Empirical Comparison of Methods of Establishing Item Difficulty Index of Test Items Using Classical Test Theory (CTT)

H. J. Osarumwense 1* and S. O. Oyedeji 2

Department of Educational Evaluation and Counselling Psychology, Faculty of Education, University of Benin, Nigeria
rowenvbiugie@gmail.com

2Demonstration Secondary school, University of Benin, Benin City

Abstract
The aim of this study was to empirically compare two methods of computing the item difficulty index of test items based on Classical Test Theory (CTT). All the students in Upper Basic certificate class in Benin Metropolis made up the population of the study. Six hundred students who were preparing for the 2013 Upper Basic Certificate examinations were randomly sampled for the study. The 2010 Upper Basic Certificate Mathematics Objective questions which were made up of eighty multiple choice items served as the instrument for data collection. After administration of the instrument, the item difficulty index of each of the items was computed using the two methods. After the item analysis, it was found that the 2010 Upper Basic Certificate Mathematics objective questions were within the recommended range of 0.30-0.70. The findings also showed a positive strong relationship between the item difficulty indices obtained by using the two methods. Finally the findings revealed that there was no significant difference between the means of the item difficulty indices obtained by using the two methods. It was recommended that test developers should use percentage sample of examinees to compute test item difficulty index in order to reduce cost in terms of finances and time as both methods are not significantly different.

Key words: CTT, item analysis, item difficulty index and item discriminating index.

Introduction
Measurement instruments are subjected to empirical analysis to ensure that the instruments are good enough to be used to measure examinees’ abilities. To establish that a measurement instrument is standardized, such instrument is expected to be subjected to item analysis. Item analysis is a process which examines student responses to individual test items (questions) in order to assess the quality of those items and of the test as a whole. Item analysis is especially valuable in improving items which will be used again in later tests, but it can also be used to eliminate ambiguous or misleading items in a single test administration. Classical Test Theory (CTT) and Item Response Theory (IRT) are some of the tools available to identify the quality of items in a test. Both CTT and IRT have been used as standard methods of item analysis. Despite the claim that IRT is theoretically “superior” to CTT, CTT-based item difficulty and discrimination indices have been found “enough” in flagging weak items
Based on findings from various studies that compare the two approaches, according to Krishnan (2013), the bottom line is: despite its attractive feature of IRT, namely the capability of analyzing the unobservable (latent) variable, the results are very similar and the frameworks are quite comparable. He affirmed that there are also instances where CTT was found better than IRT in detecting item quality. CTT as a body of theory and research could predict or explain the difficulty of questions, provides insights into the reliability of test scores, and helps us toward coming up with an assessment of how to improve the test by maintaining and developing a pool of “good” items from which future assessments can be drawn.

Courville (2004) noted that item characteristics play a vital role in all the reliability that is efficient and that item analysis is used to help in the development of test items by maximising the needed score reliability while minimizing the number of items. Detection of poor items (at least for norm-referenced test) according to Hambleton and Jones (1993) is quite straightforward and is basically accomplished through careful study of item statistics. However, for an achievement test, items with difficulty indices between 0.3 and 0.7 are considered to be appropriate (Nelson 2001; Hingorjo & Jaleel 2012; Sayyah, Vakili, Alavi, Bigdeli, Soleymani, Assarian, and Azarbad 2012). Items with discriminating indices equal to 0.3 and above are considered to be appropriate (Zurawski 1999; Mitra, Nagaraja, Ponndurai and Judson 2009).

Different authors have used either the overall examinees or percentage sample of upper and lower examinees to carry out item analysis. For instance, Sim and Rasiah (2005) investigated “Relationship between Item Difficulty and Discrimination Indices in True/False-Type Multiple Choice Questions (MCQ) of a Para-clinical Multidisciplinary Paper”. They computed the item difficulty using the formula \( P = \frac{R}{T} \), where \( P \) is the difficulty index, \( R \) is the number of correct responses and \( T \) is the total number of responses (i.e., correct + incorrect + blank responses) but computed the discriminating index using difference between the percentage of students in the upper group (\( P_U \)), i.e., the top 27% scorers, who obtained the correct response, and the percentage of those in the lower group (\( P_L \)), i.e., the bottom 27% scorers, who obtained the correct response; thus \( D = P_U - P_L \) where \( D \) is the discriminating index, \( P_U \) the upper group and \( P_L \) the lower group. The results of their study revealed that there was a wide distribution of item difficulty indices in all the MCQ papers analysed and that the relationship between the difficulty index (\( P \)) and discrimination index (\( D \)) of the MCQ items in a paper was not linear, but more dome-shaped.

According to Eaves and Erford (2009), item difficulty is determined by the number of people who answer a particular test item correctly. Citing an example, Eaves and Erford (2009) stated that if the first question on a test was answered correctly by 76% of the class, then the difficulty level (\( p \) or percentage passing) for that question is \( p = .76 \). If the second question on a test was answered correctly by only 48% of the class, then the difficulty level for that question is \( p = .48 \). The higher the percentage of people who answer correctly, the easier the item, so that a difficulty level of .48 indicates that question two was more difficult than question one, which had a difficulty level of .76. The implication of this example is that the entire examinees responses should be used to compute the item difficulty index. Zurawski (1999), in a similar view, stated that the item difficulty index, symbolized \( p \), can be computed simply by dividing the number of test takers who answered the item correctly by the total number of students who answered the item. He stressed further that as a proportion, \( p \) can range between 0.00, obtained when no examinees answered the item correctly, and 1.00, obtained when all examinees answered the item correctly.
Mitra et al (2009) in their study titled “The Levels Of Difficulty And Discrimination Indices In Type A Multiple Choice Questions Of Pre-clinical Semester 1 Multidisciplinary Summative Tests” stated that the item difficulty index is calculated as percentage of the total number of correct responses to the test item. It was calculated using the formula $P = R/T$, where $P$ is the item difficulty index, $R$ was the number of correct responses and $T$ was the total number of responses (which includes both correct and incorrect responses). They stated further that an item was considered difficult when the difficulty index value was less than 30% and the item was considered easy when the index value was greater than 80%. Boopathiraj and Chellamani (2013) in their study titled “analysis of test items on difficulty level and discrimination index in the test for research in education” described the item difficulty as the proportion of the examinees that marked the item correctly and computed the difficulty index using the formula:

$$DL = \frac{Ru + Rl}{Nu + Nl}$$

Where,
- $DL$ = the difficulty level
- $Ru$ = the number students in the upper group who responded correctly
- $Rl$ = the number students in the lower group who responded correctly
- $Nu= Number of students in the upper group$
- $Nl= Number of students in the lower group$;

Hingorjo and Jaleel (2012) in their study titled “Analysis of One-Best MCQs: the Difficulty Index, Discrimination Index and Distractor Efficiency” stated that Difficulty index (p-value), also called ease index, describes the percentage of students who correctly answered the item. It ranges from 0 - 100%. The higher the percentage, the easier the item. There commended range of difficulty is from 30% - 70%. Items having p-values below 30% and above 70% are considered difficult and easy items respectively they computed the item difficulty index using the formulae:

$$Difficulty\ index,\ p = \frac{[(H+L) / N] \times 100}{N}$$

where, $N$ was the total number of students in both high and low groups. $H$ and $L$ were the number of correct responses in the high and low groups, respectively.

Matlock-Hetzel (1997) defined the item difficulty as the percentage of students taking the test who answered the item correctly. She stated further that to compute the item difficulty, divide the number of people answering the item correctly by the total number of people answering item.

The literature reviewed above indicated that most of the authors use the overall examinees’ responses to carry out test item analysis especially to compute the item difficulty index but none has empirically compared the item difficulty index obtained when the entire examinees responses were used and when percentage sample of lower and upper groups of examinees were used. The aim of this study was, therefore, to investigate whether or not there would be difference between the item difficulty indices when the overall examinees’ responses are used to compute the test item difficulty index and when the percentage samples are used. To do this the following research questions served as guide.

**Purpose of Study**

This study was aimed at investigating empirically if there is a difference in the item difficulty indices when computed using the entire number of examinees and when computed using...
percentage sample of the examinees. The study also examined if there is relationship between the item difficulty indices obtained when the two methods were employed.

**Research Questions**

1. Do all the items of the 2010 Upper Basic Certificate Mathematics objective Examination fall within the acceptable item difficulty index range of 0.3 to 0.7?
2. Is there a relationship between the item difficulty indices obtained by using the entire examinees responses and the one obtained by using percentage sample of the examinees?
3. Is there a difference between the difficulty indices obtained when using the entire examinees responses and the one obtained by using percentage sample of the examinees?

**Hypotheses**

The following null hypotheses were tested at 0.05 level of significance.

1. There is no significant relationship between the item difficulty index obtained by using the entire examinees responses and that obtained by using percentage sample of the examinees.
2. There is no significant difference between the item difficulty index obtained by using the entire examinees responses and that obtained by using percentage sample of the examinees.

**Significance of Study**

The result of this study is expected to provide useful information to National Examination Council (NECO), the body that is responsible for conduct of the Upper basic final examinations, on the quality of the Mathematics Objective test of 2010. The result of the study is also expected to guide the test developers on the best and more economical method, in terms of finances and time, to use when computing test item difficulty.

**Conceptual Framework**

**Item Analysis**

Assessment is an integral part of teaching-learning process. The reliability of the information obtained from assessment results depends largely on the quality of the instrument used. The quality of such instrument can only be ascertained by subjecting the instrument to item analysis. Item analysis is an important phase in the development of an examination program. In this phase statistical methods are used to identify any test items that are not working well. If an item is too easy, too difficult, failing to show a difference between skilled and unskilled examinees, or even scored incorrectly, an item analysis will reveal it. Item analysis, according to Mitra et al (2009), is the process of collecting, summarizing and using information from students' responses to assess the quality of test items. It is a valuable, yet relatively simple, procedure performed after the examination that provides information regarding the reliability and validity of a test item. It also tells how difficult or easy the questions were and whether the questions were able to discriminate between students who performed well on the test, from those who did not perform well (Hingorjo & Jaleel (2012).

Item analysis, according to Thompson and Levitov (1985) investigates the performance of items considered individually either in relation to some external criterion or in relation to the remaining items on the test. It is a process which examines student responses to individual test items (questions) in order to assess the quality of those items and of the test as a whole. It
is a statistical technique that helps instructors to identify the effectiveness of their test items. Item analysis is not only valuable in improving items which will be used again in later tests, but can also be used to eliminate ambiguous or misleading items in a single test administration. In addition, item analysis is valuable for increasing instructors' skills in test construction, and identifying specific areas of course content which need greater emphasis or clarity Office of Educational Assessment (OEA 2005). Mitra et al (2009) stated that Difficulty index (P) and Discrimination index (D) are two parameters which help evaluate the standard of MCQ questions used in an examination, with abnormal values indicating poor quality. In the same view, OEA (2005) reported that the two most common statistics reported in an item analysis are the item difficulty, which is a measure of the proportion of examinees who responded to an item correctly, and the item discrimination, which is a measure of how well the item discriminates between examinees who are knowledgeable in the content area and those who are not.

Wilson (2005) noted that item difficulty index is one of the most essential components of item analysis. This could be why it is the most useful, and most frequently reported, item analysis statistics. Omoroguiwa (2010); Eaves and Erford (2009); Zurawski (1999); Mitra et al (2009); Hingorjo and Jaleel (2012) defined item difficulty index as a measure of the proportion of examinees who answered the item correctly. For this reason, it is frequently called the p-value. Item difficulty can range between 0.0 and 1.0, with a higher value indicating that a greater proportion of examinees responded to the item correctly, and it was thus an easier item. The item difficulty statistic is an appropriate choice for achievement or aptitude tests when the items are scored dichotomously (i.e., correct vs. incorrect). Thus, it can be derived for true-false, multiple-choice, and matching items, and even for essay items, where the instructor can convert the range of possible point values into the categories “passing” and “failing.”

Eaves and Erford (2009), it is desirable for a test to contain items of various difficulty levels in order to distinguish between students who are not prepared at all, students who are fairly prepared, and students who are well prepared. In other words, educators do not want the same level of success for those students who did not study as for those who studied a fair amount and those who studied exceptionally hard. Therefore, it is necessary for a test to be composed of items of varying levels of difficulty.

Matlock-Hetzel (1997) identified two indices that can be computed to determine the discriminating power of an item, the item discrimination index, D, and discrimination
coefficients. She stated that in computing the discrimination index, D, first score each student's test and rank order the test scores. Next, the 27% of the students at the top and the 27% at the bottom are separated for the analysis. Justifying the use of 27%, Wiersma and Jurs (1990) stated that "27% is used because it has shown that this value will maximize differences in normal distributions while providing enough cases for analysis. Two indicators of the item's discrimination effectiveness, according to Matlock-Hetzel (1997) are point biserial correlation and biserial correlation coefficient. She stressed that the choice of correlation depends upon what kind of question we want to answer. The advantage of using discrimination coefficients over the discrimination index (D), according to her, is that every person taking the test is used to compute the discrimination coefficients and only 54% (27% upper + 27% lower) are used to compute the discrimination index, D.

Methodology
The students in Upper Basic 3 in Benin Metropolis were used for the study. Benin Metropolis is made up of three Local Government areas. Five Junior Secondary Schools were randomly selected from each of the three Local Government Areas. Forty students who were preparing for 2013 Upper Basic final examinations were randomly selected from each of the fifteen Junior Secondary Schools selected. A total of six hundred Upper basic Students were used for the study. The instrument used was the 2010 Upper Basic Mathematics objective examination conducted by National Examination Council (NECO). The instrument was made up of eighty (80) objective multiple choice questions with four options A to D. The purpose of using this instrument was that it was the first certificate examination that would be conducted under Universal Basic Education (UBE) programme in Nigeria. The instrument was administered for two hours as stipulated on the question paper with the help of one mathematics teacher in each of the selected schools. The scripts were collected immediately the two hours elapsed and the responses of the students were scored dichotomously by assigning 1 for the correct response and 0 for the wrong response.

Item Difficulty Index using the entire number of examinees was computed using the formula:

\[ p = \frac{R}{T} \]

where \( R \) is the number of the examinees that got the item correctly and \( T \) is the total number of the examinees.

To compute the Item Difficulty Index using percentage sample, the scripts were arranged in descending order of the performance of the examinees and the first 27% of the scripts called the upper group U and the last 27% of the scripts called the lower group L were taken and the formula:

\[ p = \frac{R_U + R_L}{N_U + N_L} \]

Where \( p \) is the difficulty index, \( R_U \) was the number of Examinees who got the item correctly in the upper group, \( R_L \) was the number of examinees in the lower group who got the item correctly and \( N_U \) and \( N_L \) were the number of examinees in the upper and lower groups respectively. Microsoft Office Excel was used to execute the computations. Pearson correlation coefficient was computed using SPSS version 12 to determine the relationship between the \( p \)-value when it was computed using the entire examinees and when computed using percentage sample of the examinees. The difference between the means of the two \( p \)-values was established by conducting pair group t-test of difference between means using SPSS version 12. For the purpose of this study and in accordance with the previous literature, all the calculated item difficulty indices that fall within the range of 0.3 to 0.7 are taken to be appropriate. All the items with the item difficulty that are less than 0.3 are regarded to be too
difficulty while all the items with item difficulty greater than 0.7 are regarded to be too simple

Results and Discussion

Results

Research Question 1: Do all the items of the 2010 Upper Basic Mathematics objective Certificate Examination fall within the acceptable item difficulty index range of 0.3 to 0.7?

Table 1: Item Difficulty Indices (p-values) Analysis

<table>
<thead>
<tr>
<th>ITEM</th>
<th>p-value using the entire examinees</th>
<th>p-value using percentage sample of examinees</th>
<th>ITEM</th>
<th>p-value using the entire examinees</th>
<th>p-value using percentage sample of examinees</th>
<th>ITEM</th>
<th>p-value using the entire examinees</th>
<th>p-value using percentage sample of examinees</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1</td>
<td>0.7</td>
<td>0.69</td>
<td>Q29</td>
<td>0.52</td>
<td>0.62</td>
<td>Q57</td>
<td>0.41</td>
<td>0.43</td>
</tr>
<tr>
<td>Q2</td>
<td>0.61</td>
<td>0.62</td>
<td>Q30</td>
<td>0.34</td>
<td>0.36</td>
<td>Q58</td>
<td>0.45</td>
<td>0.53</td>
</tr>
<tr>
<td>Q3</td>
<td>0.47</td>
<td>0.52</td>
<td>Q31</td>
<td>0.45</td>
<td>0.49</td>
<td>Q59</td>
<td>0.45</td>
<td>0.52</td>
</tr>
<tr>
<td>Q4</td>
<td>0.57</td>
<td>0.61</td>
<td>Q32</td>
<td>0.43</td>
<td>0.5</td>
<td>Q60</td>
<td>0.4</td>
<td>0.48</td>
</tr>
<tr>
<td>Q5</td>
<td>0.6</td>
<td>0.62</td>
<td>Q33</td>
<td>0.31</td>
<td>0.27</td>
<td>Q61</td>
<td>0.41</td>
<td>0.45</td>
</tr>
<tr>
<td>Q6</td>
<td>0.6</td>
<td>0.69</td>
<td>Q34</td>
<td>*0.25</td>
<td>*0.26</td>
<td>Q62</td>
<td>0.39</td>
<td>0.45</td>
</tr>
<tr>
<td>Q7</td>
<td>0.54</td>
<td>0.58</td>
<td>Q35</td>
<td>0.45</td>
<td>0.52</td>
<td>Q63</td>
<td>0.44</td>
<td>0.51</td>
</tr>
<tr>
<td>Q8</td>
<td>0.5</td>
<td>0.5</td>
<td>Q36</td>
<td>0.43</td>
<td>0.48</td>
<td>Q64</td>
<td>0.46</td>
<td>0.51</td>
</tr>
<tr>
<td>Q9</td>
<td>0.53</td>
<td>0.57</td>
<td>Q37</td>
<td>0.43</td>
<td>0.45</td>
<td>Q65</td>
<td>0.42</td>
<td>0.51</td>
</tr>
<tr>
<td>Q10</td>
<td>0.6</td>
<td>0.65</td>
<td>Q38</td>
<td>0.3</td>
<td>0.6</td>
<td>Q66</td>
<td>0.43</td>
<td>0.49</td>
</tr>
<tr>
<td>Q11</td>
<td>0.38</td>
<td>0.46</td>
<td>Q39</td>
<td>0.43</td>
<td>0.48</td>
<td>Q67</td>
<td>0.46</td>
<td>0.55</td>
</tr>
<tr>
<td>Q12</td>
<td>0.37</td>
<td>0.44</td>
<td>Q40</td>
<td>0.34</td>
<td>0.35</td>
<td>Q68</td>
<td>0.46</td>
<td>0.56</td>
</tr>
<tr>
<td>Q13</td>
<td>0.46</td>
<td>0.47</td>
<td>Q41</td>
<td>*0.16</td>
<td>*0.26</td>
<td>Q69</td>
<td>0.53</td>
<td>0.61</td>
</tr>
<tr>
<td>Q14</td>
<td>0.45</td>
<td>0.55</td>
<td>Q42</td>
<td>0.41</td>
<td>0.47</td>
<td>Q70</td>
<td>0.46</td>
<td>0.52</td>
</tr>
<tr>
<td>Q15</td>
<td>0.44</td>
<td>0.49</td>
<td>Q43</td>
<td>0.44</td>
<td>0.5</td>
<td>Q71</td>
<td>0.43</td>
<td>0.48</td>
</tr>
<tr>
<td>Q16</td>
<td>0.46</td>
<td>0.54</td>
<td>Q44</td>
<td>0.47</td>
<td>0.6</td>
<td>Q72</td>
<td>0.41</td>
<td>0.49</td>
</tr>
<tr>
<td>Q17</td>
<td>0.43</td>
<td>0.46</td>
<td>Q45</td>
<td>0.42</td>
<td>0.47</td>
<td>Q73</td>
<td>0.4</td>
<td>0.5</td>
</tr>
<tr>
<td>Q18</td>
<td>0.33</td>
<td>0.39</td>
<td>Q46</td>
<td>0.38</td>
<td>0.3</td>
<td>Q74</td>
<td>*0.1</td>
<td>*0.25</td>
</tr>
<tr>
<td>Q19</td>
<td>0.45</td>
<td>0.52</td>
<td>Q47</td>
<td>0.44</td>
<td>0.52</td>
<td>Q75</td>
<td>0.46</td>
<td>0.56</td>
</tr>
<tr>
<td>Q20</td>
<td>0.53</td>
<td>0.6</td>
<td>Q48</td>
<td>0.37</td>
<td>0.43</td>
<td>Q76</td>
<td>0.42</td>
<td>0.48</td>
</tr>
<tr>
<td>Q21</td>
<td>0.46</td>
<td>0.6</td>
<td>Q49</td>
<td>0.48</td>
<td>0.62</td>
<td>Q77</td>
<td>0.43</td>
<td>0.54</td>
</tr>
<tr>
<td>Q22</td>
<td>0.48</td>
<td>0.56</td>
<td>Q50</td>
<td>0.39</td>
<td>0.48</td>
<td>Q78</td>
<td>0.41</td>
<td>0.47</td>
</tr>
<tr>
<td>Q23</td>
<td>0.44</td>
<td>0.51</td>
<td>Q51</td>
<td>0.42</td>
<td>0.48</td>
<td>Q79</td>
<td>0.4</td>
<td>0.48</td>
</tr>
<tr>
<td>Q24</td>
<td>0.47</td>
<td>0.56</td>
<td>Q52</td>
<td>0.43</td>
<td>0.46</td>
<td>Q80</td>
<td>0.49</td>
<td>0.57</td>
</tr>
<tr>
<td>Q25</td>
<td>0.41</td>
<td>0.52</td>
<td>Q53</td>
<td>0.44</td>
<td>0.46</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q26</td>
<td>0.47</td>
<td>0.53</td>
<td>Q54</td>
<td>0.48</td>
<td>0.6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q27</td>
<td>0.48</td>
<td>0.52</td>
<td>Q55</td>
<td>0.47</td>
<td>0.52</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q28</td>
<td>0.48</td>
<td>0.54</td>
<td>Q56</td>
<td>0.41</td>
<td>0.43</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Did not meet the appropriate range of values.
The difficulties index categories were set less than 0.30, 0.30 to 0.70 and above 0.70. Table 1 revealed that only three items representing 3.75% had item difficulty index less than 0.30, the remaining 77 items representing 96.25%, had item difficulty index within the appropriate range of 0.30 to 0.70 when the entire number of examinees were used and when percentage sample of examinees were used. The three items are items 34, 41 and 74 with the item difficulty indices of 0.25, 0.16 and 0.10 respectively when the entire examinees were used and 0.26, 0.26 and 0.25 when percentage sample of examinees were used. The implication of these results is that the three items that is items 34, 41 and 74 were too difficult for the examinees. Less than 30% of the examinees were able to get the items correctly.

**Research Question 2:** Is there a relationship between the item difficulty indices obtained by using the entire examinees responses and the one obtained by using percentage sample of the examinees?

**Figure 1:** The Scattered Diagram of the Two Methods

Figure 1 revealed a linear relationship between the p-value using the entire examinees response and the p-value using percentage sample of the examinees responses. The scattered diagram showed a relationship that slopes downward from the right to the left. The implication of this result is that there is a positive relationship between the item difficult indices when the entire examinees responses are used and when percentage sample of examinees responses are used.

**Hypothesis 1:** There is no significant relationship between the item difficulty index obtained by using the entire examinees responses and that obtained by using percentage sample of the examinees.
Table 2: Correlation Analysis

<table>
<thead>
<tr>
<th>Pair 1</th>
<th>N</th>
<th>Correlation</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>p-value using the entire examinees &amp; p-value using percentage sample of examinees</td>
<td>80</td>
<td>.860</td>
<td>.000</td>
</tr>
</tbody>
</table>

Table 2 revealed that the Pearson correlation coefficient is .860 and it is significant at p<.05. Hence the null hypothesis is not retained, there is therefore a significant relationship between the item difficulty index obtained using the entire examinees responses and that obtained using percentage sample of the examinees responses.

Hypothesis 2: There is no significant difference between the item difficulty index obtained by using the entire examinees responses and that obtained by using percentage sample of the examinees.

Table 3: Pair sample t-test of the difference between the mean t-valued of the two methods.

<table>
<thead>
<tr>
<th>Pair 1</th>
<th>Mean</th>
<th>N</th>
<th>Std. Deviation</th>
<th>df</th>
<th>t-value</th>
<th>Sig (2-tailed)</th>
</tr>
</thead>
<tbody>
<tr>
<td>p-value using the entire examinees</td>
<td>.439</td>
<td>80</td>
<td>.0862</td>
<td>79</td>
<td>-12.226</td>
<td>.000</td>
</tr>
<tr>
<td>p-value using percentage sample of examinees</td>
<td>.503</td>
<td>80</td>
<td>.0897</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

P<.05

Table 3 revealed that the mean of the p-values when the entire examinees responses were used was 0.439 and the standard deviation was 0.0862. The mean of the p-value when the percentage sample of the examinees responses was used was 0.503 and the standard deviation was 0.0897. The calculated t-value is 12.226 and it is significant at p<.05. hence hypothesis 2 is retained. Therefore there is no significant difference between the item difficulty index obtained by using the entire examinees responses and that obtained by using percentage sample of the examinees.

Discussion

The results in table 1 showed that three items representing 3.75% had item difficulty index less than 0.30, the remaining 77 items’ representing 96.25%, had item difficulty index within the appropriate range of 0.30 to 0.70 when the entire number of examinees were used and when percentage sample of examinees were used. This result is similar to the one obtained by (Hingorjo & Jaleel 2012). Hingorjo and Jaleel (2012) analysed a 50 items one-best MCQs having 5 options each. They found that 64% of the items have the difficulty indices ranging from 0.3-0.7. Sayyah et al (2012 also carried out An Item Analysis of Written Multiple-Choice Questions: Kashan University of Medical Sciences and found that 17.7 percent of exams in the faculty of nursing and midwifery had item difficulty less than 0.30 and 25.9 percent of the exams had item’s difficulty over 0.70 in nursing and midwifery faculty. They concluded that that the difficulty index, discrimination index and the Alpha-Cronbach values
in the faculty of nursing and midwifery were within the acceptable range recommended by experts in the field of educational measurement. Boopathiraj and Chellamani (2013) also found that thirteen items out of 60 (21%) were rejected either due to difficulty level or discrimination Index, thirty five items (58%) were accepted without revision while 12 items were accepted provided that necessary revision.

Figure 1 revealed a linear relationship between the items difficulties indices when the entire examinees responses were used and when percentage sample of examinees were used, the scatter plot slopes downward from the right to the left indicating positive relationship. Table 2 showed that the correlation coefficient of 0.860 and it is significant at p<.05. The implication of these results is that the item difficulty index when the entire examinees responses were used was strongly and positively correlated with the item difficulty index when the percentage samples of examinees were used. In other words the high item difficulty index when the entire examinees responses were used was associated with the high value of the item difficulty index when percentage sample of examinees were used while the low item difficulty index when the entire examinees responses were used was associated with low value of the item difficulty index when percentage sample of examinees were used. It therefore suffices to say that the two methods of computing the item difficulty index are comparable.

Table 3 showed that the mean values of the item difficulty indices when the entire examinees and when the percentage sample of examinees were used were respectively 0.439 and 0.503. The standard deviations were respectively 0.0862 and 0.0897. The two mean values of 0.439 and 0.503 are within the range of values of 0.30 and 0.70 suggested by Nelson 2001, Hingorjo and Jaleel 2012 and Sayyah et al 2012. However 3.75% of the test items showed item difficulties less than 0.30. According to Hingorjo and Jaleel (2012) Items having p-values below 30% and above 70% are considered difficult and easy items respectively. This condition indicates that some of the test items needed to be re-evaluated. When an item difficulty approaches a value less than 0.30 such as those found in this research, it implies that either the mathematics teacher did not cover the subject matter thoroughly or the student did not show enough interest to study it well. The table also showed that the t-calculated was -12.226 and it was significant at p<.05. The implication of these results is that there is no difference between the item difficulty indices obtained when the entire examinees’ responses are used to compute the item difficulty index and when percentage sample of examinees are used. It therefore suffices to say that using either the entire examinees responses or percentage sample of examinees to compute the item difficulty index of a test will give similar result for the same item.

Conclusion
The results of the study revealed that more than ninety five percent of the 2010 Upper Basic Certificate Mathematics Objective Test had the item difficulties that were between the recommended range of 0.30 and 0.70. It can therefore be concluded that most of the items of the examination were appropriate for the examinees. The results also revealed that the item difficulty indices obtained by using the two methods are strongly and positively correlated. It can therefore be concluded that the two methods of computing the item difficulty index are comparable. The t-test analysis of the means of the item difficulty indices obtained using the two methods showed that the means obtained are both within the range of 0.30 and 0.70 and the difference between the means is not significant at p<.05. It can therefore be concluded...
that item difficulty index can be computed using either the entire examinees responses or percentage sample of examinees.

**Recommendations**
The results obtained from this study have implications for Upper Basic Mathematics teachers, test developers and the students. It is expected that the Upper Basic Certificate objective Mathematics Examination items should cover the mathematics curriculum content for Upper Basic one, two and three. It is therefore recommended that all mathematics teachers of Upper Basic one, two and three should make every effort to expose the students to all the mathematics curriculum content. This can be done by organising extra-moral classes after the normal school hours as the periods allotted for mathematics during the official hours of schooling may not be enough to cover all the topics. Mathematics teachers are ones who implement the mathematics curriculum hence it is recommended that these teachers be involved in the preparation of the Upper Basic certificate mathematics examination questions. This will help in producing question bank where the examination body (NECO) can select to administer to the candidates.
The test developers should ensure that the mathematics objective questions to be administered are within the Upper Basic mathematics curriculum content. This can be done by preparing a table of specification before the writing of the questions. The items in the test should neither be too easy nor too difficult; hence a balance between these two must be maintained. It is therefore recommended that there should be trial testing and comprehensive item analysis should be carried out before the final items to be administered are selected. It is also recommended that test items should be arranged with easy items at the beginning of the test, leading to the more difficult items towards the end of the test. This will serve as a motivating factor, a factor which encourages students to continue with a test which they might otherwise find difficult. To reduce cost and time wastage, it is recommended that percentage sample of examinees should be used by test developers to calculate the item difficulty index of the test items.

**References**


Krishnan, V. (2013). The early child development instrument (edi): an item analysis using classical test theory (CTT) on Alberta’s data; edmonton, early child development mapping (ecmap) project, Alberta.


