score (Hambleton, et al, 1991). An alternative model that can have these characteristics is the Item Response Theory (IRT) measurement model. Modern test theory or item response theory was developed by measurement experts in psychology and education as an effort to minimize the shortcomings that exist in classical test theory. This is in line with Meyer and Zhu (2013) which states that IRT statistics are a way of estimating parameters in a model.

There are 4 models in IRT based on the number of parameters, namely one parameter logistic model (1PL), two parameter logistic model (2PL), three parameter logistic model (3PL), four parameter logistic model (4PL). This research only focuses on the 1PL model or the Rasch model. Rasch was the first to develop a one-parameter logistic model (Anunti, Vuopala, & Rusanen, 2020; Wright & Stone, 1979). The simple and accurate idea of observing elementary school student exam results in the 1950s led Georg Rasch to a new finding, namely the opportunity for students to answer one item correctly is the same as when the student's ability is compared to the level of difficulty of the questions (Bulut, 2020; Sumintono & Widhiarso, 2015).

In the Rasch model, people are given the characteristics of the level of latent ability and the items are given the characteristics of the level of difficulty. The probability of answering correctly a function is a function of the ratio between the level of ability and the difficulty of the item. An important feature of the Rasch model is that it does not contain discrimination and guess parameters. In this model, it is assumed that item difficulty is the only item characteristic that affects test performance. In addition, in the Rasch model, the problem of selecting items to construct a test is an effort to plan a quality test according to the needs and objectives of the test (Luber, FÖGELE, & Mehren, 2020; Sumintono & Widhiarso, 2015).

There are three basic assumptions that underlie item response theory, namely (1) unidimensional; (2) local independence; and (3) monotonicity. Hambleton, Swaminathan, and Rogers (1991) states that the unidimensional assumption means that only one ability is measured by a set of items in a test. The assumption of local independence is that there is no correlation between test takers' responses to different items. Monotonicity means the item characteristic function expresses the true relationship between item ability and response. The analysis with the Rasch model produces a statistical analysis of fit (fit statistics) which provides information that the data obtained describes people who have high ability to respond to items according to their level of difficulty. The parameters used are infit and outfit from the mean square and standardized values. Items that match (fit) means that the question behaves consistently with what is expected by the model. Some of the fit indices provided in the Rasch analysis are Person Infit ZSTD, Person Outfit ZSTD, Person Infit MNSQ, Person Outfit MNSQ, Item Infit ZSTD, Item Outfit ZSTD, Item Infit MNSQ, and Item Outfit MNSQ (Alagumalai, Curtis, & Hungi, 2005; Boone, Staver, & Yale, 2013).

The MNSQ value is always positive and moves from zero (0) to infinity (∞). In this case the MNSQ value is used to monitor the suitability of the data with the model. The expected mean square value is 1 (one). The mean-square value for infit or outfit that is greater than one indicates that the observed data has 30% more variation than predicted by Rasch. An infit or outfit value of less than 1 indicates that the observed data has 22% less variation than predicted by the Rasch model (Bond & Fox, 2013; Sopandi & Sukardi, 2020).

The expected z value is close to 0 (zero). When the observed data fit the model, the z-value has a mean close to 0 and the standard The deviation is 1. A ZSTD value that is too large (z > +2) or too low (z < -2) indicates that the item is not compatible with the expected model. The standardized z-value (ZSTD) on infit and outfit can be either positive or negative. A negative ZSTD value indicates less variation compared to the model. The answer response is close to the Guttman-style response string model, ie all subjects with high abilities are able to answer correctly and all subjects with low abilities answer incorrectly on the item. A positive ZSTD value indicates that the answers vary. (Bond & Fox, 2013). According to Boone et al. (2013) the value of outfit means-square (MNSQ), outfit z-standard (ZSTD), and point measure correlation (PTMEA) are used to see fit or not fit butit. The criteria are as follows.

1. Outfit Mean Square (MNSQ) value received : 0.5 < MNSQ < 1.5

2. Accepted Z-standard (ZSTD) outfit value: -2.0 < ZSTD < +2.0

3. Accepted Point Measure Correlation value: 0.4 < pt measure corr < 0.85

Rasch discriminatory power or the correlation value of item scores and Rasch scores (Pt Measure Corr) is in principle the same as the discriminatory power of items as measured by the classical test theory approach. It's just that if the classical computational test theory uses raw scores, in Pt Measure Corr the measure scores are used. The Pt Measure Corr value of 1.0 indicates that all test takers with low ability answered the item incorrectly and all test takers with high ability answered the item correctly. A negative Pt Measure Corr value indicates a misleading item because test takers with low abilities are able to answer items correctly and test takers with high abilities actually answer incorrectly. Questions with a negative correlation score should be checked to see if the answer key is incorrect, needs to be revised, or removed from the test. The correlation value between item scores and the ideal Rasch score is positive and not close to zero.

The item difficulty level in the Rasch model is basically the same as the classical test theory difficulty level, namely the ratio between the number of correct answers and the number of questions tested (odd-ratio). The difference is that the probability value is then scaled by entering a logarithmic function. The logit estimation result from this odd-ratio is called the logit or W-score or measure value. If in the classical test theory a high difficulty index value means the question is easy, in the Rasch model a high logit value indicates the item is difficult. A high logit (measure) value indicates that the item has a high level of difficulty. Measure values are arranged to range from -3 to +3. However, logit values above 2 or below -2 can be considered as extreme values. Sumintono and Widhiarso (2015) provide guidance in assessing these items into four categories, namely

1. Measure value < -1 = very easy item

2. Measure value -1 to d. 0 = easy item

3. Measure value 0 to d. 1 = difficult item

4. Measure value > 1 = very difficult item

Item bias is a test condition that is unfair, inconsistent, and polluted by factors other than the ability factor to be tested. Items that are biased cause a test to be discriminatory or in favor of certain groups whose causes can be viewed from various aspects that have absolutely nothing to do with ability factors, such as gender, ethnicity, culture, region, and others (Osterlind, 1983; Torun, 2020). An item is called biased if it is found that individuals with certain characteristics are more advantageous in answering questions than individuals with other characteristics. For example, a question can be answered more easily by a person living in a city than a person living in a village. In the Rasch model, item bias can be detected with DIF (differential item functioning). The items identified by DIF (p<0.05) are recommended to be reviewed and if deemed necessary to be revised or replaced.

Measurement information is based on the relationship between the test and the individual. Sumintono and Widhiarso (2015) explain some of the benefits of the test information function as follows. (1) The information function will show what the measurement is for. For example, for screening tests, remedial tests, and tests for Children with Special Needs, tests are centralized with information functions such as red graphs. On the other hand, if the test is carried out for strict selection, the required test information function is a test with an information function as shown in the black graph. (2) The information function shows the reliability of the measurements made. The Rasch model emphasizes the separation coefficient (item separation). The higher the peak of information that can be achieved, the higher the value of the reliability of the measurements made

3. Method

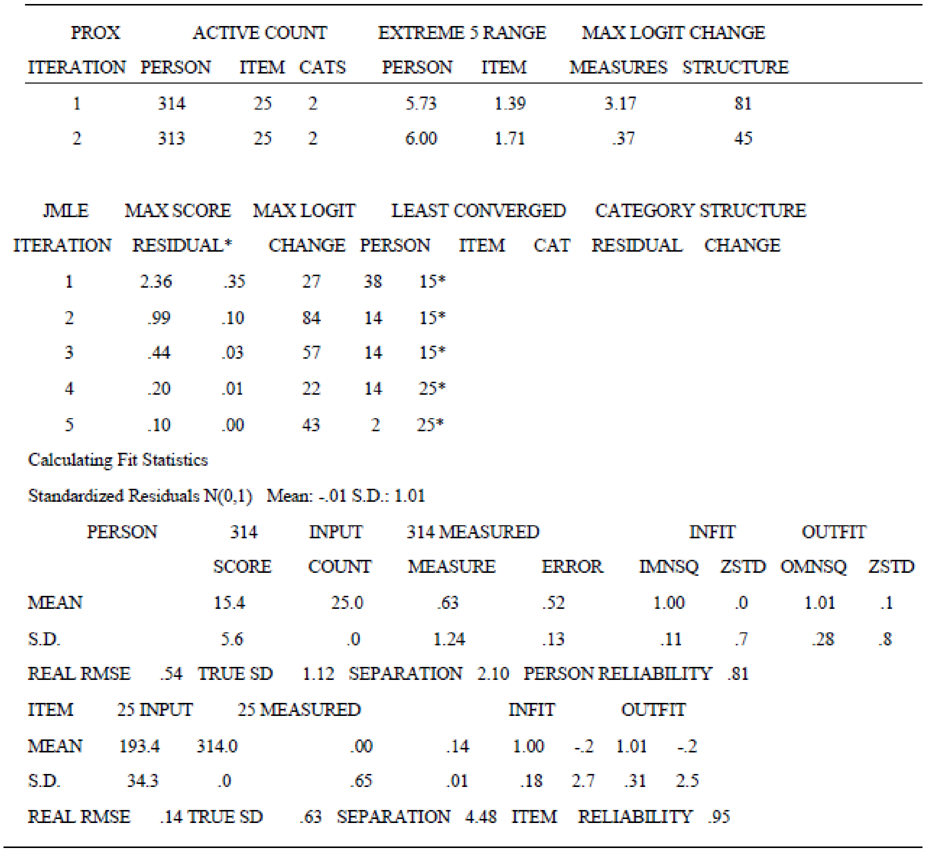
The items analyzed were the Indonesian Language Final Semester Examination questions at the PKN STAN official high school in South Tangerang. The number of questions analyzed is 25 multiple choice questions. Students who work on these items are first-year students (semester I) totaling 314 people. The test result data is then entered into the excel program and tidied up. The data that has been in excel format was then analyzed using the Winstep program (Linacre, 2011). The psychometric elements analyzed in the items of the Indonesian Final Semester Exam, assumption test, item and person fitness, bias, and test information function.

4. Results

Data Convergence

Before going into further analysis, information is needed whether the analyzed data is convergent or not. The convergence of the analyzed data can be seen in the Convergence Table.

Tabel 1. Convergence Table



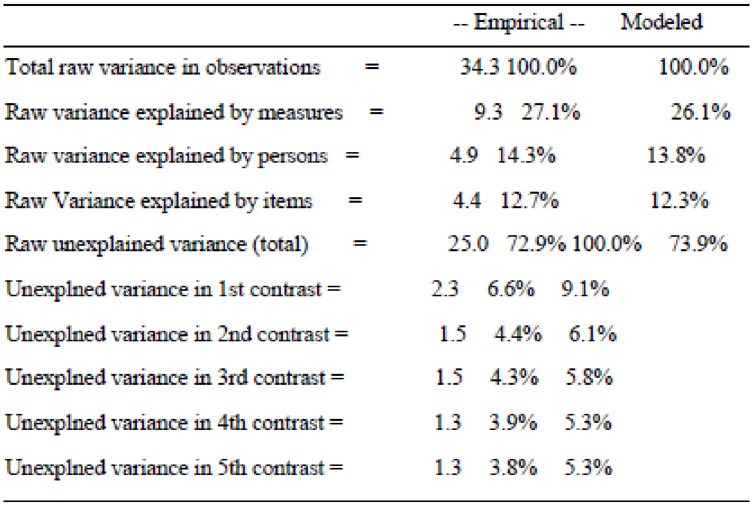
```
    PROX       ACTIVE COUNT      EXTREME 5 RANGE     MAX LOGIT CHANGE
ITERATION  PERSON    ITEM  CATS    PERSON    ITEM     MEASURES  STRUCTURE
    1        314      25    2       5.73     1.39       3.17        81
    2        313      25    2       6.00     1.71        .37        45

   JMLE     MAX SCORE   MAX LOGIT    LEAST CONVERGED    CATEGORY STRUCTURE
ITERATION   RESIDUAL*    CHANGE  PERSON    ITEM    CAT   RESIDUAL   CHANGE
    1         2.36      .35      27     38     15*
    2          .99      .10      84     14     15*
    3          .44      .03      57     14     15*
    4          .20      .01      22     14     25*
    5          .10      .00      43      2     25*
Calculating Fit Statistics
Standardized Residuals N(0,1)   Mean: -.01 S.D.: 1.01
      PERSON          314      INPUT     314 MEASURED              INFIT        OUTFIT
                      SCORE    COUNT     MEASURE       ERROR    IMNSQ  ZSTD  OMNSQ  ZSTD
MEAN                  15.4      25.0       .63          .52      1.00    .0   1.01    .1
S.D.                   5.6        .0      1.24          .13       .11    .7    .28    .8
REAL RMSE    .54  TRUE SD   1.12  SEPARATION  2.10  PERSON RELIABILITY  .81
ITEM      25 INPUT     25 MEASURED               INFIT       OUTFIT
MEAN      193.4     314.0            .00       .14     1.00   -.2   1.01   -.2
S.D.        34.3       .0            .65       .01      .18   2.7    .31   2.5
REAL RMSE    .14 TRUE SD    .63  SEPARATION  4.48  ITEM   RELIABILITY  .95
```

The convergence table 1 shows that the mean value is close to 0, which is -0.01 and the standard deviation value is close to 1, which is 1.01. This means that the data has converged. The table above also shows a summary of the test items and subjects. The test has a separation of 4.48 and item reliability of 0.95. This means that the test has functioned quite well because it has high reliability and a variety of difficulty levels. Subject has 2.10 separation and 0.82 person reliability. This means that the subject is quite varied because it has a wide or diverse range of abilities

Assumption Test (Unidimensional, Local Independence, Monotonicity)

Before proceeding with the next analysis, assumptions were tested, namely unidimensional, local independence, and monotonicity tests.

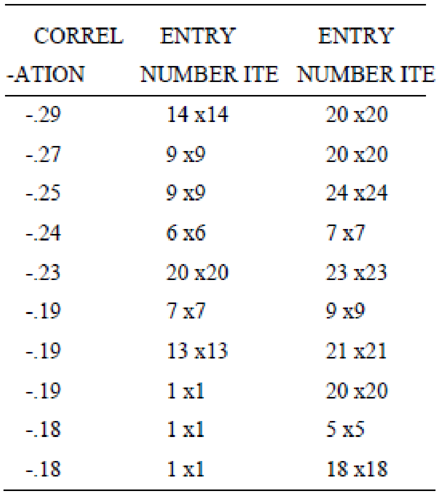
Table 2. Unidimensi



	-- Empirical --			Modeled
Total raw variance in observations =	34.3	100.0%		100.0%
Raw variance explained by measures =	9.3	27.1%		26.1%
Raw variance explained by persons =	4.9	14.3%		13.8%
Raw Variance explained by items =	4.4	12.7%		12.3%
Raw unexplained variance (total) =	25.0	72.9%	100.0%	73.9%
Unexplned variance in 1st contrast =	2.3	6.6%	9.1%	
Unexplned variance in 2nd contrast =	1.5	4.4%	6.1%	
Unexplned variance in 3rd contrast =	1.5	4.3%	5.8%	
Unexplned variance in 4th contrast =	1.3	3.9%	5.3%	
Unexplned variance in 5th contrast =	1.3	3.8%	5.3%	

To find out whether the test meets the unidimensional requirements or not, it can be seen in the table above, namely the Raw variance explained by measures section. If the value of Raw variance explained by measures is less than 30%, unidimensionality is violated or not met. Table 23.0 above shows that the Raw variance explained by measures is 27.1% (empirical) and 26.1% (modeled), less than 30%. It means that the unidimensional assumption is violated or not fulfilled.

Table 3. Local Independence



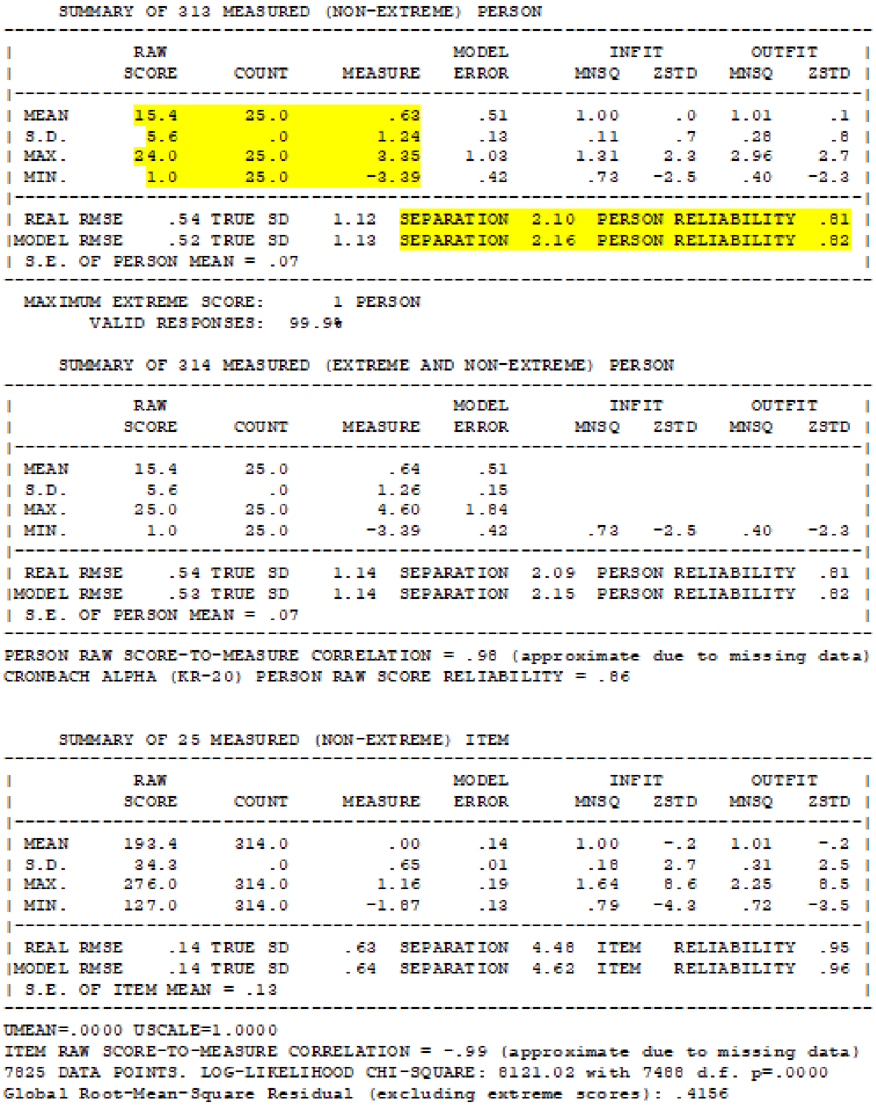
CORREL -ATION	ENTRY NUMBER ITE	ENTRY NUMBER ITE
-.29	14 x14	20 x20
-.27	9 x9	20 x20
-.25	9 x9	24 x24
-.24	6 x6	7 x7
-.23	20 x20	23 x23
-.19	7 x7	9 x9
-.19	13 x13	21 x21
-.19	1 x1	20 x20
-.18	1 x1	5 x5
-.18	1 x1	18 x18

To find out whether the test meets the local independence requirements or not, it can be seen in the table above, namely the correlation value. The assumption of local independence is met if the correlation value is less than 20% or 0.20. In the table above, it can be seen that items 9, 7, 14, 23, and 24 do not meet the local independence requirements because they are correlated (the value is above 20%) with other items.

Tabel 4. Monotonicity

ENTRY	DATA	SCORE	DATA		AVERAGE	S.E.	OUTF	PTMEA	
NUMBER	CODE	VALUE	COUNT	%	MEASURE	MEAN	MNSQ	CORR.	ITEM
8	0	0	187	60	.11	.08	1.0	-.51	x8
	1	1	127	40	1.42	.09	.8	.51	
1	0	0	175	56	-.01	.08	.8	-.58	x1
	1	1	139	44	1.47	.08	.8	.58	
11	0	0	172	55	.01	.08	.9	-.55	x11
	1	1	142	45	1.40	. 08	.8	.55	

In the table 4 above, it can be seen that the items in the test have met the monotonicity requirements as seen from the average measure values, which are ordered from low to high.



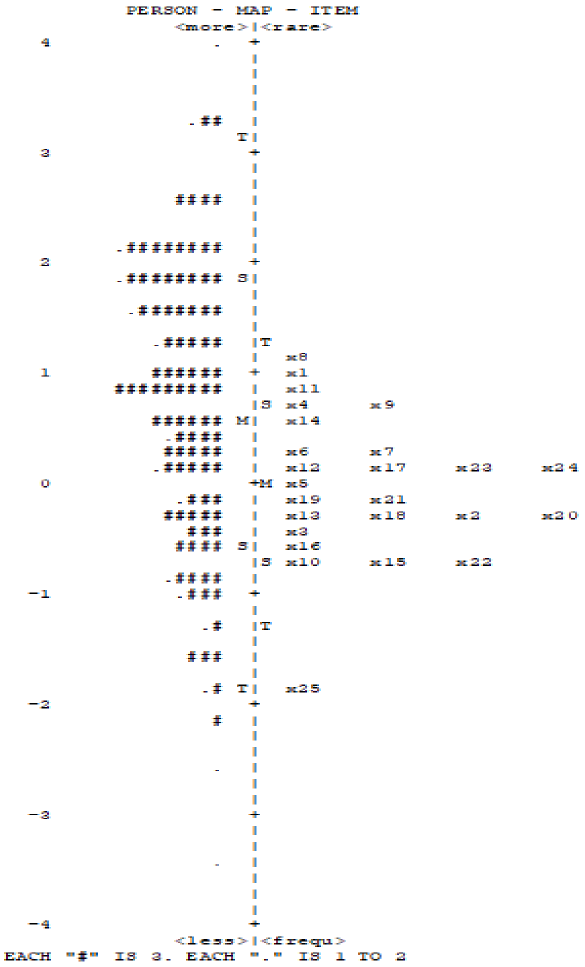
```
     SUMMARY OF 313 MEASURED (NON-EXTREME) PERSON
-------------------------------------------------------------------------------
|          RAW                          MODEL         INFIT        OUTFIT     |
|         SCORE      COUNT     MEASURE  ERROR      MNSQ   ZSTD   MNSQ   ZSTD  |
|-----------------------------------------------------------------------------|
| MEAN      15.4      25.0         .63    .51      1.00     .0   1.01     .1  |
| S.D.       5.6        .0        1.24    .13       .11     .7    .28     .8  |
| MAX.      24.0      25.0        3.35   1.03      1.31    2.3   2.96    2.7  |
| MIN.       1.0      25.0       -3.39    .42       .73   -2.5    .40   -2.3  |
|-----------------------------------------------------------------------------|
| REAL RMSE    .54 TRUE SD    1.12  SEPARATION  2.10  PERSON RELIABILITY  .81 |
|MODEL RMSE    .52 TRUE SD    1.13  SEPARATION  2.16  PERSON RELIABILITY  .82 |
| S.E. OF PERSON MEAN = .07                                                   |
-------------------------------------------------------------------------------
  MAXIMUM EXTREME SCORE:       1 PERSON
        VALID RESPONSES:  99.9%

     SUMMARY OF 314 MEASURED (EXTREME AND NON-EXTREME) PERSON
-------------------------------------------------------------------------------
|          RAW                          MODEL         INFIT        OUTFIT     |
|         SCORE      COUNT     MEASURE  ERROR      MNSQ   ZSTD   MNSQ   ZSTD  |
|-----------------------------------------------------------------------------|
| MEAN      15.4      25.0         .64    .51                                 |
| S.D.       5.6        .0        1.26    .15                                 |
| MAX.      25.0      25.0        4.60   1.84                                 |
| MIN.       1.0      25.0       -3.39    .42       .73   -2.5    .40   -2.3  |
|-----------------------------------------------------------------------------|
| REAL RMSE    .54 TRUE SD    1.14  SEPARATION  2.09  PERSON RELIABILITY  .81 |
|MODEL RMSE    .53 TRUE SD    1.14  SEPARATION  2.15  PERSON RELIABILITY  .82 |
| S.E. OF PERSON MEAN = .07                                                   |
-------------------------------------------------------------------------------
PERSON RAW SCORE-TO-MEASURE CORRELATION = .98 (approximate due to missing data)
CRONBACH ALPHA (KR-20) PERSON RAW SCORE RELIABILITY = .86


     SUMMARY OF 25 MEASURED (NON-EXTREME) ITEM
-------------------------------------------------------------------------------
|          RAW                          MODEL         INFIT        OUTFIT     |
|         SCORE      COUNT     MEASURE  ERROR      MNSQ   ZSTD   MNSQ   ZSTD  |
|-----------------------------------------------------------------------------|
| MEAN     193.4     314.0         .00    .14      1.00    -.2   1.01    -.2  |
| S.D.      34.3        .0         .65    .01       .18    2.7    .31    2.5  |
| MAX.     276.0     314.0        1.16    .19      1.64    8.6   2.25    8.5  |
| MIN.     127.0     314.0       -1.87    .13       .79   -4.3    .72   -3.5  |
|-----------------------------------------------------------------------------|
| REAL RMSE    .14 TRUE SD     .63  SEPARATION  4.48  ITEM   RELIABILITY  .95 |
|MODEL RMSE    .14 TRUE SD     .64  SEPARATION  4.62  ITEM   RELIABILITY  .96 |
| S.E. OF ITEM MEAN = .13                                                     |
-------------------------------------------------------------------------------
UMEAN=.0000 USCALE=1.0000
ITEM RAW SCORE-TO-MEASURE CORRELATION = -.99 (approximate due to missing data)
7825 DATA POINTS. LOG-LIKELIHOOD CHI-SQUARE: 8121.02 with 7488 d.f. p=.0000
Global Root-Mean-Square Residual (excluding extreme scores): .4156
```

Figure 1. Summary Of Measured Items and Person

In Figure 1, the mean measure person is 0.63. The mean value greater than 0 indicates that the tendency of the subject's ability is greater than the level of difficulty of the questions. Cronbach's Alpha value (KR-20) is a reliability coefficient calculated based on the classical test theory approach. This value is the interaction between the person and the item as a whole. Alpha value is 0.86. This shows that the reliability of the test is generally satisfactory. The value of person reliability in the table is 0.81 and the value of item reliability in the table is 0.95. This shows that the consistency of the answers from the subjects is quite good and the quality of the items in the instrument's reliability aspect is quite good.

The ideal values for INFIT and OUTFIT MNSQ are close to 1, while the ideal values for INFIT and OUTFIT ZSTD are close to 0. For the person and item tables, the mean values for INFIT and OUTFIT MNSQ and INFIT and OUTFIT ZSTD are close to ideal. The separation value also shows the quality of the instrument and the quality of the subject. The greater the value of separation, the better because it can identify a wider group of subjects (able – unable) and a wider group of items (difficult – easy). The separation value for the items in the table is 4.48 and person is 2.09. This value is relatively high, indicating that the quality of the subject and instrument is quite good.

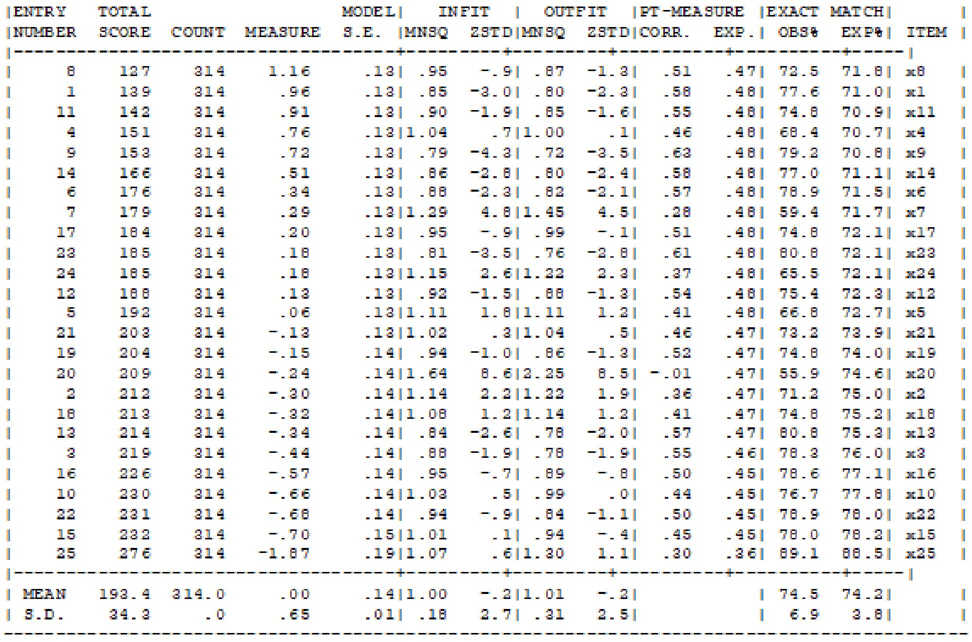
To find out the distribution of items and the ability of the subject to respond to items in general, the person-item map can be seen figure 2.



```
          PERSON - MAP - ITEM
               <more>|<rare>
   4              .  +
                     |
                     |
                     |
                     |
                .##  |
                    T|
   3                 +
                     |
                     |
               ####  |
                     |
                     |
          .########  |
   2                 +
          .######## S|
                     |
           .#######  |
                     |
             .#####  |T
                     |  x8
   1         ######  +  x1
          #########  |  x11
                     |S x4     x9
             ###### M|  x14
              .####  |
              #####  |  x6     x7
             .#####  |  x12    x17    x23    x24
   0                 +M x5
               .###  |  x19    x21
              #####  |  x13    x18    x2     x20
                ###  |  x3
               #### S|  x16
                     |S x10    x15    x22
              .####  |
  -1           .###  +
                     |
                 .#  |T
                     |
                ###  |
                     |
                 .# T|  x25
  -2                 +
                  #  |
                     |
                     |
                  .  |
                     |
                     |
  -3                 +
                     |
                     |
                  .  |
                     |
                     |
  -4                 +
               <less>|<frequ>
EACH "#" IS 3. EACH "." IS 1 TO 2
```

Figure 2. Person-Item Map

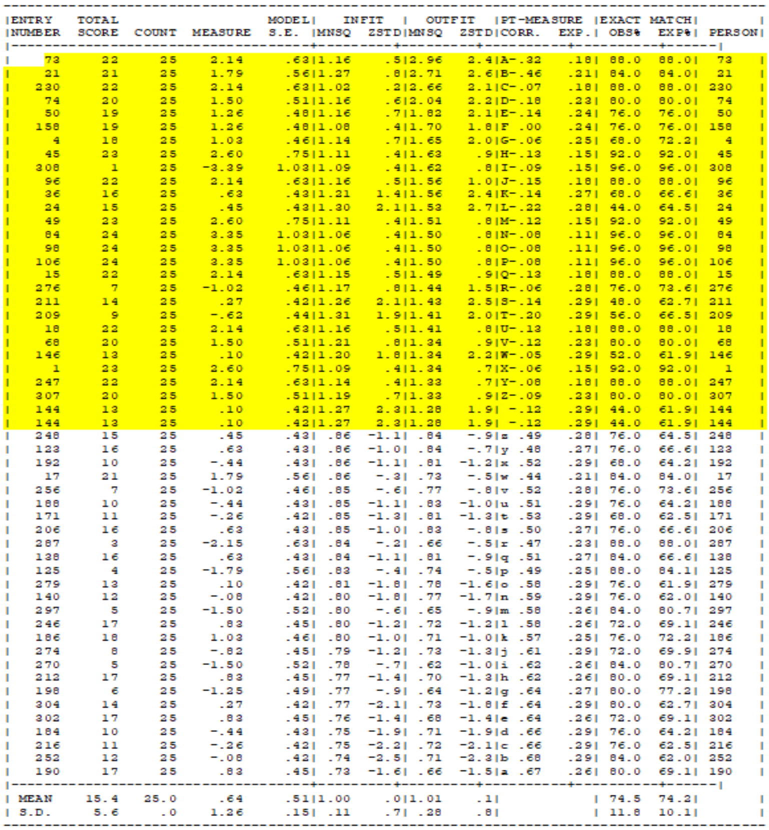
The left side is the distribution of the subject's abilities, while the right side is the distribution of items. From the map, it can be seen that the easiest item is item number 25 (x25) which is in the lowest position, while the most difficult item is item number 8 (x8). In general, the questions in the test are lower (easy) when compared to the ability of the subject. In other words, the questions are relatively too easy for test takers who generally have high abilities. To find out the level of difficulty in more detail, it can be seen from figure 3 (Measure Order) below.



```
|ENTRY   TOTAL                        MODEL|   INFIT  |  OUTFIT  |PT-MEASURE |EXACT MATCH|      |
|NUMBER  SCORE  COUNT  MEASURE  S.E. |MNSQ  ZSTD|MNSQ  ZSTD|CORR.  EXP.| OBS%  EXP%| ITEM |
|------------------------------------+----------+----------+-----------+-----------+-----|
|     8    127    314    1.16     .13| .95   -.9| .87  -1.3| .51    .47| 72.5  71.8| x8   |
|     1    139    314     .96     .13| .85  -3.0| .80  -2.3| .58    .48| 77.6  71.0| x1   |
|    11    142    314     .91     .13| .90  -1.9| .85  -1.6| .55    .48| 74.8  70.9| x11  |
|     4    151    314     .76     .13|1.04    .7|1.00    .1| .46    .48| 68.4  70.7| x4   |
|     9    153    314     .72     .13| .79  -4.3| .72  -3.5| .63    .48| 79.2  70.8| x9   |
|    14    166    314     .51     .13| .86  -2.8| .80  -2.4| .58    .48| 77.0  71.1| x14  |
|     6    176    314     .34     .13| .88  -2.3| .82  -2.1| .57    .48| 78.9  71.5| x6   |
|     7    179    314     .29     .13|1.29   4.8|1.45   4.5| .28    .48| 59.4  71.7| x7   |
|    17    184    314     .20     .13| .95   -.9| .99   -.1| .51    .48| 74.8  72.1| x17  |
|    23    185    314     .18     .13| .81  -3.5| .76  -2.8| .61    .48| 80.8  72.1| x23  |
|    24    185    314     .18     .13|1.15   2.6|1.22   2.3| .37    .48| 65.5  72.1| x24  |
|    12    188    314     .13     .13| .92  -1.5| .88  -1.3| .54    .48| 75.4  72.3| x12  |
|     5    192    314     .06     .13|1.11   1.8|1.11   1.2| .41    .48| 66.8  72.7| x5   |
|    21    203    314    -.13     .13|1.02    .3|1.04    .5| .46    .47| 73.2  73.9| x21  |
|    19    204    314    -.15     .14| .94  -1.0| .86  -1.3| .52    .47| 74.8  74.0| x19  |
|    20    209    314    -.24     .14|1.64   8.6|2.25   8.5|-.01    .47| 55.9  74.6| x20  |
|     2    212    314    -.30     .14|1.14   2.2|1.22   1.9| .36    .47| 71.2  75.0| x2   |
|    18    213    314    -.32     .14|1.08   1.2|1.14   1.2| .41    .47| 74.8  75.2| x18  |
|    13    214    314    -.34     .14| .84  -2.6| .78  -2.0| .57    .47| 80.8  75.3| x13  |
|     3    219    314    -.44     .14| .88  -1.9| .78  -1.9| .55    .46| 78.3  76.0| x3   |
|    16    226    314    -.57     .14| .95   -.7| .89   -.8| .50    .45| 78.6  77.1| x16  |
|    10    230    314    -.66     .14|1.03    .5| .99    .0| .44    .45| 76.7  77.8| x10  |
|    22    231    314    -.68     .14| .94   -.9| .84  -1.1| .50    .45| 78.9  78.0| x22  |
|    15    232    314    -.70     .15|1.01    .1| .94   -.4| .45    .45| 78.0  78.2| x15  |
|    25    276    314   -1.87     .19|1.07    .6|1.30   1.1| .30    .36| 89.1  88.5| x25  |
|------------------------------------+----------+----------+-----------+-----------+-----|
| MEAN  193.4  314.0     .00     .14|1.00   -.2|1.01   -.2|           | 74.5  74.2|      |
| S.D.   34.3     .0     .65     .01| .18   2.7| .31   2.5|           |  6.9   3.8|      |
----------------------------------------------------------------------------------------------
```

Figure 3. Measure Order

The most difficult item is item number 8 with a measure value of 1.96. Other items included in the difficult category are items numbered 1, 11, 14, 4, 9, 14, 6, 7, 17, 23, 24, 12, and 5. Item questions number 21, 19, 20, 2, 18 , 13, 3, 16, 10, 22, and 15 are easy questions. Items that are included in the very easy category are item number 25 with a measure of -1.87. To find out, out of 314 subjects, which subjects or persons are not fit, see figure 4.



```
-------------------------------------------------------------------------------------------
|ENTRY   TOTAL                        MODEL|   INFIT  |  OUTFIT  |PT-MEASURE |EXACT MATCH|        |
|NUMBER  SCORE  COUNT  MEASURE  S.E. |MNSQ  ZSTD|MNSQ  ZSTD|CORR.  EXP.| OBS%  EXP%| PERSON|
|------------------------------------+----------+----------+-----------+-----------+------|
|    73     22     25    2.14     .63|1.16    .5|2.96   2.4|A-.32   .18| 88.0  88.0|  73    |
|    21     21     25    1.79     .56|1.27    .8|2.71   2.6|B-.46   .21| 84.0  84.0|  21    |
|   230     22     25    2.14     .63|1.02    .2|2.66   2.1|C-.07   .18| 88.0  88.0| 230    |
|    74     20     25    1.50     .51|1.16    .6|2.04   2.2|D-.18   .23| 80.0  80.0|  74    |
|    50     19     25    1.26     .48|1.16    .7|1.82   2.1|E-.14   .24| 76.0  76.0|  50    |
|   158     19     25    1.26     .48|1.08    .4|1.70   1.8|F .00   .24| 76.0  76.0| 158    |
|     4     18     25    1.03     .46|1.14    .7|1.65   2.0|G-.06   .25| 68.0  72.2|   4    |
|    45     23     25    2.60     .75|1.11    .4|1.63    .9|H-.13   .15| 92.0  92.0|  45    |
|   308      1     25   -3.39    1.03|1.09    .4|1.62    .8|I-.09   .15| 96.0  96.0| 308    |
|    96     22     25    2.14     .63|1.16    .5|1.56   1.0|J-.15   .18| 88.0  88.0|  96    |
|    36     16     25     .63     .43|1.21   1.4|1.56   2.4|K-.14   .27| 68.0  66.6|  36    |
|    24     15     25     .45     .43|1.30   2.1|1.53   2.7|L-.22   .28| 44.0  64.5|  24    |
|    49     23     25    2.60     .75|1.11    .4|1.51    .8|M-.12   .15| 92.0  92.0|  49    |
|    84     24     25    3.35    1.03|1.06    .4|1.50    .8|N-.08   .11| 96.0  96.0|  84    |
|    98     24     25    3.35    1.03|1.06    .4|1.50    .8|O-.08   .11| 96.0  96.0|  98    |
|   106     24     25    3.35    1.03|1.06    .4|1.50    .8|P-.08   .11| 96.0  96.0| 106    |
|    15     22     25    2.14     .63|1.15    .5|1.49    .9|Q-.13   .18| 88.0  88.0|  15    |
|   276      7     25   -1.02     .46|1.17    .8|1.44   1.5|R-.06   .28| 76.0  73.6| 276    |
|   211     14     25     .27     .42|1.26   2.1|1.43   2.5|S-.14   .29| 48.0  62.7| 211    |
|   209      9     25    -.62     .44|1.31   1.9|1.41   2.0|T-.20   .29| 56.0  66.5| 209    |
|    18     22     25    2.14     .63|1.16    .5|1.41    .8|U-.13   .18| 88.0  88.0|  18    |
|    68     20     25    1.50     .51|1.21    .8|1.34    .9|V-.12   .23| 80.0  80.0|  68    |
|   146     13     25     .10     .42|1.20   1.8|1.34   2.2|W-.05   .29| 52.0  61.9| 146    |
|     1     23     25    2.60     .75|1.09    .4|1.34    .7|X-.06   .15| 92.0  92.0|   1    |
|   247     22     25    2.14     .63|1.14    .4|1.33    .7|Y-.08   .18| 88.0  88.0| 247    |
|   307     20     25    1.50     .51|1.19    .7|1.33    .9|Z-.09   .23| 80.0  80.0| 307    |
|   144     13     25     .10     .42|1.27   2.3|1.28   1.9| -.12   .29| 44.0  61.9| 144    |
|   144     13     25     .10     .42|1.27   2.3|1.28   1.9| -.12   .29| 44.0  61.9| 144    |
|   248     15     25     .45     .43| .86  -1.1| .84   -.9|z .49   .28| 76.0  64.5| 248    |
|   123     16     25     .63     .43| .86  -1.0| .84   -.7|y .48   .27| 76.0  66.6| 123    |
|   192     10     25    -.44     .43| .86  -1.1| .81  -1.2|x .52   .29| 68.0  64.2| 192    |
|    17     21     25    1.79     .56| .86   -.3| .73   -.5|w .44   .21| 84.0  84.0|  17    |
|   256      7     25   -1.02     .46| .85   -.6| .77   -.8|v .52   .28| 76.0  73.6| 256    |
|   188     10     25    -.44     .43| .85  -1.1| .83  -1.0|u .51   .29| 76.0  64.2| 188    |
|   171     11     25    -.26     .42| .85  -1.3| .81  -1.3|t .53   .29| 68.0  62.5| 171    |
|   206     16     25     .63     .43| .85  -1.0| .83   -.8|s .50   .27| 76.0  66.6| 206    |
|   287      3     25   -2.15     .63| .84   -.2| .66   -.5|r .47   .23| 88.0  88.0| 287    |
|   138     16     25     .63     .43| .84  -1.1| .81   -.9|q .51   .27| 84.0  66.6| 138    |
|   125      4     25   -1.79     .56| .83   -.4| .74   -.5|p .49   .25| 88.0  84.1| 125    |
|   279     13     25     .10     .42| .81  -1.8| .78  -1.6|o .58   .29| 76.0  61.9| 279    |
|   140     12     25    -.08     .42| .80  -1.8| .77  -1.7|n .59   .29| 76.0  62.0| 140    |
|   297      5     25   -1.50     .52| .80   -.6| .65   -.9|m .58   .26| 84.0  80.7| 297    |
|   246     17     25     .83     .45| .80  -1.2| .72  -1.2|l .58   .26| 72.0  69.1| 246    |
|   186     18     25    1.03     .46| .80  -1.0| .71  -1.0|k .57   .25| 76.0  72.2| 186    |
|   274      8     25    -.82     .45| .79  -1.2| .73  -1.3|j .61   .29| 72.0  69.9| 274    |
|   270      5     25   -1.50     .52| .78   -.7| .62  -1.0|i .62   .26| 84.0  80.7| 270    |
|   212     17     25     .83     .45| .77  -1.4| .70  -1.3|h .62   .26| 80.0  69.1| 212    |
|   198      6     25   -1.25     .49| .77   -.9| .64  -1.2|g .64   .27| 80.0  77.2| 198    |
|   304     14     25     .27     .42| .77  -2.1| .73  -1.8|f .64   .29| 80.0  62.7| 304    |
|   302     17     25     .83     .45| .76  -1.4| .68  -1.4|e .64   .26| 72.0  69.1| 302    |
|   184     10     25    -.44     .43| .75  -1.9| .71  -1.9|d .66   .29| 76.0  64.2| 184    |
|   216     11     25    -.26     .42| .75  -2.2| .72  -2.1|c .66   .29| 76.0  62.5| 216    |
|   252     12     25    -.08     .42| .74  -2.5| .71  -2.3|b .68   .29| 84.0  62.0| 252    |
|   190     17     25     .83     .45| .73  -1.6| .66  -1.5|a .67   .26| 80.0  69.1| 190    |
|------------------------------------+----------+----------+-----------+-----------+------|
| MEAN    15.4   25.0     .64     .51|1.00    .0|1.01    .1|           | 74.5  74.2|        |
| S.D.     5.6     .0    1.26     .15| .11    .7| .28    .8|           | 11.8  10.1|        |
-------------------------------------------------------------------------------------------
```

Figure 4. Person Misfit Order

From the misfit person table above, it is known that the person or subject who is not fit based on the PTMEA Corr value is 28 people, namely person serial numbers 73, 21, 230, 74, 50, 158, 4, 45, 308, 96, 36, 24, 49, 84, 98, 106, 15, 276, 211, 211, 209, 18, 68, 146, 1, 347, 307, and 144. The person or subjects are categorized as not fit based on the PTMEA Corr because they have negative values and are not in within the specified range, namely 0.4 < Pt Measure Corr < 0.85. Persons who are not fit because their scores are not in the MNSQ fit category (0.5 < MNSQ < 1.5) are persons or subjects with serial numbers 73, 21, 230, 74, 50, 158, 4, 45, 308, 96, 36, 24, 49. Persons who are not fit because they do not meet the ZSTD category (-2.0 < ZSTD < +2.0), namely person numbers 73, 21, 230, 74, 50, 158, 36, 24, 211, and 146.

Persons who do not meet the criteria for PTMEA Corr as well as MNSQ, namely person serial numbers 73, 21, 230, 74, 50, 158, 4, 45, 308, 96, 36, 24, 49, while persons who do not meet the criteria for PTMEA Corr as well as ZSTD are persons 73, 21, 230, 74, 50, 158, 4, 45, 308, 96, 36, 24, 49. Persons who do not meet the MNSQ criteria as well as ZSTD are person numbers 73, 21, 230, 74, 50, 158 , 36, and 24. Persons who do not meet the MNSQ, ZSTD, and PTMEA Corr criteria are persons with numbers 73, 21, 230, 74, 50, 158, 36, and 24. Thus, persons who do not fit or do not meet the three These categories can be omitted.

Bias (DIF)

To detect the presence or absence of bias in the items analyzed, it can be seen and observed the Person DIF Plot Graph and figure 5 below.

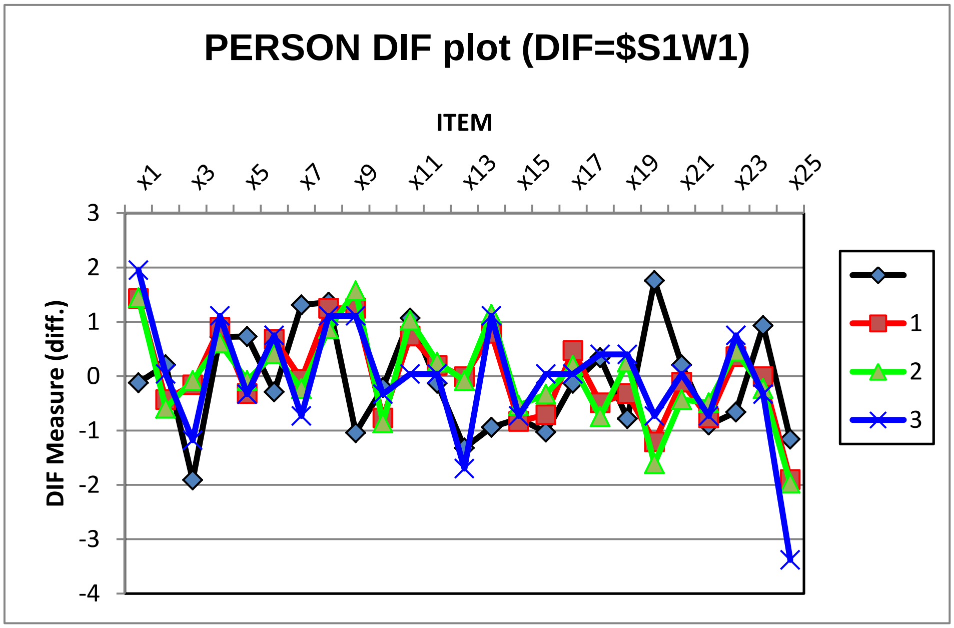


Figure 5. Person DIF Plot

This plot is a visual version of the statistical analysis already shown in the output figure 5. The graph shows the relative difficulty of the items for each group. The further away the graph points are from the mean, the more difficult the item is for the group. There are four curve lines that describe the four groups formed. The red line represents group 1, the green line is group 2, the blue line is group 3, and the black line shows the average value. From the graph it can be seen roughly that the distance of the DIF measure values between the three groups that are farthest from the average is item number 1, 7, 9, 14, and 20. In other items, some are far away and some are not too far away.

The furthest distance of the items from the average indicates that there is a fairly large difference in the level of difficulty between the groups. In this case, there are groups who are more advantaged and there are groups who are disadvantaged because an item appears to be more difficult for that group than other groups. The graph shows that item number 1 is more profitable for groups 1 and 2. Item number 7 benefits group 1, while item number 9 is more profitable for group 3 when compared to group 2 and group 1. item number 20, group 3 benefits the most.

Test Information Function

The test analyzed in this paper is a mid-semester test in the field of Indonesian language studies for first semester students. The mid-semester test aims to determine the extent to which students understand the course material for half a semester. The multiple-choice Indonesian language exam, as analyzed in this paper, has long been administered as a summative and formative test on campus. The students themselves are strictly selected students from all over Indonesia. Therefore, it can be roughly predicted that their ability in academics tends to be high. By analyzing these Indonesian questions, it can be seen whether the questions given so far are in accordance with their level of ability or not. Therefore, it is necessary to know in general the test information function of the analyzed questions. The function of test information from the analysis of students' Indonesian mid-semester examination questions is presented in the form of a plot as follows.

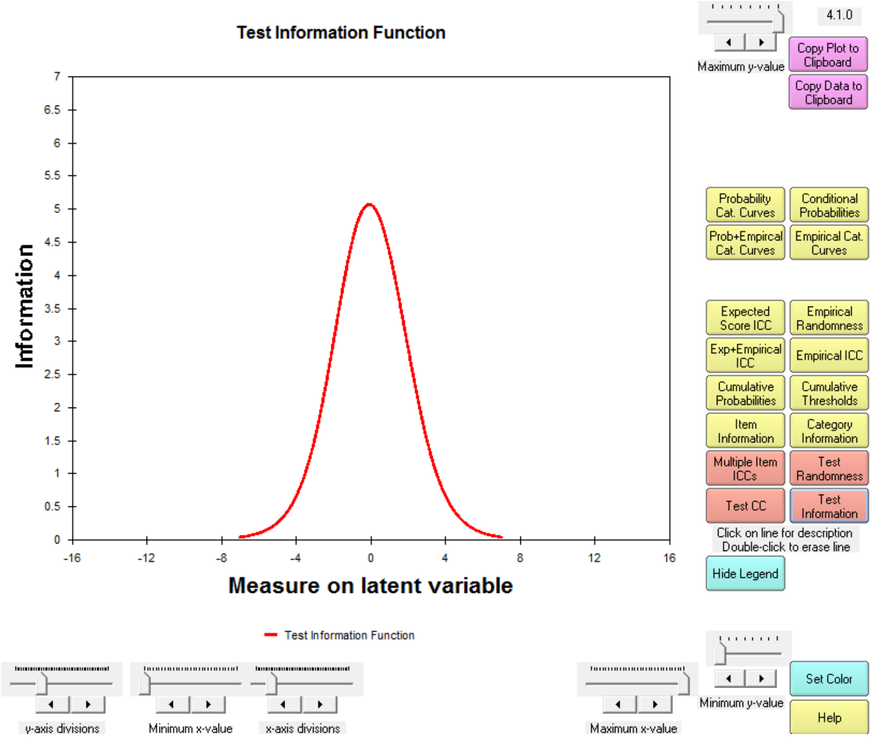


Figure 6. Test Information Function Graph

From the graph above, it can be concluded that of the 25 questions presented to 314 subjects (students) it shows that the items are suitable for low-ability students. This is indicated by the height of the curve which only reaches an information value of around 5. Thus, the items in the Indonesian final exam are not suitable to be given to students as an exam because the questions are too easy for those with high average ability.

5. Conclusion

The conclusions obtained from the item analysis are as follows. The data used in this study is convergent data. The unidimensional condition is not met by this test. There are several items in the test that do not meet the local independence requirements, while the monotonicity requirements have been met. The ability of the subject on this test is greater than the level of difficulty of the questions indicated. The reliability of the test is generally satisfactory with the quality of the items and the consistency of the subject's answers are quite good.

The test questions are relatively too easy for test takers who generally have high abilities. Persons or subjects who are not fit based on the PTMEA corr score are 28 people, while the items that are not fit are 5 items. There are 11 items indicated by DIF or bias. The items on this test are suitable for students with moderately low abilities. Thus, the items in the Indonesian Language Exam test are not suitable to be given to students as a test because the questions are too easy for those who have a high average ability.

Acknowledgements

I am very grateful to Promoter Dr. Muchlas Suseno M.Pd and co-Promoter Bahrul Hayat, Ph.D, who have contributed to this article.

References

Abulela, M. A., & Harwell, M. M. (2020). Data Analysis: Strengthening Inferences in Quantitative Education Studies Conducted by Novice Researchers. *Educational Sciences: Theory and Practice, 20*(1), 59-78. https://doi.org/10.12738/jestp.2020.1.005

Aiken, L. R. (1985). Three Coefficients for Analyzing the Reliability and Validity of Ratings. *Educational and Psychological Measurement, 45*(1), 131-142. https://doi.org/10.1177/0013164485451012

Aksoy-Pekacar, K., Kanat-Mutluoğlu, A., & HAKKI-ERTEN, İ. S. (2020). " I am just shy and timid": Student teachers' explanations for their performances of their presentations. *Eurasian Journal of Applied Linguistics, 6*(3), 459-480. https://doi.org/10.32601/ejal.834657

Alagumalai, S., Curtis, D. D., & Hungi, N. (2005). *Applied Rasch measurement: A book of exemplars*. Dordrecht: Springer. https://doi.org/10.1007/1-4020-3076-2

Anunti, H., Vuopala, E., & Rusanen, J. (2020). A portfolio model for the teaching and learning of GIS competencies in an upper secondary school: A case study from a finnish geomedia course. *Review of International Geographical Education Online, 10*(3), 262-282. https://doi.org/10.33403/rigeo.741299

Barabadi, E., Robatjazi, M. A., & Bayat, M. (2020). A phraseological examination of research articles in the field of environment using key phrase frame. *Eurasian Journal of Applied Linguistics, 6*(1), 81-100. https://doi.org/10.32601/ejal.710217

Bond, T. G., & Fox, C. M. (2013). *Applying the Rasch model: Fundamental measurement in the human sciences* (2 ed.). New York: Psychology Press. https://doi.org/10.4324/9781410614575

Boone, W. J., Staver, J. R., & Yale, M. S. (2013). *Rasch analysis in the human sciences* (1 ed.). London: Springer doi: https://doi.org/10.1007/978-94-007-6857-4

Bulut, A. (2020). Teacher Opinions about Children's Awareness of Zero-Waste and Recycling in the Pre-School Education Years. *Review of International Geographical Education Online, 10*(3), 351-372. https://doi.org/10.33403/rigeo.689426

Caliskan, A., & Zhu, C. (2020). Organizational Culture and Educational Innovations in Turkish Higher Education: Perceptions and Reactions of Students. *Educational Sciences: Theory and Practice, 20*(1), 20-39. https://doi.org/10.12738/jestp.2020.1.003

Demir, S. (2020). The role of self-efficacy in job satisfaction, organizational commitment, motivation and job involvement. *Eurasian Journal of Educational Research, 20*(85), 205-224.

Deveci, T. (2020). The introduction section of research articles in English and Turkish: The case of educational sciences–a preliminary study. *Eurasian Journal of Applied Linguistics, 6*(1), 119-140. https://doi.org/10.32601/ejal.710233

Erdil-Moody, Z., & Thompson, A. S. (2020). Exploring motivational strategies in higher education: Student and instructor perceptions. *Eurasian Journal of Applied Linguistics, 6*(3), 387-413. https://doi.org/10.32601/ejal.834670

Erturk, A., & Ziblim, L. (2020). Is the perception of organizational deviation affected by the organizational climate? Research in schools. *Eurasian Journal of Educational Research, 20*(85), 1-22. https://doi.org/10.14689/ejer.2020.85.1

Hambleton, R. K., & Swaminathan, H. (1985). Assumptions of Item Response Theory. In R. K. Hambleton & H. Swaminathan (Eds.), *Item Response Theory: Principles and Applications* (pp. 15-31). Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-017-1988-9_2

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory* (Vol. 2). California: SAGE Publications. Retrieved from https://psycnet.apa.org/record/1991-98425-000.

Hassan, A. B. (2020). Exploring global citizenship as a cross-curricular theme in Moroccan ELT textbooks. *Eurasian Journal of Applied Linguistics, 6*(2), 229-242. https://doi.org/10.32601/ejal.775801

Kasalak, G., & Dagyar, M. (2020). The Relationship between Teacher Self-Efficacy and Teacher Job Satisfaction: A Meta-Analysis of the Teaching and Learning International Survey (TALIS). *Educational Sciences: Theory and Practice, 20*(3), 16-33. https://doi.org/10.12738/jestp.2020.3.002

Linacre, J. (2011). *A user's guide to WINSTEPS [Computer Manual]*. Chicago: Winsteps.

Luber, L., FÖGELE, J., & Mehren, R. (2020). How Do Students Experience a Deprived Urban Area in Berlin? Empirical Reconstruction of Students' Orientations. *Review of International Geographical Education Online, 10*(4), 500-532. https://doi.org/10.33403/rigeo.763170

Mameche, Y., OMRI, M. A., & Hassine, N. (2020). Compliance of Accounting Education Programs with International Accounting Education Standards: The Case of IES 3 in Tunisia. *Eurasian Journal of Educational Research, 20*(85), 225-246. https://doi.org/10.14689/ejer.2020.85.11

Mardapi, D. (2008). Techniques for preparing test and non-test instruments. Yogyakarta: Mitra Cendikia Press.

Mardapi, D. (2017). *Measurement of Educational Assessment and Evaluation* (2nd ed.). Yogyakarta: Parama Publishing.

Meyer, J. P., & Zhu, S. (2013). Fair and equitable measurement of student learning in MOOCs: An introduction to item response theory, scale linking, and score equating. *Research & Practice in Assessment, 8*, 26-39. Retrieved from https://www.learntechlib.org/p/157939/

Osterlind, S. J. (1983). *Test item bias*. Beverly Hills, CA: Sage Publication Inc. https://doi.org/10.4135/9781412986090

Rasyid, H., & Mansur. (2008). *Assessment of Learning Outcomes*. Bandung: Wacana Prima.

Sopandi, W., & Sukardi, R. R. (2020). Using four-tier diagnostic tests to understand the conceptions held by pre-service primary school teachers about sea pollutant migration. *Review of International Geographical Education Online, 10*(2), 13-29. https://doi.org/10.33403/rigeo.629388

Stiggins, R. J., & Chappuis, J. (2012). *An introduction to student-involved assessment for learning*. Boston: Pearson.

Sukardi, H. (2008). *Evaluation of principle and operational education.* Jakarta: Bumi Aksara.

Sumintono, B., & Widhiarso, W. (2015). *Application of rasch modeling on educational assessment.* Cimahi, Indonesia: Trim komunikata. Retrieved from http://eprints.um.edu.my/id/eprint/14228

Suprananto, K. (2012). *Educational measurement and assessment*. Yogyakarta: Graha Ilmu.

Torun, F. (2020). The effect of a textbook preparation process supported by instructional technology tools on the TPACK self-confidence levels of prospective social studies teachers. *Review of International Geographical Education Online, 10*(2), 115-140. https://doi.org/10.33403/rigeo.691943

Wright, B. D., & Stone, M. H. (1979). *Best Test Design. Rasch Measurement*. Chicago, IL: MESA Press. Retrieved from https://core.ac.uk/download/pdf/212418787.pdf

Copyrights