Abstract: - The traditional process of scale anchoring only set several specific scores as anchoring points. It only causes ambiguous grading results. Furthermore, the skills and abilities used in solving tests, the applied indicators are abstract and obscure. Thus, in this paper, a smoothing method of arbitrary anchoring revised from scale anchoring is proposed to make tests more accurate in distinguishing test levels. It reflects the difficulty of test scores by the sum and hence easily shows the test levels. In addition, a library containing test items is set up to be shared and reused for users. The paper comes out with two conclusions: smoothing method of arbitrary anchoring can represent the levels of test items by all possible score points. Its scale anchoring could indicate exam’s difficulty as well as the tester’s ability of passing the test; the test items library could provide references for teachers doing proposition and give students a better adaptive learning environment.

Key-Words: - Scale anchoring, Smoothing method of arbitrary anchoring, Test Items

1 Introduction
Test items are the backbone of examination and they play a crucial role on whether test can function well. That is, without good test items, even though tests have ideal examination patterns, or excellent system designs, the results are still suspect. To understand the test quality, item analysis is required and quantitative analysis can be used to judge the test quality. In quantity aspect, some pointers such as difficulty, discriminator, and indicators will be calculated by the testers’ responses. But these pointers all focus on the properties of test items, and the relationship of testers and these test items did not consider. Thus, we cannot judge the testers’ ability from their scores.

In a Trend of International Mathematics and Science Study (TIMSS) report, the scale anchoring is usually used to present the relations between the test items and testers’ scores. Further, they discriminate on tests according to difficulty of different levels. They simultaneously discuss the skills and the ability to solve different levels of tests. That is, it explains the meaning of scores based on the content reference viewpoint and fully understands the significance of tests and its relation with tester. Therefore, one of the motives of this study is to decide the anchoring scale in analyzing the test and establish one indicator. It should be able to reflect the test and the tester while inquire into the examination function of the tests.

After the test items have been analyzed and showed the important characteristic material, to make the test items be able to reuse and share, test items library is needed and that is the other motives of the study.

2 Related Work

2.1 Test Question Analysis
American instructional evaluator Kibler[3] proposed a basic model to dividing instructional processes into goals, pre-assessment, activity, evaluation. In this pattern, the teacher uses the tool (usually is examination) to evaluate students’ performance and improve their teaching and modify the study goal by...
the feedback functions. Thus, the good test items are major resources for instruction.

2.2 Quantity Analysis
The most basic unit is the test items which constitute an examination. Good test items makes good examinations. In most of the schools nearly all examination are achievement by tests. Thus, carrying on a qualitative analysis in the contents of test items and carrying on a quantitative analysis of the statistical characteristics of test items can screen good test items.

In quantitative analysis, no matter if it is the classical test theory (CTT) or the item response theory (IRT) the calculated index of correlation, all based on the tester’s response to calculate the difficulty, discriminator, or indicators for guiding answers. These quantified data represent characteristics of tests, and do not consider test items or its relation with tester. It implies that these test items are unable to present the difference among the testers as well as unable to understand which test items can examine tester and to what degree. In order to solve this problem, we may analyze the test content and clarify the ability and concept which the test item to examine. To understand respondent’s ability and its relation with the score (total score range) is the major function of anchoring scale.

2.3 Test Items Response Theory
The classical test theory (CTT) locates tester’s ability based on score. Passing rate (of all items) indicates the test difficulty and is a basic method to calculate raw score. If the exam is simple, scores may be high; or it may be low and difficult. There is no standard for the tests, the exam characteristics determines the score; this is called “exam dependence” (or test dependant).

Item response theory (IRT) is also called as “latent trait theory”. It assumes tester is under the single characteristic or specialty responses could be formulated into mathematical function. While testing, the test difficulty and the ability of the tester won’t be influenced by sample or test items.

Item response theory (IRT) takes probability of the tester’s passing rate as an ordinate axis and the ability of the testers as the abscissa axis then draws the curve, which comes out as item characteristic curve (ICC). According to the test items’ characteristic curve in ICC parameters it may divide the IRT pattern into the single parameter logistic model (1PLM), the double parameters logistic pattern (2PLM) and three parameters logistic pattern (3PLM).

Two parameters and three parameters patterns are generally called multi-parameter model. The single parameter pattern is Rasch pattern. Wang[6] suggested that Rasch is the right pattern for measurement, two or three parameters patterns is a statistical pattern; it does not have the significant meaning of measurement. Based on Wang’s study, this research will apply the Rasch pattern to analyze test results, estimate the test item difficulty, and the ability value of tester’s passing probability characteristics.

2.4 Rasch Dichotomous Scoring Pattern
Let:
\[ n: \] the n\textsuperscript{th} tester,
\[ i: \] the i\textsuperscript{th} test item,
\[ B_n: \] present the latent characteristics or ability,
\[ D_i: \] present the test characteristics or difficulty,
\[ P_{ni}: \] present the passing probability of n\textsuperscript{th} tester to i\textsuperscript{th} item,

According to Rasch[5] in dichotomous score pattern, the inferential reasoning may refer to formula 1:

\[
P_{ni} = \frac{\exp(B_n - D_i)}{1 + \exp(B_n - D_i)}
\]

(1)

In the formula exp represents the exponential function, Bn and Di are logarithmic numbers. Formula 1, indicates the passing probability of a tester, once we get tester’s ability Bn and test item difficulty Di. For example: one tester’s abilities is +1 logit, test item difficulty - 2 logit, then the passing probability equal to

\[
P_{ni} = \frac{\exp(1 - (-2))}{1 + \exp(1 - (-2))} = 0.95
\]

(2)

In Rasch pattern, take the passing probability as the ordinate axis, take the Bn-Di value as the abscissa axis, formula 1 may draw the item characteristic curve (ICC) as shown in Fig. 1, this curve succinctly expresses the changeable relation between test item parameter and the ability clearly.
This research applied anchoring scale based on Rasch pattern for calculation. Then used the Rasch analysis to measure item difficultness and tester’s ability value. After that study we use the smoothing procedures of anchoring scale, to obtain quantitative indicator of a test item.

2.5 The Anchoring Scale
Traditionally, the test analysis based on tester’s response to calculates the test item difficultness, discriminator, and indicators for guiding answers. These quantitative data represent the test item’s characteristics and do not reflect its relation with the score. Through this type of analysis we can hardly know what test items could apply to testers or could investigate what type of ability. Beaton & Allen [1] proposed the anchoring scale. It chooses several points as the anchoring spot, and use statistical analysis to select test items as anchoring items. Testers have higher passing scores in these anchoring items if they reach the anchoring point; otherwise they have less chance to pass.

The anchoring scale has two types of calculation: direct method and the smoothing method. The difference between two methods lies in the calculation process of correctness. Although the direct method is simple in small sample calculating, it causes major errors. Hence, this study favors the smoothing procedures. In the following, a smoothing method of arbitrary anchoring revised from scale anchoring is proposed to make tests more accurate in distinguishing test levels.

2.6 Threshold of Passing Rate
In the anchoring scale, no matter it is direct method or the smoothing procedures they both compute the test items to determine whether it belongs to the level of correct threshold, it means students who reach this score have the higher probability to pass this test question, this threshold is an important indicator as well as a meaningful factor to all learners. Dana [2] organized several different choices of standards. This study decides that both methods involved in the anchoring test item, its passing threshold standard set to 65% after referring various ranges of the above standards.

3 Research Methodology
This study uses empirical data analysis together with content analysis. The test items go through item analysis before being digitized and stored to make sure it obtains test item characteristics. Then induct the Wiki system[4] and apply knowledge organization, to establish test items library.

The flow chart of this study is shown in Fig. 2. The first stage is the test item analysis, mainly to induct the anchoring scale. It carries out the quantity analysis of the test items and produces the characteristic data. The second stage is the test items library, mainly to analyze the characteristic of the test item. The Wiki system helps with digitization and classification before storage as well as achieves the goal of re-use, and share.

3.1 Target Population
This study obtains data that originates from a committee (junior high student basic ability measurement working committee). It provides exam score from Nature Science in 2001 as the first examination and 2005 as the second exam. Altogether there are ten times of the Nature Science scores accompanied with the statistical data. Each time includes 5000 stochastic response patterns, and has been transformed to dichotomous scoring pattern (right answer 1, wrong answer 0). In the basic measurement test the examination contains several subjects, such as: biology, physics and chemistry, geosciences, health education (for 2005 test not included). The examination pattern is four choices in a test item (only one correct answer).
Computer scans test papers for scoring. Various subjects cover different topics and have different number of test items. When analyzing test item content effects, this study takes physics and chemistry as the target subjects.

3.2 Tools for the Test Item Analysis
1. This study uses the WINSTEPS for computation, while carrying out the test item response theory using Rasch pattern analysis, to measure the test item difficultness and the tester’s ability value.
2. Microsoft EXCEL and VBA macro assembler (designed by the research team) was applied to calculate the anchoring scale level.

3.3 Test Item Analysis Sequence
The major sample of the study is getting from “junior high student basic ability measurement”. This study processes the student responses through the quantitative analysis and processing test item content through qualitative analysis. Figure 3 dashed lines square express flow chart of test item analysis sequence.

3.4 Quantitative Analysis
1. Organize and input response data: getting the answer sheet from “junior high student basic ability measurement” transforms to dichotomous scoring pattern (right answer 1, wrong answer 0), and reorganizes to the Rasch pattern for analysis.
2. Rasch pattern analysis: here applies the Rasch pattern to carry out the analysis of the antecedent item. WINSTEPS was applied to measure the difficultness and the tester’s ability value. These values will be quoted as the base numbers for calculate the anchoring scale.
3. The calculation of anchoring scale: this study improves the original anchoring scale by setting all scores as anchoring scale without setting extra anchoring points. This specific characteristic makes the test items evolve more discriminate levels and causes more precise results. This study called it “arbitrary anchoring smoothing method”.

The main step is as follows:
• Step 1. Decide the test item response and the scale relations of test item characteristic descriptive function.
  This study uses function of Rasch pattern to express the test item difficulty and the relations between testers’ ability.

\[
P_{nt} = \frac{\exp(B_n - D_i)}{1 + \exp(B_n - D_i)}
\]

n: the n\textsuperscript{th} tester, i: the i\textsuperscript{th} test item, B\textsubscript{n}: present the latent characteristics or ability, D\textsubscript{i}: present the test characteristics or difficultness, P\textsubscript{nt}: present the passing probability of n\textsuperscript{th} tester to i\textsuperscript{th} item.

When analyzing data from the “junior high student basic ability measurement”, the related scale refers to passing rate from basic ability measurement and scale score table.
• Step 2 Calculate the lowest ability value to pass the test item.
  To consider passing rate P and test item difficultness Di as known value, when calculating the tester’s ability value may through inferential reasoning as following:

\[
\begin{align*}
1 - P &= \exp(D_i - B_n) \\
\ln \frac{1 - P}{P} &= D_i - B_n
\end{align*}
\]

According to formula 2, replace passing rate threshold value P and various test items difficultness Di, then may obtain the lowest ability value needed to pass the item (expressed by). To contrast this score with the basic ability measurement scale may obtain the lowest passing number of test items and the
score satisfied. This is the score of scale level. Take Fig. 4 as example:

<table>
<thead>
<tr>
<th>Item No.</th>
<th>Ability</th>
<th>Scale score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-0.496898</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>-0.496898</td>
<td>6</td>
</tr>
<tr>
<td>3</td>
<td>-0.496898</td>
<td>6</td>
</tr>
<tr>
<td>4</td>
<td>-0.496898</td>
<td>6</td>
</tr>
<tr>
<td>5</td>
<td>-0.496898</td>
<td>6</td>
</tr>
<tr>
<td>6</td>
<td>-0.496898</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>-0.496898</td>
<td>6</td>
</tr>
<tr>
<td>8</td>
<td>-0.496898</td>
<td>6</td>
</tr>
<tr>
<td>9</td>
<td>-0.496898</td>
<td>6</td>
</tr>
<tr>
<td>10</td>
<td>-0.496898</td>
<td>6</td>
</tr>
</tbody>
</table>

Fig. 4 Scale level calculation steps for arbitrary points smoothing method

1. To obtain the topic \( B_i \) value - 0.490960792;
2. To found in the contrast scale the number larger than \( B_i \) and check the smallest ability value in the scale, its passing items is 22;
3. Its corresponding scale score 19, it is the scale level of this year.

Step 3. Create test item grading result and set the test item “anchoring scale indicator”.

According to the steps mentioned above, may obtain test item scale level in different years; take ten mean values as “the anchoring indicator”. The assigned level of the passing rate threshold is one of the most important attributes affecting result. Several different choices were discussed in the literature. This study refers to the TIMSS frequently use standard and decides the passing rate threshold is 65%.

4 Research Findings and Discussion

Study uses Rasch pattern to obtained test item difficulty and compares with the traditional smoothing anchoring scale procedures. Take 2001 the first “basic ability measurement” test item for example, Fig. 5 shows, the abscissa axis represents the year (2 digits) + semester (1 digit) + test item number (2 digits), left side the ordinate axis represents the test item difficulty which Rasch method measured (unit is logit), at the right flank the ordinate axis represents the calculated anchor point through smoothing procedures scale.

Above analysis shows:

1. The traditional anchoring scale using smoothing procedures results are shown in Fig. 5. When the anchoring indicator is big, usually its test item difficulty is higher. But in turn, when the test item difficulty is high, it’s anchoring scale will not necessarily be higher; this is the weakness of anchor point smoothing procedures it could not precisely present level grading in the test item.
2. Selecting an arbitrary point in the smoothing procedures, like Fig. 6 shows; detect that arbitrary anchoring method by smoothing procedures obtain the anchoring scale value quite consistent with the test item difficulty. Therefore, it can make more precise level discrimination for the test items.
3. Taking the anchor point by smoothing procedures, the amount of test items belonging to scale level 40 is 26 as shown in Table 1. These test items belong to 11 different levels if an arbitrary method of smoothing procedure is used. Therefore, it is obviously on test item level discrimination. The arbitrary point surpasses the anchor point smoothing procedures, as shown as Fig. 7.
### Table 1 the test items of scale level 40

<table>
<thead>
<tr>
<th>Item No</th>
<th>07</th>
<th>12</th>
<th>34</th>
<th>51</th>
<th>37</th>
<th>36</th>
<th>42</th>
<th>41</th>
<th>23</th>
</tr>
</thead>
<tbody>
<tr>
<td>Difficulty</td>
<td>0.06</td>
<td>0.01</td>
<td>-0.01</td>
<td>0.01</td>
<td>0.03</td>
<td>0.04</td>
<td>0.05</td>
<td>0.07</td>
<td>0.09</td>
</tr>
<tr>
<td>Index of scale anchoring</td>
<td>20</td>
<td>30</td>
<td>28</td>
<td>21</td>
<td>26</td>
<td>21</td>
<td>24</td>
<td>21</td>
<td>33</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Item No</th>
<th>21</th>
<th>55</th>
<th>35</th>
<th>44</th>
<th>43</th>
<th>47</th>
<th>22</th>
<th>49</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>Difficulty</td>
<td>0.1</td>
<td>0.25</td>
<td>0.30</td>
<td>0.45</td>
<td>0.44</td>
<td>0.55</td>
<td>0.63</td>
<td>0.64</td>
<td>0.45</td>
</tr>
<tr>
<td>Index of scale anchoring</td>
<td>30</td>
<td>33</td>
<td>34</td>
<td>35</td>
<td>36</td>
<td>37</td>
<td>37</td>
<td>37</td>
<td>37</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Item No</th>
<th>25</th>
<th>52</th>
<th>15</th>
<th>44</th>
<th>20</th>
<th>29</th>
<th>46</th>
<th>28</th>
</tr>
</thead>
<tbody>
<tr>
<td>Difficulty</td>
<td>0.71</td>
<td>0.73</td>
<td>0.75</td>
<td>0.75</td>
<td>0.77</td>
<td>0.82</td>
<td>0.85</td>
<td>0.9</td>
</tr>
<tr>
<td>Index of scale anchoring</td>
<td>30</td>
<td>38</td>
<td>38</td>
<td>38</td>
<td>38</td>
<td>39</td>
<td>39</td>
<td>40</td>
</tr>
</tbody>
</table>

The above result indicates that the arbitrary point smoothing procedures calculates indicators to reflect scores points of the test item difficulty on scale, therefore, may clearly demonstrate two-way meaning, namely, the test item difficulty and passing score level.

### 5 The Conclusion and the Prospective

In two types of the anchoring scale computed results, this study proposes the improvement method -- arbitrary point smoothing procedures, as good for the test item grading effect. The scale anchoring of the test item produces anchor indicators, it may represent the test item difficulty, and also may demonstrate tester reaches this score and how many probabilities to pass this item.

In addition, from annual statistics may find that majority of test items distributed in the middle of diverse difficulty. Thus, the test item of Junior High Base Exam wins good reputation. In test question grouping pattern, the scale anchoring analysis of this study may clearly reveal the proportion of test item difficulty, provides teachers important reference while proposition; it also may give students a direction in exam preparation.

The anchoring scale indicator used in different exam demand more test item choice for effective measurement purpose. For example: school mathematics competition, is to exam the ability of outstanding competitor, then may choose the scale level its indicator score 40 and above test items; if it is the formation evaluation, it shall target on students’ learning mastery condition, then may choose the indicator score 30 or below test items.

Due to sample limitations, this study is unable to discuss the high level test item or the different subject test item, this part needs further research.

### 6 Acknowledgement

This research is supported in part by the National Science Council of Republic of China under the contract number NSC96-2520-S-218-001.

### References:


