EVALUATION OF PSYCHOMETRIC PROPERTY ITEMS
OF COMPUTER-BASED NATIONAL EXAMS BASED ON
ITEM RESPONSE THEORY AND RESPONSE TIME

Deni Hadiana
Jakarta State University, Indonesia

Bahrul Hayat
Islamic State University Syarif Hidayatullah, Indonesia

Burhanuddin Tola
Jakarta State University, Indonesia
denihadiana_pep14s3@mahasiswa.unj.ac.id

ABSTRACT

The quality of UNBK results is strongly determined by psychometric properties of items and response time of each examinee. This study aims to analyze and evaluate the psychometric properties of UNBK items and find the relationship between response time and difficulty index (b) using IRT 1 Parameter. The data used in this study were in the form of log file data from each UNBK examinees in IPA 2017/2018 Academic Year in DKI Jakarta. The difficulty index (b) is related to response time, then the relationship is described. The results of the psychometric property analysis of items obtained difficulty of items: easy 6 items; moderately difficult 28 items; and difficult 6 items. Psychometric property evaluation results obtained all items fit with the model. Person measure shows that examinee's abilities are still above the difficulty items, item reliability indicates that the reliability or consistency of the examinees' answers is in the criteria are sufficient while the item reliability is very good, the separation value indicates the quality of the instrument is good, the information function of the test package is very close to the normal curve shape. An increase of ability was not followed by a decrease in response time. Item difficulty, level cognitive items, and ability are not linear relationship with response time.

KEYWORDS

UNBK, Classical Test Theory, Item Response Theory, Response Time.

1. INTRODUCTION

Board of National Education Standards (BSNP) and Center for Educational Assessment (Puspendik) collaborate in organizing the National Examination (UN) and improving the quality of UN results. UN results can be used for consideration of selection, mapping of education quality, and policy intervention to improve education quality. The use of UN results will be meaningful, appropriate, and effective if UN scores obtained by examinees are accurate. This means that the score obtained by examinees reflects the real ability of examinees. To improve the accuracy of UN scores, BSNP and Puspendik endeavor to improve the quality of UN governance, one of which is through the implementation of Computer-Based National Examination (UNBK) since 2015.

UNBK can reduce the level of fraud during the test, according to Haladyna & Downing (2005), using computer-based test can minimize test fraud, especially cheating during the test, such as answer copying and memorization of response string. In computer-based test, the items are drawn randomly from the item pool so that the examinee takes a different set of items making it difficult for examinee to cheat each other during the exam. In addition to the implementation of UNBK, UN results will be increasingly able to predict accurately or provide more accurate information.
about the ability of examinees through efforts to improve the quality of items. The quality of items can be seen from the psychometric property analysis of items related to the validity and reliability.

To know the validity and reliability of items, according to Magno (2009), psychometric theory offers two approaches in analyzing test data: classical test theory (CTT) and item response theory (IRT). Both theories enable to predict outcomes of psychological tests by identifying parameters of item difficulty and the ability of test takers. Both of these approaches provide measures of validity and reliability. Hambleton and Jones (1993), measurement specialists have a choice of working within a classical test theory framework or an item response theory framework, or even a combination of frameworks. Next, Hambleton and Jones (1993) said that classical test models are often referred to as "weak models" because the assumptions of these models are fairly easily met by test data. Item response models are referred to as strong models too, because the underlying assumptions are stringent and therefore less likely to be met with test data. For example, the popular one-, two-, and three-parameter logistic models make the strong assumption that the set of items that compose the test are measuring a single common trait or ability. Therefore, the quality of items can be analyzed through the classical test theory approach and item response theory.

In addition to psychometric property information, UNBK provides information on the response time of each examinee to the items during the exam. This response time is very useful for getting information on the behavior of each examinee during the exam. This study aims to evaluate the psychometric properties of UNBK items and find the relationship between response time and difficulty index (b) using item response theory 1 parameter (Rasch Model).

1.1. Classical Test Theory

Classical test theory (CTT) is a theory about test scores, there are three concepts of test scores: observed score, true score, and error score. Observed score often called the test score. Hambleton and Jones (1993), within theoretical framework, models of various forms have been formulated, for example classical test model. It is a simple linear model that linking the observable test score \(X\) to the sum of two unobservable (latent) variable, true score \(T\) and error score \(E\) by the equation:

\[
X = T + E
\]

In this equation, we can see two unknown variables, so the equation is not solvable unless some simplifying assumptions are made. The assumptions in the classical test model are (1) true scores and error scores are uncorrelated, (2) the average error scores in the population of examinees is zero, (3) error scores on parallel test are uncorrelated.

Most of the work in classical test theory has focused on models at the test-score level. The models have linked test scores to true scores rather than item scores to true scores. Item statistics in classical test theory (item difficulty and item discriminating power) have been connected to test statistics such as test-score mean, standard deviation, and reliability. These item statistics have proven valuable in the test development process. Still, one main shortcoming is that they are sample dependent, and this dependency reduces their utility. To minimize this shortcoming, test developer has to ensure the examinee sample is similar to the examinee population. The sample differs in some unknown way from the population could easily happen in a field test, this condition make the utility of the item statistics may be reduced. The use of anchor items in a field test that also appeared in an actual test administration can be used to partially resolve sampling problems. In classical test theory, person parameters (true scores) and item parameters (item difficulty and item discrimination) are dependent on the test and the examinee sample, these dependencies can limit the utility of the person and item statistics.
1.2. Item Response Theory

In the item response theory (IRT), person parameters and item parameters are independent and how performance relates to the abilities that are measured by the items in the test, the ability is symbolized $\theta$. The item response may be dichotomous or two categories, such as yes or no, right or wrong, and multiple choice scores (0 or 1). Or, it may be polytomous or more two categories, such as a rating from a judge. Item parameters in IRT take into account the item difficulty and item discrimination. The item difficulty is useful to match the trait levels of a target population, such as the items on a fifth grade natural science test should not be so easy that the average fifth-grader answers nearly all the items correctly, nor should they be so difficult that the average student answers nearly all of them incorrectly. Item discrimination ($a$) is useful for selecting items that differentiate well between examinees with low and high levels of the proficiency measured by the test items. This index is sometimes called the slope, because it indicates how steeply the probability of correct response changes as the proficiency or trait increases. Items that are more discriminating or more reliable are weighted more heavily, so IRT scores can be more reliable than classical test theory.

In IRT, the difficulty index ($b$) is on the same metric as the ability. The item difficulty identifies the ability at which about 50% of the examinees are expected to answer the item correctly. For example, if $b = 0.2$, then about 50% of examinees with ability $= 0.2$ would get the item right. A larger percent of examinees with ability $= 0.5$ would get the item right. According to Demars (2010), in IRT, the information function is used to calculate the standard error of measurement and the reliability. The test information is a function of proficiency and the items on the test. Thus, test information varies with the proficiency level, as shown in Figure 1.

The standard error of measurement is the inverse of the square root of information, so that the greater the information, the smaller the standard error and the greater the reliability. An advantage of IRT is that the information function can also be defined at the item level, and the item information functions sum to the test information function. This is useful because items from different test forms can be assembled in different configurations, and the test information can be calculated for each new test form before the test is administered.

Also, in IRT, subsets of items can be removed from a test form and the corresponding item information functions can be subtracted to quickly calculate the new test information. Finally, if a reliability estimate of the IRT scores is desired for a sample or population of examinees, it can be estimated based on the item parameters and the distribution of the trait in that group of examinees. Another advantage of IRT is population invariance of the item parameters. This means that the item parameters should be the same in different populations of examinees. In IRT, the item difficulty, $b$, is the same (invariant) across samples, up to a linear transformation. A linear transformation means that the $b$’s from one population of examinees are multiplied/divided by a
constant and another constant is added or subtracted, to put them on the same metric as the b’s from another population of examinees.

Dichotomous items are only two categories that show the probability of score of 1, the probability of a score of 0 is 1 minus the probability of a score of 1. On academic test like UNBK items, correct responses are scored 1. The probability of a correct response is as a function of θ. The typical models used for dichotomous items are the three-parameter logistic (3PL), the two-parameter logistic (2PL), and the one-parameter logistic (1PL). These models are named by the number of item parameters used in the function, 3PL is for discrimination, difficulty, and guessing, 2PL is for discrimination and difficulty, 1PL is for difficulty. These parameters are typically labeled a, b, and c. In this article, we will focus on IRT 1LP.

The difficulty parameter, b, tells how difficult the item is. Its value equals the θ value where the slope of the function is steepest. About 50% of examinees with θ = b would score 1.

Figure 2. Items with different b-parameters, item 1 is more difficult than item 2.

![Figure 2](image)

Figure 2 shows two items with different b, item 1 more difficult than item 2, it is mean that the probability of getting item 1 right is lower than the probability for getting item 2 right. The mathematical form of 1PL model is:

\[
P(\theta) = \frac{1}{1+e^{(\theta-b)\beta}}
\]

The mathematical form of 1PL is equivalent to the Rasch model which was developed separately by George Rasch. Demars (2010) said the Rasch model was originally specified in terms of odd [probability/(1- probability)] or log-odds –the natural log of odds is called logits), it is now often specified in terms of probability. Next, the notation system in the Rasch model is different, using δ in place of b and β in place of θ. The mathematical form of Rasch model is:

\[
P(\beta) = \frac{1}{1+e^{(\beta-\delta)\beta}}
\]

\[
\ln\left(\frac{P(\beta)}{1-P(\beta)}\right) = \beta - \delta
\]

P(β) is the probability of correct response given β and δ. For consistency with the rest of this article, the symbol θ and the symbol b are used in this article. Still according to Demars (2010), values for θ range from −∞ to +∞ theoretically, but most examinees will have values -3 to +3.

There are three assumptions of IRT: unidimensionality, local independence, and correct model specification. Unidimensionality, a test that is unidimensional consists of items that measure only
one dimension or one ability. It is an implicit assumption that the items share a common primary construct, the model has a single \( \theta \) for each examinee and any other factors affecting the item response are treated as random error. It is mean in item characteristic curve is one for all group members at a certain level of ability. **Local independence**, a second assumption is that the responses to an item are independent of any other item. Ayala (2009) said that how a person responds to a question is determined solely by his or her location on the latent continuum and not by how he or she responds to any other question on the examination. Local Independence which means that the examinee’ response to an item does not affect examinee response to any other item. Local independence implicitly is achieved if the unidimensionality assumption is achieved (Hambleton & Swaminathan, 2010). **Correct model specification**, a third assumption is the functional form which states that the data follow the function specified by the model. This curve represents the relationship between the ability of a respondent and the probability of the correct response to the item, and it varies depending on the model used in the IRT.

1.3. Response Time

Test theorists have always been interested by the relationship between responses to the test items and the time used by the examinee to response them. In educational testing, it was rather difficult to record response time than item response before tests were administered on computers. In computerized testing like UNBK, the examinee’ responses as well as their response times on the items are recorded. The information in the item response help to evaluate the psychometric property of item to improve item, test, and testing administration, such as item calibration, ability estimation. The information in the response times is used to predict anomaly behavior of the examinee during the exam. Lee and Chen (2011) defined the term response time (RT), they stated that RT refers to the time an examinee spends on an item in test. In high-stakes test like UNBK, the information of RT is very useful to improve high-stakes testing situations. There are easy ways to incorporate RT into operations. According to Lee and Chen (2011), afterward examination of test speededness can be devised using charted RTs of each item or descriptive RT statistics such as mean, percentiles, standard deviation, etc. In this simple situation with only one fixed item sequence, each item only appears in one position in the test. Therefore, if RTs seem to diminish with later item positions, it implies that at least some examinees may be rushing their responses when looking at the later items. It is also possible to use RT to ensure that time limits are generous enough to avoid speededness, in addition to test development experience, to post test feedback indicating that most (if not all) examinees answer all items, or to descriptive RT statistics when available.

Fox, Entink, and Linden (2007) discuss the issue of how response time has been approached from three different angles. One approach is to model the response times with time parameters added to a regular item response theory (IRT) model. A second approach is characterized by modeling the response times separately from the responses. Linden (2006) discusses a selection of these models for response times on test items. In a third approach, introduced in Linden (2007), the response times and responses are modeled hierarchically. At the first level, both the distributions of the responses and response times are assumed to follow separate models, each with a different set of person and item parameters. The person parameters represent the speed and accuracy (or ability) of the test taker on the items. A test taker’s choice of speed and accuracy is generally constrained by a tradeoff. But since the speed and accuracy is assumed to be stationary during the test, the tradeoff can be ignored. Hence, at this first level of modeling, the item responses and response times can be assumed to be conditionally independent given the speed and accuracy parameters. However, at the second level, these parameters are allowed to be dependent. This leads to a hierarchical modeling framework in which the relation between speed and accuracy is dependent on the level of modeling. Focus of this paper is to describe response time of examinee’ UNBK.
2. METHOD

The data used in this study were in the form of log file data from 966 UNBK examinees in IPA subject for 2017/2018 Academic Year in DKI Jakarta. Log file data is converted into structured data and then analyzed through item response theory 1 parameter and response time statistically descriptive.

3. RESULT AND DISCUSSION

Item and person map or The Wright Map in Figure 3 shows the logit scale, which is the measurement unit common to both person ability and item difficulty, person and items are located on the map according to their ability and difficulty estimates. From the map we can see that item 10 is the most difficult item while item 35 is the easiest item. There are six items below one standard deviation and above one standard deviation. Most items are in -1 to +1 standard deviation range. Item 6 and item 23 are calculated as being at the mean of item difficulty (0).

Figure 3. Items person map for UNBK Examinee.

There are three categories of ability person: high ability with measure >+1 SD (>0.70), low ability with measure <-1SD (<-0.70), and mid ability with -0.70≤measure≤0.70. Six items are difficult, six items are easy, and 28 items are moderately difficult.

There are around 20 examinees have the logit above the logit of item 10 (the most difficult item), it is mean their probability to answer all item correctly. In Figure 4, we can see the summary of the analysis results.

| RAW | MODEL | INFIT | OUTFIT |
|-----|-------|-------|--------|
| SCORE | MEASURE | ERROR | MNSQ | ISTD | MNSQ | ISTD |
| S.D. | .00 | .00 | .06 | 3.1 | .08 | 3.0 |
| MAX. | 725.0 | 966.0 | 1.33 | 6.6 | 1.16 | 6.1 |
| MIN. | 222.0 | 966.0 | -1.21 | .83 | -8.8 | .80 |

Figure 4. SUMMARY OF 40 MEASURED, 966 Persons
The reliability index of 0.99 suggest that we can quite readily rely on this order of item estimates to be replicated when we give the UNBK Exam to other sample for whom it is suitable. Before we discuss the item fit of UNBK, in the Rasch model, fit is at the core of Rasch measurement. According to Bond and Fox (2015), the beautiful properties of Rasch measurement exits only to the extent that the data fit the Rasch model. Next, according to Boone et al. (2014), outfit mean square (MNSQ) can be used to assess misfits of item, Boon suggest that item with 0.5<MNSQ<1.5 is fit to the model. Based on Figure 5 we concluded that all items of UNBK IPA are fit to the model.

Figure 5. Item difficulty and MNSQ

Other useful information from Figure 5 is the categorization of items, there are three categories of difficulty items: the items is difficult with measure >1 SD (>0.65), easy items with measure <1SD (<-0.65), and moderately difficult items with -0.65≤measure≤0.65. Six items are difficult, six items are easy, and 28 items are moderately difficult.

Figure 6 shows differential item functioning and item bias. According to Ayala (2009), although the term bias has a statistical interpretation, the systematic under or overestimation of a parameter, in the layperson’s mind bias is typically associated with the issue of unfairness. That is, an instrument that has an adverse impact on different ethnic or racial groups. As such, the terms item bias and test bias have certain culturally negative connotations. To identify items as biased involves using differential item functioning (DIF) methods to detect items that are functioning differently across manifest groups of individuals. In Figure 6, we can see that there are five items (item 38, item 21, item 17, item 16, and item 10) are identified as biased items because the items that display different statistically properties (prob. < 0.005) between SMP students and MTs students. That is, item 21 (0.039), item 10 (0.031), item 16 (0.007), item 17 (0.001), and item 38 (0.000). From Figure 6, the MTs students are disadvantaged by item 10, item 16, and item 38.

Figure 6. Person DIF Plot SMP and MTs Students
In Figure 7, we can find the information about average of response time, item difficulty, and the level of item. Here, the level of item is divided into three categories: level 1 is knowledge and understanding (black number), level 2 is applying (brown number), and level 3 is reasoning (red number). Figure 7 shows that there is not a linear relationship between item difficulty and level of item to response time.

Both an increase in the cognitive process from level 1 to level 3 and an increase of the item difficulty from -1.4 to +1.4 were not followed by an increase in response time for the items. Item 26 is the item with the fastest average time while item 40 is the item is the longest, item 26 and item 40. Both items have the same level of difficulty. Item 26 measures knowledge and understanding while item 40 is reasoning item. Item 35 is the easiest item while item 10 is the most difficult. Both items have almost the same average time. Item 35 is knowledge and understanding while item 10 is application.

The next pictures show the relationship of ability, difficulty level, and average time. Figure 8 displays the distribution of ability for all items (1), easy items (2), moderately items (3), and difficult items (4).
Examinees with mid ability tend to take more time to response to the items than those who have high ability or low ability for easy items, moderately items, and difficult items. An increase of ability was not followed by a decrease in response time. Response time increases from low ability to mid ability and decreases from mid ability to high ability for easy, moderately, and difficult items.

Figure 9 displays distribution response time for 3 examinee’s low ability (-2.09; -1.06; 0.8), mid ability (-0.68; 0; 0.68), and high ability (2.69; 1.20; 0.80). The information of this figure is consistent with information in Figure 8 that it concluded an increase of ability was not followed by a decrease in response time. One examinee with mid ability spent the longest time to respond easiest item and most difficult item.

According to Linden (2009), one of the very first to address the relation between responses and RTs from a perspective now known as item response theory (IRT) was Thurstone (1937). The results of this study relevant to Thurstone in 1937 as quoted by Linden (2009) that a decrease of the probability of success with the difficulty of the item but an increase of the probability with time. The probability of success on a more difficult item can always be compensated by spending more time on it. The probability of success is as a function of the difficulty of the item. Related to relationship response time and cognitive level, Sab et al. (2011) concluded that pictures in the answer options of test items can reduce the time needed to respond to test items. Thus, by using pictures in the answer options of test items, it seems to be possible to increase the efficiency of testing without compromising the accuracy with which test performance is being measured. This seems to be particularly true for test items that are relatively low in complexity. More time-efficient testing formats might be relevant in large-scale assessments because the costs associated with employing large-scale tests might be lowered when testing time can be reduced. Therefore the response time to the items which high in complexity spend more time than the items which low in complexity.

Figure 9. Distribution of time for nine examinees.

Based on distribution of nine examinees like in Figure 9, we concluded that item difficulty, level cognitive items, and ability are not linear relationship with response time. According to Sab et al. (2011), the increase in response time is affected by the complexity of item.
This research is still focused on describing ability, item difficulty, level of cognitive process, and response time separately. Therefore future research need to investigate the modeling of responses and response times.

4. CONCLUSION

The results of the psychometric property analysis of items obtained difficulty of items: easy 6 items; moderately difficult 28 items; and difficult 6 items. Psychometric property evaluation results obtained all items fit with the model. Person measure shows that examinees' abilities are still above the difficulty items, item reliability indicates that the reliability or consistency of the examinees' answers is in the criteria are sufficient while the item reliability is very good, the separation value indicates the quality of the instrument is good, the information function of the test package is very close to the normal curve shape. An increase of ability was not followed by an decrease in response time. Item difficulty, level cognitive items, and ability are not linear relationship with response time.

ACKNOWLEDGEMENT

The author thank to Center for Educational Assessment as an institution that helps author to collect the data.

REFERENCES

Ayala, (2009). The Theory and Practice of Item Response Theory. New York: The Guilford Press.
Demars, (2010). Item Response Theory: Understanding Statistics Measurement. New York: Oxford University Press.
Fox et al., 2007. Modeling of Responses and Response Times with the Package cirt. Journal of Statistical Software, Vol. 20, Issue 7. http://www.jstatsoft.org/
Hamleton & Jones, 1993. Comparison of Classical Test Theory and Item Response Theory and Their Applications to Test Development. An NCME Instructional Module on University of Massachusetts at Amherst, Washington DC., USA.
Haladyna & Downing, 2006. Handbook of Test Development. New Jersey: Lawrence Erlbaum Associates.
Hambleton, R., & Swaminathan, H. (2010). Item Response Theory. Principles and Application. Boston: Kluwer-Nigh off Publishing.
Linden, 2006. A Lognormal Model for Response Times on Test Items. Journal of Educational and Behavioural Statistics, Vol. 31, pp. 181–204.
Linden, 2007. A Hierarchical Framework for Modeling Speed and Accuracy on Test Items. Psychometrika. In press.
Linden, et al., 2007. IRT Parameter Estimation with Response Times as Collateral Information. Applied Psychological Measurement. Submitted.
Linden, 2009. Conceptual Issues in Response-Time Modeling. Journal of Educational Measurement, Vol. 46, No. 3, pp. 247–272.
Lee & Chen, 2011. A Review of Recent Response-Time Analyses in Educational Testing. Psychological Test and Assessment Modeling, Vol. 53, pp. 359-379.
Magno, 2009. Demonstrating the Difference between Classical Test Theory and Item Response Theory Using Derived Test Data. The International Journal of Educational and Psychological Assessment, Vol. 1, Issue 1, pp. 1-11.
Sab et al., 2012. Pictures in Test Items: Effects on Response Time and Response Correctness. Applied Cognitive Psychology, Appl. Cognit. Psychol., Vol. 26, pp. 70–81.