ALGORITHM DETERMINING ITEM DIFFICULTY ANALYSIS IN COMPUTER BASED TEST (CBT)

RAHMAT HIDAYAT, SISKA SULISTYORINI

State Islamic University Sunan Kalijaga Yogyakarta, Islamic University of Indonesia, Yogyakarta
Email: rahmat.hidayat@uin-suka.ac.id, Ichahidayat86@gmail.com, 154220102@uui.ac.id

Abstract- This research examines algorithm designing to determine item difficulty automatically of each answer denoted by examinee in Computer Based Test. As a quality can be identified reviewed from how many times the items could be answered correctly by the students based on the number of the answered items. Some researchers performed some application about item analysis but still need more energy to input all the data manually.

The research uses qualitative research with educational quantitative analysis test item analysis theory approach in CBT. It also required experiment to see the result of how algorithm working to solve the problem.

The result shows that (1) the design of algorithm could be use to analyze test result as well as the test difficulty. (2) The strength of the system are working fast, instant result, as a basic consideration to have good item quality and fair item distribution. The weakness of the algorithm design: it has not meet examinee character and need good socialization to make educator understand how to compose and operate CBT.

Keywords: Algorithm design, Item difficulty, CBT.

I. INTRODUCTION

Item Analysis describes the statistical analyses which allow measurement of the effectiveness of individual test items. (Scrolla, 2003). Item difficulty analysis is important because it reveals whether an item is too easy or too hard. In either case, the optimal item difficulty depends on the question-type and on the number of possible distractors.

There have been many item analysis application to help teachers in analyzing testing items such as ITEMAN, RASCAL, ASCAL, BILOG, FACETS dan CONQUEST. But particularity in developing country like Indonesia, applications are relatively did not work well due to language understanding problems and IT mastery problem.

One of local item analysis was invited and designed by Ali Muhsin (2014)and became popular to help teacher to do their item analysis problems. But it is a desktop based application where educators must input all of the exam result one by one that this research purpose to seek for the answer on: (1) how to design an algorithm to determine item test difficulty automatically of every answer denoted by examinee in CBT; and (2) the strength and the weakness of this algorithm, so the test result as well as the item difficulty analysis could be automatically seen some minutes after examinee answering the questions.

II. LITERATURE REVIEW

Previous Works

A number of researches examined the analysis of item test such as a paper by Susan, Matlock and Hetzel (1997). These analyses evaluate the quality of the items and of the test as a whole. The purpose of the present paper is to summarize the recommendations for item and test analysis practices, as these are reported in commonly-used measurement textbooks (Crocker & Algina, 1986).

Second is by Jerard Kehoe, (1995: 1-5) This article offers some suggestions for the improvement of multiple-choice tests using “item analysis” statistics. These statistics are typically provided by a measurement services, where tests are machine-scored, as well as by testing software packages.

Another interesting article is by Chris Elvin, about item analysis using microsoft excel spreadsheet. This article is written for teachers and researchers whose budgets are limited and who do not have access to purposely designed item analysis software such as Iteman (2003).

It describes how to organize a computer spreadsheet such as Microsoft Excel in order to obtain statistical information about a test and the students who took it. Using a fictitious example for clarity, and also a real example of a personally written University placement test, I will show how the information in a spreadsheet can be used to refine test items and make judicious placement decisions. Included is the web address for accessing the sample Excel files for the class of fictitious students (Elvin, 2003a, 2003b).

Although previous studies have also examined constructs that are conceptually linked with perceived theory, the articles haven’t discussed about the application of test difficulty and test quality reviewed from CBT. It motives the author to explore more in the terms of item quality as a base theory for item analysis application.

Theoretical Framework

Multiple Choice

Multiple choice items are considered to be among the most versatile of all item types. They can be used to...
Algorithm Determining Item Difficulty Analysis in Computer Based Test (CBT)

test factual recall as well as levels of understanding and ability to apply learning. (Clay, 2001: 13)
Multiple choice tests can be used to test the ability to: 1) recall memorized information 2) apply theory to routine cases 3) apply theory to novel situations 4) use judgment in analyzing and evaluating. Clay(2001: 14).
Unfortunately, multiple choice items are difficult and time consuming to construct well. They may also appear too discriminating (picky) to students, especially when the alternatives are well constructed and are open to misinterpretation by students who read more into questions than is there.

**Item Difficulty**

Item difficulty is simply the percentage of students taking the test who answered the item correctly. The larger the percentage getting an item right, the easier the item. The higher the difficulty index, the easier the item is understood to be (Wood, 1960).
To compute the item difficulty, divide the number of people answering the item correctly by the total number of people answering item. The proportion for the item is usually denoted as p and is called item difficulty (Crocker & Algina, 1986).
An item answered correctly by 85% of the examinees would have an item difficulty, or p value, of .85, whereas an item answered correctly by 50% of the examinees would have a lower item difficulty, or p value, of .50.
One cannot determine which item is more difficult simply by reading the questions. One can recognize the name in the second question more readily than that in the first. But saying that the first question is more difficult than the second, simply because the name in the second question is easily recognized, would be to compute the difficulty of the item using an intrinsic characteristic. This method determines the difficulty of the item in a much more subjective manner than that of a p value.
Another implication of a p value is that the difficulty is a characteristic of both the item and the sample taking the test. For example, an English test item that is very difficult for an elementary student will be very easy for a high school student. A p value also provides a common measure of the difficulty of test items that measure completely different domains. It is very difficult to determine whether answering a history question involves knowledge that is more obscure, complex, or specialized than that needed to answer a math problem. When p values are used to define difficulty, it is very simple to determine whether an item on a history test is more difficult than a specific item on a math test taken by the same group of students. (Matclock, Hetzel, 1997)
At the end of the Item Analysis report, test items are listed according their degrees of difficulty (easy, medium, hard) and discrimination (good, fair, poor). These distributions provide a quick overview of the test, and can be used to identify items which are not performing well and which can perhaps be improved or discarded (Washington Edu, 2017)

**Item Distractors**

One important element in the quality of a multiple choice item is the quality of the item's distractors. In a simple approach to distractor analysis, the proportion of examinees who selected each of the response options is examined. For the key, this proportion is equivalent to the item p-value, or difficulty.
If the proportions are summed across all of an item's response options they will add up to 1.0, or 100% of the examinees' selections. (proftesting, 2010).
The proportion of examinees who select each of the distractors can be very informative. (proftesting, 2010) For example, it can reveal an item mis-key. Whenever the proportion of examinees who selected a distractor is greater than the proportion of examinees who selected the key, the item should be examined to determine if it has been mis-keyed or double-keyed. A distractor analysis can also reveal an implausible distractor.

**III. RESEARCH METHOD**

**Research Sort**
The sort of this research performs qualitative research. Qualitative methods are first and foremost research method. They are ways of finding out what people do, know, think, and feel, by observing, interviewing, and analyzing documents. (Quinn, 2001: 295). This research describes the statistical analyses which allow measurement of item difficulties and measure item distractors quality.

**Data Collecting Method**
The Author uses observation focusing on the analysis of item question in part of quality and analysing material culture. Material Culture constitute particularly rich source of information about many organizations and program. (Quinn, 2001: 293)

**Data Analysis Method**
Analysis of this research will use the model f Miles and Huberman where they explained the process of data analysis process would be in these steps: (Miles and Huberman, 1984, 10-12). First is Data Collection is an activity to collect the data from the field focused of and suited to the procedure determined and the researcher needed.
Secondly is Data Reduction, it’s to summarize, to separate the data then be arranged according to the focus of the research quality item analysis.
Thirdly is Data Display, a neat and systematic data arrangement to be displayed focused on item analysis approached from how to write effective test items.
The Last is Data Verification, the author will take conclusion from the data displayed then they will
be completed by other supportive data which will make the research results perfect.

IV. RESULTS

5.1. Algorithm Design of item difficulty in CBT

This research was conducted in Indonesia. The subject was two classes of students studying Information Technology Security in Informatic engineering department, Faculty of Science and technology State Islamic University of Sunan Kalijaga Yogyakarta, Indonesia in 2015 and 2016.

Each class which consist of 40 students female and male. The mid semester and final test used computer based test.

Before designing the algorithm, authors previously did a six month in which students were provided with 70 question item in midterm test and 80 questions in final term test using CBT.

Some steps used to denote a CBT quick test are:
1. Teacher composes the multiple choice questions.
2. Examiner answered the questions.
3. For every submitted answer, system counted variable b and variable n.
4. b is number of correct answer
5. n is number of total answer.
6. Item difficulty would count every number \( P = \frac{b}{n} \) right away so the result can be seen soon.

It describes the algorithm pattern used to determine item difficulty.

\[
p = p(n), \quad b = b(n);
\]

\[
\text{if (answer is correct) then} \quad \{ b = b+1; \} \quad \text{n} = n+1; \quad \text{p} = \frac{b}{n};
\]

\( n = \text{number of answered item} \)

After the experiment of mid term test. The results showed in the table below.

<table>
<thead>
<tr>
<th>No</th>
<th>Item ID</th>
<th>U</th>
<th>F</th>
<th>T</th>
<th>Total</th>
<th>T / (U+F)</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>33</td>
<td>6</td>
<td>37</td>
<td>43</td>
<td>27</td>
<td>0.736</td>
<td>Medium</td>
</tr>
<tr>
<td>2</td>
<td>35</td>
<td>4</td>
<td>52</td>
<td>48</td>
<td>11</td>
<td>0.458</td>
<td>Medium</td>
</tr>
<tr>
<td>3</td>
<td>30</td>
<td>4</td>
<td>29</td>
<td>34</td>
<td>44</td>
<td>0.773</td>
<td>Medium</td>
</tr>
<tr>
<td>4</td>
<td>34</td>
<td>5</td>
<td>19</td>
<td>30</td>
<td>40</td>
<td>0.750</td>
<td>Medium</td>
</tr>
<tr>
<td>5</td>
<td>34</td>
<td>5</td>
<td>6</td>
<td>41</td>
<td>37</td>
<td>0.732</td>
<td>Medium</td>
</tr>
<tr>
<td>6</td>
<td>36</td>
<td>5</td>
<td>17</td>
<td>32</td>
<td>49</td>
<td>0.816</td>
<td>Medium</td>
</tr>
<tr>
<td>7</td>
<td>39</td>
<td>4</td>
<td>10</td>
<td>53</td>
<td>37</td>
<td>0.638</td>
<td>Medium</td>
</tr>
<tr>
<td>8</td>
<td>32</td>
<td>4</td>
<td>24</td>
<td>36</td>
<td>41</td>
<td>0.809</td>
<td>Medium</td>
</tr>
<tr>
<td>9</td>
<td>35</td>
<td>2</td>
<td>24</td>
<td>37</td>
<td>44</td>
<td>0.868</td>
<td>Medium</td>
</tr>
</tbody>
</table>

Table. example of algorithm for item difficulty in the system

In this table U is for unanswered questions, F is for false item. And T is for item answered true. In the next column, examiner put the algorithm system. So, everytime a student answer the item of question, the system would automatically count and report to examiner about real score.

This analysis used three scales of measurement: easy, medium and difficult. Easy id for result which is more than 0.70 above, medium is for result range 0.5 - 0.70, and difficult for result with range 0.1 - 0.5. For items with one correct alternative worth a single point, the item difficulty is simply the percentage of students who answer an item correctly.

In this case, it is also equal to the item mean. (washington.edu, 2017) The item difficulty index ranges from 0 to 100; the higher the value, the easier the question. When an alternative is worth other than a single point, or when there is more than one correct alternative per question, the item difficulty is the average score on that item divided by the highest number of points for any one alternative.

Items that are functioning very poorly should usually be removed from consideration and the exams re-scored before the test results are released. In other cases, items may still be usable, after modest changes are made to improve their performance on future exams.(proftesting, 2010)

One of the ways to improve item test is by improving item options. One of the techniques is through analyzing distractor quality. It is then important to discuss about item distractors.

Item Distractor algorithm

Item difficulty are usually also will connect to item discrimination and item distractors. The good analysis of item distractor would lead a teacher to have better quality items.

1. For every submitted answer, system counted variable d and variable n.
2. d is number of examinee answering the item.
3. n is number of total answer.
4. Item distractor would count every number of examinee who chose the answer over all the items.

The algorithm pattern for distractor difficulties are :

\[
d = d(n), \quad c = c(n);
\]

\[
\text{if (chosen item) then} \quad \{ c = c+1; \} \quad \text{n} = n+1; \quad \text{d} = \frac{c}{n};
\]

\( n = \text{number of answered item} \)

Everytime examinee submit the answer, the system will count the index of distractor in every chosen item.

5.2. Algorithm Strength and Weakness

As a new system made, author see some of the strengths as well as the weaknesses such as:
Algorithm Determining Item Difficulty Analysis in Computer Based Test (CBT)

Algorithm Strengths
The strength of the system are:
1. This system works fast.
2. Teacher could know the score as well as the analysis result of item difficulty and item distractor right away after student submitting the test.
3. Teacher don’t need to input all data first to know answer.
4. The data can be a basic consideration to have good item quality and fair item distributions.

Algorithm Weaknesses
The weakness of the algorithm design:
1. Teacher must understand how to use the system.
2. It needs more detail tutorial to work this item analysis.
3. The application may need internet access which some place has poor internet connectivity.
4. It has not meet examinee character.
5. It need good socialization to make educator understand how to compose and operate CBT questions.

CONCLUSION
It can be concluded that (1) the design of algorithm could be use to analyze test result as well as the test difficulty and create more for item distractors analysis. (2) The strength of the system are working fast, instant result, as a basic consideration to have good item quality and fair item distributions. The weakness of the algorithm design: it has not meet examinee character and need good socialization to make educator understand how to compose and operate CBT.

Further attention to this research is to explore more to item discrimination, etc to give more valuable benefit to ease teachers in creating qualified items.

REFERENCES