**The Use of Stocking-Lord and Haebara Methods in Horizontal Equating:**

**A Case of Indonesian Madrasah Competence Assessment**

**Kusaeri1, Ali Ridho2, Noor Wahyudi1**

Universitas Islam Negeri Sunan Ampel Surabaya Indonesia1

Universitas Islam Negeri Maulana Malik Ibrahim Malang, Indonesia2

kusaeri@uinsa.ac.id

**Abstract**

Indonesian Madrasah Competence Assessment (AKMI) is a national assessment implemented each year held by the Ministry of Religious Affairs. One of the uniqueness of the AKMI is the use of different tests every year. AKMI focuses on capturing the development of learning in Madrasa by comparing the test scores of the current year with the previous year. An equating process is crucial for valid results when comparing scores. Therefore, this research aims to (a) equate the scientific literacy assessment tools at AKMI in 2022 with 2023 and (b) evaluate the business process of developing AKMI scientific literacy instruments (along with the MSAT design), which has implications for the equating process. This study adopted a Non-Equivalent Anchor Test (NEAT) design because the two test sets were parallel years, and the participants were from a diverse population. The data is from the AKMI Science Literacy of the Ministry of Religious Affairs, with 303,987 participants in 2022 and 342,987 in 2023 from the Islamic elementary school level. A total of 674 scientific literacy instrument items in 2022 and 1,392 items in 2023, with 90 items used as anchor items. There are 3 stages of analysis: pre-equalization, equalization calibration, and post-equalization analysis. The results show that there are differences in item parameter estimation results between 2022 and 2023, where 2022 has a higher level of item difficulty. Furthermore, the Stocking-Lord and Haebara methods had proven to be effective and had produced estimates with minimal differences in the equating process. In addition, the anchor items used as the basis for the equating do not represent the items as a whole in the item pool. These findings indicate the need for firm, careful standardization based on psychometric principles of the process at AKMI, from developing items to assembling items, testing, determining anchor items, and assembling items in the MSAT application.

**Keywords**: Horizontal equating, AKMI, Stocking-Lord dan Haebara

**Introduction**

Assessment plays a very important role in education because it provides a valuable portrait of students (Ayanwale, 2023). It is also an effective instrument for educational change or reform (Alonzo et al., 2021; Looney, 2014). In this context, the concept of assessment-driven instruction (Fischer et al., 2023) is appropriate. This concept means that with good assessment leads to a good learning process (Park & Park, 2012). Educational reform is expected to occur from improving the learning process in the classroom (Supovitz, 2009).

Refers to the above arguments, in 2020, the Indonesian government reformed the national assessment from assessing the achievement of national standards to classroom assessment practices that are student learning progress-oriented (Aditomo et al., 2019). Similarly, the Ministry of Religious Affairs, as the institution that is responsible for madrasa education around Indonesia, supports this policy through the AKMI (Indonesian Madrasah Competency Assessment) program (Kusaeri, Dwisanti, et al., 2022). AKMI is an assessment designed to produce information that can be used to provide feedback to students. This feedback is a reference for teachers who are teaching the learning process in madrasas. With the treatment, student learning outcomes and literacy skills are expected to improve every year (Kusaeri, Yudha, et al., 2022).

AKMI uses several different test sets (measuring the same construct) every year. According to Wei (2013), standardization needs to be carried out using the equating process of the test score. By equating, scores from different years can be converted into parameter items on the same scale (M. Kolen & Brennan, 2014; Nisa & Retnawati, 2018), so the scores between test sets from different years are compared (M. Kolen & Brennan, 2014; Moghadamzadeh et al., 2011). The result reveals the relationship between the raw scores of two sets of tests from two parallel years so that the scores of the previous year can be compared with current year scores (Rodrigues et al., 2022; Stoolmiller et al., 2013). Thus, the result captures the impact of the madrasa teachers learning interventions during the year.

Researchers have been keen on equating test scores to large-scale data assessment in the past decade. As a matter of fact, Majoros et al. (2021) and Strietholt & Rosén (2016) use data from the IEA (International Association for the Evaluation of Educational) for mathematical literacy data. Similarly, equating test scores for the TIMSS and PIRLS assessment programs on reading and mathematics literacy between different countries (Khorramdel et al., 2022). Chmielewski (2019) and Majoros (2023) use equating test scores from regional, national, and international assessments over a long period. They generally apply anchor items for different test items using an IRT approach. However, from the studies above, few, if any, have revealed the equating process for preparing large-scale tests (large assessments) with various forms of questions using Multistage Adaptive Testing (MSAT), such as AKMI.

AKMI uses five kinds of questions (multiple choice, complex multiple choice, matching, true-false, and short answer). The challenge is how to choose anchor items for the equating process to represent item parameters (such as difficulty level) as well as the items in the question bank (M. J. Kolen & Brennan, 2014; Magis et al., 2017). In addition, Fink & Born (2018) propose that the content of the anchor item must be able to represent the items in the question bank. Indeed, this equating process is more complex, especially when administering AKMI using MSAT. This is the focus of this research, which has not been explored previously.

MSAT is an exam administration method that has recently become popular in assessment (Cai et al., 2021; Li et al., 2021; MacGregor et al., 2022; Shin et al., 2021). Test administration using the MSAT model can increase measurement efficiency (Berger et al., 2019). Each test taker will get different question items according to their abilities. In this way, they will get question items with a difficulty level that matches their abilities (Ersen & Lee, 2023). Furthermore, in large-scale assessments, this method can minimize fraud or cheating during the exam. The MSAT is more efficient in the number of items to estimate the test taker's ability, more precise measurement results (with smaller measurement error), and high predictive validity (MacGregor et al., 2022; Steinfeld & Robitzsch, 2021). Thus, with these advantages, such as effectiveness in accommodating a balance between content, difficulty level, and security, the AKMI development team in the Indonesian Ministry of Religious Affairs has implemented MSAT in madrasas.

There are two types of equating, horizontal and vertical (Ayanwale, 2023; van der Linden, 2000). Horizontal equating functions to equalize two scores from two different test sets but measure the same object of the research (Nisa & Retnawati, 2018). The purpose of horizontal equating is to compare two or more groups of test takers using two or more different test devices but measuring the same thing. On the other hand, vertical equating is equating on test instruments that have different levels of difficulty and different grade levels but measure the same thing. Vertical equating is used to reveal the development of students' abilities, even though these students are at different grade levels and have different ability levels, as long as the test equipment used measures the same thing (Cumming et al., 2020). In the AKMI context, the equating that is more suitable and needed in the field is horizontal equating. Because of horizontal equating, developments in learning outcomes and literacy skills can be detected between years. Therefore, this study will focus on the horizontal equating type.

In horizontal equating, several equating methods based on Item Response Theory (IRT) can be applied, such as Haebara, Stocking-Lord (SL), mean-mean, mean-sigma, and concurrent calibration (Rahmawati & Mardapi, 2015; Uysal & Kilmen, 2016). Several studies show that certain equating methods provide better results compared to others (Battauz, 2023; Kilmen & Demirtasli, 2012; Rahmawati & Mardapi, 2015; Setiawan, 2019). For example, Setiawan (2019) compared the Haebara method and mean-sigma in the 2018 national exam data, and the results show that the Haebara method is better than the mean-sigma method. In addition, Kilmen dan Demirtasli (2012) compared various IRT-based equating methods and revealed that the Haebara and SL methods were more precise with lower error rates. Both methods also have more moderate error estimates, and the results are more accurate when using more anchor items (Born et al., 2019). Similarly, Yusron, Retnawati, & Rafi (2020) show that the Haebara method can produce the smallest average RMSE compared to the mean-mean, mean-sigma, and Stocking Lord methods.

The studies above have shown that the Haebara and SL equating methods are more consistent and accurate, with smaller errors than other methods. However, there are still slight differences in results between the Haebara and SL methods from one researcher to another. This fact is certainly interesting to test and implement further the two methods in equating AKMI test scores in 2022 with 2023. Therefore, this research aims to (a) equate the scientific literacy test equipment at AKMI in 2022 with 2023. From this process, it is expected that AKMI 2022 and 2023 result scores can be compared well; (b) evaluate the AKMI instrument development business process (along with the MSAT design), which has implications for the AKMI test score equating process.

The findings of this research can significantly contribute to policymakers at the Ministry of Religious Affairs standardizing the processes at AKMI so they comply with existing stages and standards. Again, the process of developing items, assembling items, testing, and determining anchor items, as well as assembling items in the MSAT design, needs to be done carefully, precisely, and in line with psychometric scientific principles. As a result, ongoing processes (especially the equating process) can provide valid information so that the progress of AKMI results provides a complete picture of AKMI implementation. This valid information is very useful in providing treatment to intervene in the learning process in madrasas at the next stage, preparing or revising madrasa textbooks, and finding models for giving assignments, projects, or homework that suit students' needs. In the end, AKMI can be a diagnostic tool for the progress of madrasas in Indonesia.

**Methods**

There are three different stages for score equalization in this study. The process begins with equalizing test item parameters in each 2022 and 2023 administrative year, then continues with horizontal equalization across years. The initial and second phases used a simultaneous calibration approach across the test sets. The final phase consists of cross-year equalization by using Battauz’s framework (2017), starting with the determination of the conversion coefficient. The conversion coefficient is calculated by considering the item discrimination parameters and the difficulty level of the items involved. Next, these coefficients are applied to the scale equation according to the Haebara and SL equalization framework.

This study adopted a non-equivalent anchor test (NEAT) design, which is very important for correlating and equating between different years (2022 and 2023). This design was chosen because the test sets for both years were parallel, and the test participants came from a diverse population (from madrasas spread throughout Indonesia). With this design, an illustration of differences in test takers' abilities can be seen from the proportion of test takers who answered correctly on the anchor item. By referring to this proportion, differences in the difficulty level on unique questions in each test set can be adjusted.

The object of this study is the AKMI instrument set on scientific literacy for class V MI level in 2022 with 674 items and in 2023 with 1,392 items, along with data on overall participant responses to the test instrument set. Of the number of items in 2022 and 2023, 90 items function as joint items (anchors or common items) to facilitate the equating process between test sets for the two years. These shared items were strategically selected to assess constructs that were consistent across both years.

***Participants***

This research utilizes information obtained from the Science Literacy conducted by the Indonesian Ministry of Religious Affairs, spanning the years 2022 to 2023. Participants were selected using cluster sampling from all provinces in Indonesia. The number of students participating in 2022 was 303,987, while in 2023, there were 342,987 participants. They are in class V MI level, whose exams are held in October (the middle of the odd semester) every year. An overview of AKMI test participants in 2022 and 2023 is presented in Table 1.

**Table 1**. Descriptive Statistics of 2022 and 2023 AKMI Participants on Scientific Literacy

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Description | | 2022 (Sum) | 2022 (Percentage) | 2023 (Sum) | 2023 (Percentage) |
| Origin of the Region | Sumatera | 55.465 | 18,2% | 58.959 | 17,2% |
| Java | 208.717 | 68,7% | 233.233 | 68,0% |
| Bali | 799 | 0,3% | 2.564 | 0,7% |
| West Nusa Tenggara | 7.424 | 2,4% | 9.001 | 2,6% |
| East Nusa Tenggara | 1.606 | 0,5% | 1.910 | 0,6% |
| Borneo | 14.500 | 4,8% | 18.623 | 5,4% |
| Celebes | 11.936 | 3,9% | 13.804 | 4,0% |
| Maluku | 2.092 | 0,7% | 3.161 | 0,9% |
| Papua | 1.448 | 0,5% | 1732 | 0,5% |
| Total | | 303.987 | 100% | 342.987 | 100% |

***Instrument***

The instruments that are equated are scientific literacy instruments used in AKMI 2022 and 2023 by referring to the framework with the following link: [https://drive.google.com/file/d/1u4QwbsbZZ6mehm 4Q21el0zWvKHCxmXbh/](https://drive.google.com/file/d/1u4QwbsbZZ6mehm%204Q21el0zWvKHCxmXbh/) view?usp=drive\_link. The instrument consists of five variations of question format: multiple choice, complex multiple choice, matching, true-false, and short answer. These variations in question form are an integrative part of the stimulus (in the form of text or discourse), which is a characteristic of AKMI questions. This instrument has gone through a very strict development process, starting from writing, review (both internal and external), and readability testing (on students and teachers) to field trials. Through a rigorous process, it is hoped that the resulting literacy score will truly reflect the portrait of the test taker's abilities.

A diagram of a process

Description automatically generated

**Figure 1**. MSAT Design in AKMI

The selected items whose parameters have been known through a series of previous stages are assembled into a test set in the MSAT to be administered to test takers by considering the Test Information Function (TIF). Each test taker will go through 3 stages (stages) with different paths according to their respective abilities. At each stage, 10 (ten) items are tested with varying levels of difficulty, context, question form, and content. The context used in scientific literacy includes personal, local or national, and global, with various forms of multiple-choice questions, complex multiple-choice, matching, true-false, and short answers. Meanwhile, the content being worked on at AKMI 2022 and 2023 includes health and disease, natural resources, environmental quality, mitigation, and science and technology. Thus, each test taker will be faced with 30 question items according to their respective paths. Through this method, it is hoped that it can test the consistency of the test taker's thinking abilities and that the exam questions will not leak easily.

***Data Analysis***

With the help of the R program in the "mirt" and "plink" libraries, the Haebara and SL equating methods are implemented with the following stages. First, pre-equalization analysis. At this stage, initial item analysis uses the Item Response Theory (IRT) model, which is carried out separately for each year's data collection (2022 and 2023). This is done to assess the functionality of the test items and identify items that function well or not.

Second, equalization calibration. This stage is carried out using the IRT method, specifically the one-parameter IRT model (1-PL) for dichotomous responses and the generalized partial credit model (GPCM) for polytomous responses. In accordance with the AKMI mixed model approach, it is used to calibrate the parameters of the items on the same scale, with a focus on the 90 anchor items to establish the relationship between the two forms. Third, post-equalization analysis: the accuracy and fairness of the equalization results are evaluated to determine whether the scores from both forms are accurate or not so as not to harm either group. In this way, it is hoped that the results of the equating carried out will be able to provide a complete portrait of the condition of the test equipment and the abilities of the participants in 2022 and 2023.

**Results**

Information on the evaluation results of the distribution of difficulty of anchor items used by AKMI 2022 and 2023, along with their characteristics for the two years, is presented in Table 2. This evaluation refers to four statistical measures: Average (M), Standard Deviation (SD), Maximum (Max), and Minimum (Min), each of which provides a unique perspective on the difficulty level. The complete information on the anchor items, which is used as the basis for the analysis in this research, is presented in Appendix 1, while the visualization is described in Figure 1. This article refers to Ayanwale (2023) to describe the item difficulty level index. Very easy items are marked with a negative difficulty index (less than -2). On the other hand, an item is said to be very difficult if it has an index of more than +2.

From Table 2, the AKMI 2022 difficulty level has an average of 0.510, which shows that, on average, the questions are quite difficult. A standard deviation of 0.815 shows a fairly large variation in the difficulty level of the items, which indicates a mix of simple and complex items. This variability is further emphasized by the (maximum) difficulty level of 3.919, which implies the presence of very difficult questions. On the other hand, the difficulty level (minimum) at -0.921 indicates that there are items that may be considered less challenging or even too easy for participants.

**Table 2** Summary of Item Anchor Parameters (Difficulty Level Aspect)

AKMI Question in 2022 and 2023

|  |  |  |
| --- | --- | --- |
| **Detail/Aspect** | **AKMI in 2022** | **AKMI in 2023** |
| Average | 0.510 | 0.479 |
| Deviation Standard | 0.815 | 0.832 |
| Maximum | 3.919 | 5.061 |
| Minimum | -0.921 | -4.030 |

In the following year, AKMI 2023 showed a slightly lower average difficulty level at 0.479, indicating a slight decrease in the difficulty at the test level accumulated from the test items. However, the standard deviation increased to 0.832, reflecting a wider spread in item difficulty levels. This increase shows a greater difference between the easiest and most difficult items in 2023 compared to 2022. The (maximum) difficulty level increased significantly to 5.061, highlighting the introduction or increase in more complex questions. On the other hand, the difficulty level (minimum) decreased significantly to -4.030, which indicates that there are questions that are potentially much easier compared to the questions in the previous year.

The visualization in Figure 2 also supports the previous description. It is characterized by the difficulty level of the anchor items, which is not the same in the 2022 and 2023 implementations. This can be seen in the location of the difficulty levels in 2022 and 2023, which are above, right in the middle, and below the regression line. The majority of data points cluster around the ascending line of best fit, indicating a positive correlation. Items that are difficult in one year tend to maintain their difficulty level in the following year. Outliers, or data points that are far from the best match line, certainly require deeper examination from the participant's side, the implementation process in the field, or other aspects that could potentially cause this to happen.

A graph with black dots and a blue line

Description automatically generated

**Figure 2**. Distribution of Difficulty Level of Anchor Items 2022 and 2023

The slope of the line indicates that the relationship is fairly proportional, although not perfect, as there is marked variability around the line. The plot also suggests a potential ceiling effect for the most difficult questions, as indicated by the less dense clustering of points at the upper end of the 2023 axis. This could indicate that the upper range of question difficulty experiences more variation and a potential increase in difficulty from 2022 to 2023. Thus, the scatter plot reveals that although there is consistency in the level of item difficulty from year to year, certain items show striking changes.

In another aspect, the position of the item difficulty level of the anchor items, which are on the theta scale in the 2022 AKMI implementation, forms the Test Characteristic Curve (TCC) for 2022. Likewise, the anchor items will have a TCC in 2023, which is not the same as in 2022. Through SL or Haebara equalization, these two TCCs can be rescaled by using one as a reference, with the 2023 TCC being used as a reference. The 2022 TCC shifts are depicted in Figure 3 and provide a comprehensive visualization of the role of the equalization process in maintaining the comparability of scores between years.

|  |  |
| --- | --- |
|  |  |
| **a** | **b** |

**Figure 3**. TCC Anchor items before and after equalization

Figure 3a displays the expected TCC on the original scale (before equalization) for 2022 and 2023. The two curves are closely aligned, especially at the lower end of the ability spectrum, indicating that before equalization, the assessments were already relatively equal in terms of difficulty. Because the curves differ slightly at higher ability levels, this may reflect fluctuations in item difficulty that need to be addressed by equalization. Figure 3b represents the scaled scores (after equalization), showing the convergence of the curves across the ability spectrum. This shows that the equalization process has effectively adjusted the 2022 and 2023 scores, particularly across the average ability range, to ensure that the two forms of the test are comparable.

Finally, equalization through calibration of the two TCCs was carried out to ensure fairness of scores and comparability between the two forms of the test. Although the average difficulty level, as discussed previously, remains stable from 2022 to 2023, the equalization process has addressed differences that may arise from variations in item difficulty. Thus, it is able to maintain the integrity of assessment measurements across different administrations.

The equating coefficient at the next stage was obtained in both years, and the results are presented in Table 3. The value of 1 in both methods is due to the analysis process using the one-parameter IRT method (1-PL), which only takes into account the difficulty level (b value). The equalization process appears to have a greater effect on mid-range abilities, indicating that the original 2023 form of the test may be slightly easier for the average test taker compared to 2022. The scaled curve implies that after equalization, individual ability levels are expected to yield the same score regardless of the year the test was taken.

**Table 3** Test Sets Coefficient Equating in 2022 and 2023

|  |  |  |
| --- | --- | --- |
|  | **Stocking-Lord** | **Haebara** |
| **a** | 1 | 1 |
| **b** | -0.101 | -0.190 |
| **Mean** | 746 | 740 |

The results of equalizing the anchor items are implemented on all questions so that the expected true score E(τ) is obtained, namely the actual score if the student answers all the questions. Even though students do not answer all the questions, based on the overall characteristics of the questions in each implementation, E(τ) can still be estimated. Based on Table 3, the equating scores produced by both methods are close to the same, with an average E(τ) value of 746 for SL and an average of 740 for Haebara. As an illustration, some of the score conversions from 2022 to 2023 are described in Appendix 2. For example, an AKMI 2022 test taker who has θ = -3, then E(τ) is 45. If we look at the implementation scale of AKMI 2023 without equalization, we will get E(τ) = 90. However, if you apply the equalization coefficient, you will get E(τ) = 98 if you use the SL method and E(τ) = 104 if you use the Haebara method. From 90 to 98 or 104, this indicates an increasing E(τ).

**Discussion**

The results of the research show that the overall average difficulty level is relatively stable, but the range of difficulty levels will expand in 2023. The improvement of maximum and minimum scores in 2023 indicates diversification in the difficulty level of the questions, with more difficult and easier questions in the assessment. On the other hand, although the average difficulty level has been relatively consistent, there has been an addition of the difficulty range from 2022 to 2023. This change may imply a deliberate effort to fulfill a wider range of abilities or to introduce more variability in the questions tested based on this result. It is expected that diverse disparities in the abilities of madrasa students from all over Indonesia can be fully represented (Kusaeri & Aditomo, 2019; Umar et al., 2022).

The description above indicates how important the item parameters (in this context, the level of item difficulty) are in the equating process. However, the equating process in this research used the MSAT, where the difficulty level is a crucial issue. When designing the MSAT, the difficulty level information is vital as a basis for determining the stage that each student must pass to suit their abilities. Certainly, an initial stage is needed in the form of a field test in order to obtain information on item parameters in the form of the difficulty level (Widhiarso & Ridho, 2022). The problem is how to ensure that field testing participants are answering the questions as in a real test. Similarly, Steinfeld & Robitzsch (2021) are concerned about the large number of test participants who are often taking the test as if it is a real test. As a result, the information on the difficulty level of the items obtained from the field testing results will be biased and cannot fully describe the real abilities of the participants. In fact, Ersen & Lee (2023) suggest that the parameter estimation results must be as accurate as possible because they are very important in estimating the participants' actual abilities at a later stage (during the official exam). The results of estimating participants' abilities have implications for the equating process or results (Kilmen & Demirtasli, 2012).

The analysis above can be explained through the MSAT stage in Figure 1. At stage 1, all participants will receive scientific literacy questions with a medium item difficulty level. The score obtained at this stage will determine the path and stage that a participant will go through (Widhiarso & Ridho, 2022). If they succeed in exceeding the minimum score, they will get a question set that is relatively more difficult at stage 2. On the other hand, they will get easier question sets when they fail to exceed the minimum predetermined score. Similarly, the participant's success or failure in exceeding the cut-off score in stage 2 will affect the path they must take in stage 3. After they have completed all the test stages, their abilities will be estimated. Thus, at each stage, the role of item parameter estimation in the form of item difficulty level cannot be denied. Inaccuracy in the parameter estimation process will have a fatal impact on the participants' ability to estimate results, which ultimately affects the equating results.

The research data shows that there are differences in the results of item parameter estimation (on 90 anchor items) between 2022 and 2023. The items in 2022 are more difficult than in 2023. This fact is certainly interesting to reveal from various perspectives: trial participants, process implementation, and readiness of madrasas. In terms of trial participants, 2022 participants appear to be still experiencing shock due to the transition process from online to offline learning, so they need an adjustment process. After they had experienced online learning for more than 2 years, there seems to be a decline in various aspects such as learning motivation, learning outcomes, and learning effectiveness (Umar et al., 2022). In the first year of online learning, there was a learning loss of 10-20%; in the second year, the learning loss improved to 70-80% (Lestari et al., 2023). On the other hand, the 2023 participants have become accustomed to the offline learning process again. Thus, it is logical that the difficulty level of anchor items in 2022 is higher than in 2023.

From the aspect of the implementation process, the 2023 trial is better prepared than 2022. Madrasahs, the test sample, have been informed previously so that computer or laptop devices are well prepared (Kemenag, 2023). Meanwhile, the madrasah's readiness to take part in the AKMI stages in 2023 is better than the previous year. This is marked by the learning process, assignments, and practice questions given in madrasas, which are starting to use AKMI questions and their various forms. In this way, they have more adequate provisions compared to the 2022 participants. The 2022 participants are also facing AKMI model questions for the first time (questions that begin with a stimulus in the form of text, reading, or infographics), which tend to be long, and they have not encountered them during learning. In the classroom (Le Hebel et al., 2017). As a result, the 2022 participants were shocked and ended up having difficulty solving questions like this. Thus, this fact greatly influences the results of estimating item difficulty levels in 2022.

Furthermore, research data shows how the equating coefficients produced by both methods (both SL and Haebara) have produced similar estimations in the equating process. Previously, the researcher provided evidence that the two methods differ by only 6 points (98 for SL and 104 for Haebara). These results confirm the findings of previous research conducted by Kilmen & Demirtasli (2012) and Rahmawati & Mardapi (2015), which showed the effectiveness of using IRT-based equating methods such as SL and Haebara to compare test scores in various forms or test administrations. In addition, the research results show a strong correlation between the equating scores from the two methods. This result means that there is a reasonable level of agreement between the two methods. Therefore, the focus of this research is comparing the harmony between the two methods rather than their goodness of fit.

These results are in line with Setiawan (2019), who attempted to compare the Haebara and SL equating methods using National Examination results data in Indonesia in 2018, where the result shows that the Haebara method has a higher mean value compared to SL. Similarly, Lee & Ban (2009) use a random group design. They found that the Haebara method gave better results than the SL method. However, these findings are different from several other studies, such as (Aksekioglu, 2017) and Mutluer & Çakan (2023), who revealed that the SL method outperformed the Haebara method. Both research groups, either supporting or against the result of this current research, certainly enrich the scientific knowledge related to the SL and Haebara methods to help the equating process (Özdemir & Atar, 2022).

Based on the results of the analysis and discussion described above, the various differences in existing research have not been able to provide strong evidence that one equating method is better than another. Differences in findings between studies are very likely to occur as a result of differences in several things, such as the characteristics of test takers and the types of questions tested. Differences in findings caused by differences in the characteristics of test participants and the types of questions give opportunities for other researchers to explore the topic of equating methods. Thus, It should be emphasized that there is no single method that can be used for all conditions and provide the best results.

On the other hand, research data does not yet provide information on the distribution of the 90 anchor items used in the 2022 and 2023 equating process. That is, do the anchor items represent all the items used? Referring to the argument of Fink & Born (2018), the anchor items used in the equating process must represent the characteristics of the entire item, starting from the content, context, and form of the question. Kolen & Brennan (2014) also stated that the anchor item must be able to represent the statistical characteristics of the entire item, such as the distribution of difficulty levels. In the AKMI context, have the three contexts in scientific literacy (personal, local or national, and global) been accommodated in the anchor items? Including various forms of questions (multiple choice, complex multiple choice, matching, true-false, and short answers) as well as the content used (health and disease, natural resources, environmental quality, mitigation, and science and technology). These three aspects are an inherent part of scientific literacy that must be represented in anchor items.

Using more items as anchor items (taking into account various aspects) is very important to ensure the accuracy of the equating results. However, the use of a large number of anchor items may raise concerns regarding testing safety (Wang, 2013). Care and accuracy in the process of selecting anchor items is a necessity. The trial design and process of assembling items into the MSAT system needs to be done carefully. Careless trial design and assembly processes can result in the three aspects inherent in scientific literacy above potentially not all appearing. The seriousness of the test participants is another prerequisite that must be met in order to obtain true item characteristics. Without all of this, it is impossible to fulfill the conditions stated by Fink & Born (2018) and (M. Kolen & Brennan, 2014).

**Conclusion**

Some important points of this research are highlighted as follows. First, the item difficulty level parameter as a reference in the equating process in 2022 is more difficult than in 2023. However, there is greater variability in items between the easiest and most difficult questions in 2023 compared to 2022. This shows that the question design in 2023 had made adjustments with more variations in the difficulty level of the questions. Furthermore, both methods - SL and Haebara - have produced estimates that are similar in the equating process. In addition, there is a strong correlation between the equating scores from the two methods. Thus, it indicates a reasonable level of agreement between the two. Second, research data does not provide information about the distribution of anchor items used in the equating process. This means that information on the representation of all items used in Scientific Literacy, such as context, various forms of questions, and content, has not been concluded. These findings indicate the need for strict, careful standardization and following psychometric principles from item development, ordering items, testing, determining anchor items, and ordering items in the MSAT application.

**Acknowledgment**

This work is part of the research funded by UIN Sunan Ampel Surabaya under the programme of Research and Community Service Grant 2023 (The Rector's Decree Number 180 of 2023). The researchers are also grateful to the Ministry of Religious Affairs of the Republic of Indonesia, for provision access to publish the AKMI 2023 data.

**References**

Aditomo, A., Rahmawati, N., Felicia, N., Shihab, M., Psi, F., & Handayani, M. B. A. (2019). *Academic Study and Recommendations for National Assessment System Reform*. https://pusmendik.kemdikbud.go.id/pdf/file-137

Aksekioglu, B. (2017). *Comparison of Test Equating Methods Based on Item Response Theory: PISA 2021 Science Test Sample* [Akdeniz Universities]. https://acikbilim.yok.gov.tr/bitstream/handle/20.500.12812/40289/yokAcikBilim\_10138163.pdf?sequence=-1&isAllowed=y

Alonzo, D., Leverett, J., & Obsioma, E. (2021). Leading an Assessment Reform: Ensuring a Whole-School Approach for Decision-Making. *Frontiers in Education*, *6*. https://doi.org/10.3389/feduc.2021.631857

Ayanwale, M. A. (2023). Test score equating of multiple-choice mathematics items: techniques from characteristic curve of modern psychometric theory. *Discover Education*, *2*(1), 30. https://doi.org/10.1007/s44217-023-00052-z

Battauz, M. (2023). Testing for differences in chain equating. *Statistica Neerlandica*, *77*(2), 134–145. https://doi.org/10.1111/stan.12277

Berger, S., Verschoor, A. J., Eggen, T. J. H. M., & Moser, U. (2019). Improvement of Measurement Efficiency in Multistage Tests by Targeted Assignment. *Frontiers in Education*, *4*. https://doi.org/10.3389/feduc.2019.00001

Born, S., Fink, A., Spoden, C., & Frey, A. (2019). Evaluating Different Equating Setups in the Continuous Item Pool Calibration for Computerized Adaptive Testing. *Frontiers in Psychology*, *10*(JUN). https://doi.org/10.3389/fpsyg.2019.01277

Cai, L., Albano, A. D., & Roussos, L. A. (2021). An Investigation of Item Calibration Methods in Multistage Testing. *Measurement: Interdisciplinary Research and Perspectives*, *19*(3), 163–178. https://doi.org/10.1080/15366367.2021.1878778

Chmielewski, A. K. (2019). The Global Increase in the Socioeconomic Achievement Gap, 1964 to 2015. *American Sociological Review*, *84*(3), 517–544. https://doi.org/10.1177/0003122419847165

Cumming, J., Goldstein, H., & Hand, K. (2020). Enhanced use of educational accountability data to monitor educational progress of Australian students with focus on Indigenous students. *Educational Assessment, Evaluation and Accountability*, *32*(1), 29–51. https://doi.org/10.1007/s11092-019-09310-x

Ersen, R. K., & Lee, W. (2023). Pretest Item Calibration in Computerized Multistage Adaptive Testing. *Journal of Educational Measurement*, *60*(3), 379–401. https://doi.org/10.1111/jedm.12361

Fink, A., & Born, S. (2018). A Continuous Calibration Strategy for Computerized Adaptive Testing. *Psychological Test and Assessment Modeling*, *60*(3), 327–346. http://www.iacat.org/content/operational-cat-programs

Kemenag, R. (2023). Technical Report Asesmen Kompetensi Madrasah Indonesia (AKMI). In *Direktorat Kurikulum, Sarana, Kelembagaan, dan Kesiswaan Madrasah Direktorat Jenderal Pendidikan Islam* . Direktorat Kurikulum.

Khorramdel, L., Yin, L., Foy, P., Jung, J. Y., Bezirhan, U., & Davier, M. (2022). *Rosetta Stone analysis report: Establishing a concordance between PASEC and TIMSS/PIRLS*. TIMSS & PIRLS International Study Center. https://tcg.uis.unesco.org/wp-content/uploads/sites/4/2022/07/Rosetta-Stone\_PASEC\_Analysis-Report\_2022.pdf

Kilmen, S., & Demirtasli, N. (2012). Comparison of Test Equating Methods Based on Item Response Theory According to the Sample Size and Ability Distribution. *Procedia - Social and Behavioral Sciences*, *46*, 130–134. https://doi.org/10.1016/j.sbspro.2012.05.081

Kolen, M., & Brennan, R. (2014). Test equating, scaling, and linking. Methods and practices. 3rd revised ed. In *Test Equating, Scaling, and Linking: Methods and Practices: Third Edition*. https://doi.org/10.1007/978-1-4939-0317-7

Kolen, M. J., & Brennan, R. L. (2014). *Test Equating, Scaling, and Linking* (3rd ed.). Springer New York. https://doi.org/10.1007/978-1-4939-0317-7

Kusaeri, K., & Aditomo, A. (2019). Pedagogical Beliefs about Critical Thinking among Indonesian Mathematics Pre-service Teachers. *International Journal of Instruction*, *12*(1), 573–590. https://doi.org/10.29333/iji.2019.12137a

Kusaeri, K., Dwisanti, C., Yanti, A., & Ridho, A. (2022). Indonesian Madrasah Competency Assessment: Students’ numeracy based on age. *Beta: Jurnal Tadris Matematika*, *15*(2), 148–156. https://doi.org/10.20414/betajtm.v15i2.558

Kusaeri, K., Yudha, Y. H., Kadarisman, Y. P., & Hidayatullah, A. (2022). Do Instructional Practices by Madrasah Teachers Promote Numeracy? *International Conference on Madrasah Reform 2021 (ICMR 2021*, 1–5. https://doi.org/10.2991/assehr.k.220104.001

Le Hebel, F., Montpied, P., Tiberghien, A., & Fontanieu, V. (2017). Sources of difficulty in assessment: example of PISA science items. *International Journal of Science Education*, *39*(4), 468–487. https://doi.org/10.1080/09500693.2017.1294784

Lee, W. C., & Ban, J. C. (2009). A comparison of irt linking procedures. *Applied Measurement in Education*, *23*(1), 23–48. https://doi.org/10.1080/08957340903423537

Lestari, M., Johar, R., Mailizar, M., & Ridho, A. (2023). Measuring Learning Loss Due to Disruptions from COVID-19: Perspectives from the Concept of Fractions. *Jurnal Didaktik Matematika*, *10*(1), 131–151. https://doi.org/10.24815/jdm.v10i1.28580

Li, G., Cai, Y., Gao, X., Wang, D., & Tu, D. (2021). Automated Test Assembly for Multistage Testing With Cognitive Diagnosis. *Frontiers in Psychology*, *12*(1347). https://doi.org/10.3389/fpsyg.2021.509844

Looney, A. (2014). Assessment and the Reform of Education Systems. In C. Wyatt-Smith, V. Klenowski, & P. Colbert (Eds.), *Designing Assessment for Quality Learning: The Enabling Power of Assessment* (pp. 233–247). Springer Science Business Media. https://doi.org/10.1007/978-94-007-5902-2\_15

MacGregor, D., Yen, S. J., & Yu, X. (2022). Using Multistage Testing to Enhance Measurement of an English Language Proficiency Test. *Language Assessment Quarterly*, *19*(1), 54–75. https://doi.org/10.1080/15434303.2021.1988953

Magis, D., Yan, D., & von Davier, A. A. (2017). *Computerized Adaptive and Multistage Testing with R*. Springer International Publishing. https://doi.org/10.1007/978-3-319-69218-0

Majoros, E. (2023). Linking the first- and second-phase IEA studies on mathematics and science. *Large-Scale Assessments in Education*, *11*(1), 14. https://doi.org/10.1186/s40536-023-00162-y

Majoros, E., Rosén, M., Johansson, S., & Gustafsson, J.-E. (2021). Measures of long-term trends in mathematics: linking large-scale assessments over 50 years. *Educational Assessment, Evaluation and Accountability*, *33*(1), 71–103. https://doi.org/10.1007/s11092-021-09353-z

Moghadamzadeh, A., Salehi, K., & Khodaie, E. (2011). A comparison Method of Equating Classic and Item Response Theory (IRT): A Case of Iranian Study in the University Entrance Exam. *Procedia - Social and Behavioral Sciences*, *29*, 1368–1372. https://doi.org/10.1016/j.sbspro.2011.11.375

Mutluer, C., & Cakan, M. (2023). Comparison of Test Equating Methods Based on Classical Test Theory and Item Response Theory. *Journal of Uludag University Faculty of Education*, *36*(3), 866–906. https://doi.org/10.19171/uefad.1325587

Nisa, C., & Retnawati, H. (2018). Comparing the methods of vertical equating for the math learning achievement tests for junior high school students. *Research and Evaluation in Education*, *4*(2), 164–174. https://doi.org/10.21831/reid.v4i2.19291

Özdemir, G., & Atar, B. (2022). Investigation of the Missing Data Imputation Methods on Characteristic Curve Transformation Methods Used in Test Equating. *Journal of Measurement and Evaluation in Education and Psychology*, *13*(2), 105–116. https://doi.org/10.21031/epod.1029044

Park, J. S., & Park, J. H. (2012). The changes of assessment at middle school level in Korea. *ZDM*, *44*(2), 201–209. https://doi.org/10.1007/s11858-012-0408-z

Rahmawati, R., & Mardapi, D. (2015). Modified Robust Z method for equating and detecting item parameter drift. *Research and Evaluation in Education*, *1*(1), 100. https://doi.org/10.21831/reid.v1i1.4901

Rodrigues, B., Cadime, I., Freitas, T., Choupina, C., Baptista, A., Viana, F. L., & Ribeiro, I. (2022). Assessing oral reading fluency within and across grade levels: Development of equated test forms. *Behavior Research Methods*, *54*(6), 3043–3054. https://doi.org/10.3758/s13428-022-01806-7

Setiawan, R. (2019). A Comparison of Score Equating Conducted Using Haebara and Stocking Lord Method for Polytomous. *European Journal of Educational Research*, *8*(4), 1071–1079. https://doi.org/10.12973/eu-jer.8.4.1071

Shin, H. J., Yamamoto, K., Khorramdel, L., & Robin, F. (2021). *Increasing Measurement Precision of PISA Through Multistage Adaptive Testing* (pp. 325–334). Springer Proceeding in Mathematics & Statistics. https://doi.org/10.1007/978-3-030-74772-5\_29

Steinfeld, J., & Robitzsch, A. (2021). Item Parameter Estimation in Multistage Designs: A Comparison of Different Estimation Approaches for the Rasch Model. *Psych*, *3*(3), 279–307. https://doi.org/10.3390/psych3030022

Stoolmiller, M., Biancarosa, G., & Fien, H. (2013). Measurement Properties of DIBELS Oral Reading Fluency in Grade 2. *Assessment for Effective Intervention*, *38*(2), 76–90. https://doi.org/10.1177/1534508412456729

Strietholt, R., & Rosén, M. (2016). Linking Large-Scale Reading Assessments: Measuring International Trends Over 40 Years. *Measurement: Interdisciplinary Research and Perspectives*, *14*(1), 1–26. https://doi.org/10.1080/15366367.2015.1112711

Supovitz, J. (2009). Can high stakes testing leverage educational improvement? Prospects from the last decade of testing and accountability reform. *Journal of Educational Change*, *10*(2–3), 211–227. https://doi.org/10.1007/s10833-009-9105-2

Umar, A., Kusaeri, K., Ridho, A., Yusuf, A., & Asyhar, A. H. (2022). Does opportunity to learn explain the math score gap between madrasah and non-madrasah students in Indonesia? *Jurnal Cakrawala Pendidikan*, *41*(3), 792–805. https://doi.org/10.21831/cp.v41i3.40169

Uysal, İ., & Kilmen, S. (2016). Comparison of Item Response Theory Test Equating Methods for Mixed Format Tests. *International Online Journal of Educational Sciences*, *8*(2), 1–11. https://doi.org/10.15345/iojes.2016.02.001

van der Linden, W. J. (2000). A test-theoretic approach to observed-score equating. *Psychometrika*, *65*(4), 437–456. https://doi.org/10.1007/BF02296337

Wang, W. (2013). *Mixed-format test score equating* [University of Iowa]. https://doi.org/10.17077/etd.kvqyo3b2

Wei, W. (2013). *Mixed-format test score equating: Effect of item-type multidimensionality, length and composition of common-item set, and group ability difference* [The University of Iowa]. https://www.proquest.com/docview/1495946546

Widhiarso, W., & Ridho, A. (2022). *Validation of Setting and Design of Multi-Stage Testing (MST) to Portray Students’ Achievement on Reading Literacy in AKMI 2021*. https://doi.org/10.2991/assehr.k.220104.002

Yusron, E., Retnawati, H., & Rafi, I. (2020). Bagaimana hasil penyetaraan paket tes USBN pada mata pelajaran matematika dengan teori respon butir? *Jurnal Riset Pendidikan Matematika*, *7*(1), 1–12. https://doi.org/10.21831/jrpm.v7i1.31221

**Appendix 1.**  Average Item difficulties in Anchor Items

|  |  |  |
| --- | --- | --- |
| **Item** | **b\_2023** | **b\_2022** |
| X40091 | 0.759 | 1.713 |
| X40095 | 0.190 | 0.656 |
| X40103 | -0.631 | -0.650 |
| X40099 | 0.471 | 0.490 |
| X40107 | -0.184 | 0.550 |
| X40121 | 0.692 | 0.991 |
| X40137 | -0.241 | 0.609 |
| X40129 | 0.316 | -0.666 |
| X40125 | 0.623 | 0.614 |
| X40133 | 0.454 | 0.831 |
| X40122 | 0.948 | 0.537 |
| X40130 | 0.189 | -0.403 |
| X40126 | 0.570 | 0.896 |
| X40134 | 0.292 | 0.764 |
| X40142 | 0.547 | 0.913 |
| X40154 | -0.203 | 0.919 |
| X40145 | 0.666 | 0.815 |
| X40151 | 0.452 | 0.713 |
| X40148 | 0.521 | -0.921 |
| X40143 | 0.273 | 0.641 |
| X40146 | 0.801 | 0.735 |
| X40155 | -0.080 | 0.840 |
| X40149 | 0.234 | -0.605 |
| X40144 | 0.780 | 1.118 |
| X40156 | 0.674 | 0.872 |
| X40147 | 0.702 | 0.719 |
| X40150 | -0.605 | -0.532 |
| X40157 | 3.171 | 0.999 |
| X40160 | 0.279 | 0.936 |
| X40169 | 0.377 | 0.125 |
| X40158 | 0.629 | 0.855 |
| X40161 | -0.240 | 0.383 |
| X40167 | 0.444 | 0.256 |
| X40164 | 0.739 | 0.817 |
| X40170 | 0.228 | 0.464 |
| X40180 | 1.003 | 0.546 |
| X40192 | 0.360 | 0.606 |
| X40186 | 0.006 | -0.298 |
| X40195 | 5.061 | 1.435 |
| X40198 | 0.517 | -0.381 |
| X40204 | -0.055 | 0.259 |
| X40201 | 0.827 | 0.787 |
| X40207 | 0.332 | 0.289 |
| X40380 | 0.870 | 0.829 |
| X40395 | 0.617 | 0.662 |
| X40386 | 0.199 | -0.598 |
| X40392 | 0.449 | 0.297 |
| X40389 | 0.607 | 0.396 |
| X40673 | 0.081 | 1.952 |
| X40697 | 0.676 | 0.175 |
| X40685 | 0.575 | -0.866 |
| X40667 | 0.657 | 0.457 |
| X40679 | 0.613 | 0.487 |
| X40675 | 1.500 | 0.248 |
| X40693 | 0.438 | -0.624 |
| X40687 | 0.389 | -0.632 |
| X40669 | 0.165 | 0.503 |
| X40681 | 0.571 | 0.607 |
| X40729 | 0.057 | 0.989 |
| X40735 | 0.334 | 0.984 |
| X40723 | 0.000 | -0.505 |
| X40705 | 0.477 | 0.809 |
| X40211 | 2.234 | 1.642 |
| X40217 | 0.638 | 0.153 |
| X40223 | 0.465 | 0.252 |
| X40421 | 0.204 | -0.301 |
| X40416 | 0.340 | 0.826 |
| X40431 | 0.474 | 0.305 |
| X40422 | 0.548 | -0.160 |
| X40428 | 0.644 | 0.678 |
| X40425 | 0.685 | 0.581 |
| X40745 | 0.708 | 2.420 |
| X40757 | 0.070 | -0.368 |
| X40751 | 0.438 | 0.116 |
| X41198 | -4.030 | -0.772 |
| X41190 | 0.019 | 0.180 |
| X40815 | 0.751 | 1.997 |
| X40827 | 0.887 | 1.035 |
| X40823 | 0.097 | 0.159 |
| X40811 | 0.230 | 3.919 |
| X40819 | 0.889 | -0.284 |
| X40475 | 0.676 | 0.739 |
| X40485 | 0.610 | 0.594 |
| X40483 | 0.609 | 3.258 |
| X40481 | 0.807 | 0.700 |
| X40479 | 0.059 | -0.566 |
| X40476 | 0.498 | 1.360 |
| X40486 | 0.663 | -0.131 |
| X40484 | 0.525 | 0.878 |
| X40480 | 0.215 | -0.739 |
|  |  |  |
| M | 0.479 | 0.510 |
| SD | 0.832 | 0.815 |
| Min | -4.030 | -0.921 |
| Max | 5.061 | 3.919 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Pelaksanaan 2023** | | |  | **Pelaksanaan 2022** | | |
| **Butir** | **b1** | **b2** | **b3** |  | **b1** | **b2** | **b3** |
| X40091 | 0.759 |  |  |  | 1.713 |  |  |
| X40095 | 0.19 |  |  |  | 0.656 |  |  |
| X40103 | -1.983 | -0.534 | 0.623 |  | -2.501 | -0.437 | 0.987 |
| X40099 | 6.996 | -6.054 |  |  | 0.972 | 0.007 |  |
| X40107 | 9.584 | -9.952 |  |  | 9.154 | -8.054 |  |
| X40121 | 0.692 |  |  |  | 0.991 |  |  |
| X40137 | -0.241 |  |  |  | 0.609 |  |  |
| X40129 | -0.912 | 0.152 | 1.708 |  | -2.403 | -0.514 | 0.92 |
| X40125 | 6.99 | -5.745 |  |  | 1.32 | -0.092 |  |
| X40133 | 10.148 | -9.241 |  |  | 10.337 | -8.676 |  |
| X40122 | 0.948 |  |  |  | 0.537 |  |  |
| X40130 | -0.843 | 0.032 | 1.377 |  | -2.271 | -0.088 | 1.151 |
| X40126 | 5.991 | -4.852 |  |  | 1.36 | 0.431 |  |
| X40134 | 9.19 | -8.606 |  |  | 9.245 | -7.718 |  |
| X40142 | 0.547 |  |  |  | 0.913 |  |  |
| X40154 | -0.203 |  |  |  | 0.919 |  |  |
| X40145 | 6.976 | -5.644 |  |  | 1.509 | 0.121 |  |
| X40151 | 10.548 | -9.645 |  |  | 1.343 | 0.082 |  |
| X40148 | -0.675 | 0.294 | 1.944 |  | -2.31 | -0.694 | 0.24 |
| X40143 | 0.273 |  |  |  | 0.641 |  |  |
| X40146 | 7.09 | -5.488 |  |  | 1.46 | 0.009 |  |
| X40155 | -0.08 |  |  |  | 1.402 | 0.277 |  |
| X40149 | -0.673 | 0.147 | 1.228 |  | -2.112 | -0.456 | 0.752 |
| X40144 | 0.78 |  |  |  | 1.118 |  |  |
| X40156 | 0.674 |  |  |  | 0.872 |  |  |
| X40147 | 7.014 | -5.61 |  |  | 1.405 | 0.032 |  |
| X40150 | -2.334 | -0.367 | 0.887 |  | -2.03 | -0.535 | 0.968 |
| X40157 | 3.171 |  |  |  | 0.999 |  |  |
| X40160 | 0.279 |  |  |  | 0.936 |  |  |
| X40169 | 9.737 | -8.983 |  |  | 8.825 | -8.575 |  |
| X40158 | 0.629 |  |  |  | 0.855 |  |  |
| X40161 | -0.24 |  |  |  | 0.383 |  |  |
| X40167 | -0.725 | 0.258 | 1.798 |  | -1.298 | 0.266 | 1.801 |
| X40164 | 7.193 | -5.715 |  |  | 1.554 | 0.079 |  |
| X40170 | 10.683 | -10.23 |  |  | 9.04 | -8.111 |  |
| X40180 | 1.003 |  |  |  | 0.546 |  |  |
| X40192 | 0.36 |  |  |  | 0.606 |  |  |
| X40186 | -1.826 | 0.087 | 1.756 |  | -1.398 | -0.28 | 0.784 |
| X40195 | 5.061 |  |  |  | 1.435 |  |  |
| X40198 | 0.517 |  |  |  | -0.381 |  |  |
| X40204 | -1.827 | -0.01 | 1.673 |  | -1.333 | 0.358 | 1.753 |
| X40201 | 7.43 | -5.777 |  |  | 1.488 | 0.086 |  |
| X40207 | 10.477 | -9.813 |  |  | 8.963 | -8.385 |  |
| X40380 | 0.87 |  |  |  | 0.829 |  |  |
| X40395 | 0.617 |  |  |  | 0.662 |  |  |
| X40386 | -1.785 | 0.007 | 2.376 |  | -1.92 | -0.515 | 0.641 |
| X40392 | 5.225 | -4.327 |  |  | 0.835 | -0.241 |  |
| X40389 | 9.732 | -8.518 |  |  | 9.02 | -8.228 |  |
| X40673 | 0.081 |  |  |  | 1.952 |  |  |
| X40697 | 0.676 |  |  |  | 0.175 |  |  |
| X40685 | -0.775 | 0.15 | 2.351 |  | -2.486 | -0.671 | 0.558 |
| X40667 | 6.312 | -4.999 |  |  | 1.405 | -0.492 |  |
| X40679 | 10.659 | -9.433 |  |  | 9.118 | -8.145 |  |
| X40675 | 1.5 |  |  |  | 0.248 |  |  |
| X40693 | 0.438 |  |  |  | -0.624 |  |  |
| X40687 | -1.326 | 0.437 | 2.057 |  | -2.675 | -0.326 | 1.104 |
| X40669 | 4.818 | -4.488 |  |  | 1.478 | -0.473 |  |
| X40681 | 7.57 | -6.429 |  |  | 9.207 | -7.994 |  |
| X40729 | 0.057 |  |  |  | 0.989 |  |  |
| X40735 | 0.334 |  |  |  | 0.984 |  |  |
| X40723 | -1.318 | 0.345 | 0.972 |  | -2.193 | -0.388 | 1.066 |
| X40705 | 4.925 | -3.972 |  |  | 1.343 | 0.274 |  |
| X40211 | 2.234 |  |  |  | 1.642 |  |  |
| X40217 | 6.962 | -5.686 |  |  | 1.924 | -1.619 |  |
| X40223 | 10.855 | -9.926 |  |  | 5.487 | -4.983 |  |
| X40421 | -1.496 | 0.006 | 2.102 |  | -2.872 | -0.178 | 2.146 |
| X40416 | 0.34 |  |  |  | 0.826 |  |  |
| X40431 | 0.474 |  |  |  | 0.305 |  |  |
| X40422 | -1.035 | 0.314 | 2.366 |  | -1.484 | -0.256 | 1.26 |
| X40428 | 7.344 | -6.056 |  |  | 1.332 | 0.023 |  |
| X40425 | 7.863 | -6.494 |  |  | 9.953 | -8.791 |  |
| X40745 | 0.708 |  |  |  | 2.42 |  |  |
| X40757 | -1.906 | 0.341 | 1.774 |  | -2.639 | -0.045 | 1.58 |
| X40751 | 7.177 | -6.301 |  |  | 5.659 | -5.427 |  |
| X41198 | -4.03 |  |  |  | -0.772 |  |  |
| X41190 | -0.033 | -0.603 | 0.693 |  | 7.118 | -6.758 |  |
| X40815 | 0.751 |  |  |  | 1.997 |  |  |
| X40827 | 0.887 |  |  |  | 1.035 |  |  |
| X40823 | -1.858 | 0.179 | 1.971 |  | -1.061 | 0.419 | 1.12 |
| X40811 | 6.225 | -5.766 |  |  | 3.919 |  |  |
| X40819 | 6.843 | -5.066 |  |  | 4.642 | -5.21 |  |
| X40475 | 0.676 |  |  |  | 0.739 |  |  |
| X40485 | 0.61 |  |  |  | 0.594 |  |  |
| X40483 | 6.446 | -5.228 |  |  | 1.37 | -0.14 | 8.545 |
| X40481 | 10.741 | -9.127 |  |  | 1.408 | -0.009 |  |
| X40479 | -1.724 | 0.06 | 1.841 |  | -1.838 | -0.415 | 0.555 |
| X40476 | 0.498 |  |  |  | 1.36 |  |  |
| X40486 | 0.663 |  |  |  | -0.131 |  |  |
| X40484 | 6.225 | -5.175 |  |  | 2.41 | -0.654 |  |
| X40480 | -1.516 | 0.088 | 2.074 |  | -1.842 | -0.472 | 0.098 |

**Appendix 2.**  Expected Score 2022 as 2023

T 2022 on T2023 Scale (SL)

T 2022 on T2023 Scale (H)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Theta** | **T 2023** | **T 2022** | **T 2022**  **Lord** | **T 2022**  **Haebara** |
| 1 | -4.0 | 42 | 19 | 45 | 48 |
| 2 | -3.9 | 45 | 20 | 49 | 52 |
| 3 | -3.8 | 49 | 22 | 52 | 56 |
| 4 | -3.7 | 52 | 25 | 57 | 61 |
| 5 | -3.6 | 57 | 27 | 61 | 66 |
| 6 | -3.5 | 61 | 29 | 66 | 71 |
| 7 | -3.4 | 66 | 32 | 72 | 77 |
| 8 | -3.3 | 72 | 35 | 77 | 83 |
| 9 | -3.2 | 77 | 38 | 84 | 90 |
| 10 | -3.1 | 84 | 42 | 90 | 97 |
| 11 | -3.0 | 90 | 45 | 98 | 104 |
| 12 | -2.9 | 97 | 49 | 105 | 113 |
| 13 | -2.8 | 105 | 54 | 114 | 122 |
| 14 | -2.7 | 114 | 58 | 123 | 131 |
| 15 | -2.6 | 123 | 63 | 132 | 142 |
| 16 | -2.5 | 132 | 69 | 143 | 153 |
| 17 | -2.4 | 143 | 74 | 154 | 165 |
| 18 | -2.3 | 154 | 81 | 166 | 177 |
| 19 | -2.2 | 166 | 87 | 179 | 191 |
| 20 | -2.1 | 179 | 94 | 193 | 206 |
| 21 | -2.0 | 192 | 102 | 207 | 221 |
| 22 | -1.9 | 207 | 110 | 223 | 238 |
| 23 | -1.8 | 223 | 118 | 240 | 256 |
| 24 | -1.7 | 239 | 128 | 258 | 275 |
| 25 | -1.6 | 257 | 138 | 277 | 295 |
| 26 | -1.5 | 276 | 148 | 297 | 317 |
| 27 | -1.4 | 297 | 160 | 319 | 340 |
| 28 | -1.3 | 319 | 172 | 342 | 365 |
| 29 | -1.2 | 342 | 185 | 367 | 391 |
| 30 | -1.1 | 367 | 199 | 394 | 420 |
| 31 | -1.0 | 394 | 214 | 423 | 450 |
| 32 | -0.9 | 423 | 231 | 454 | 483 |
| 33 | -0.8 | 453 | 249 | 487 | 518 |
| 34 | -0.7 | 486 | 268 | 522 | 556 |
| 35 | -0.6 | 522 | 288 | 561 | 597 |
| 36 | -0.5 | 560 | 310 | 602 | 641 |
| 37 | -0.4 | 601 | 333 | 646 | 688 |
| 38 | -0.3 | 646 | 359 | 694 | 739 |
| 39 | -0.2 | 693 | 385 | 745 | 794 |
| 40 | -0.1 | 745 | 414 | 800 | 852 |
| 41 | 0.0 | 800 | 444 | 859 | 914 |
| 42 | 0.1 | 858 | 475 | 921 | 980 |
| 43 | 0.2 | 921 | 508 | 988 | 1050 |
| 44 | 0.3 | 987 | 542 | 1057 | 1122 |
| 45 | 0.4 | 1057 | 576 | 1130 | 1197 |
| 46 | 0.5 | 1129 | 612 | 1205 | 1274 |
| 47 | 0.6 | 1205 | 647 | 1282 | 1352 |
| 48 | 0.7 | 1282 | 682 | 1360 | 1430 |
| 49 | 0.8 | 1359 | 717 | 1438 | 1507 |
| 50 | 0.9 | 1437 | 752 | 1515 | 1582 |
| 51 | 1.0 | 1514 | 785 | 1591 | 1655 |
| 52 | 1.1 | 1590 | 817 | 1663 | 1725 |
| 53 | 1.2 | 1663 | 848 | 1733 | 1792 |
| 54 | 1.3 | 1732 | 877 | 1799 | 1854 |
| 55 | 1.4 | 1798 | 904 | 1860 | 1911 |
| 56 | 1.5 | 1859 | 929 | 1917 | 1964 |
| 57 | 1.6 | 1917 | 953 | 1970 | 2013 |
| 58 | 1.7 | 1970 | 975 | 2018 | 2058 |
| 59 | 1.8 | 2018 | 995 | 2063 | 2099 |
| 60 | 1.9 | 2062 | 1014 | 2103 | 2136 |
| 61 | 2.0 | 2103 | 1031 | 2140 | 2169 |
| 62 | 2.1 | 2139 | 1047 | 2173 | 2200 |
| 63 | 2.2 | 2173 | 1061 | 2203 | 2228 |
| 64 | 2.3 | 2203 | 1074 | 2231 | 2253 |
| 65 | 2.4 | 2230 | 1086 | 2256 | 2276 |
| 66 | 2.5 | 2255 | 1097 | 2278 | 2297 |
| 67 | 2.6 | 2278 | 1107 | 2299 | 2316 |
| 68 | 2.7 | 2299 | 1116 | 2318 | 2334 |
| 69 | 2.8 | 2318 | 1124 | 2336 | 2350 |
| 70 | 2.9 | 2336 | 1132 | 2352 | 2365 |
| 71 | 3.0 | 2352 | 1139 | 2366 | 2379 |
| 72 | 3.1 | 2366 | 1145 | 2380 | 2391 |
| 73 | 3.2 | 2380 | 1151 | 2392 | 2403 |
| 74 | 3.3 | 2392 | 1156 | 2404 | 2413 |
| 75 | 3.4 | 2404 | 1161 | 2414 | 2423 |
| 76 | 3.5 | 2414 | 1165 | 2424 | 2432 |
| 77 | 3.6 | 2424 | 1169 | 2433 | 2440 |
| 78 | 3.7 | 2433 | 1173 | 2441 | 2448 |
| 79 | 3.8 | 2441 | 1176 | 2449 | 2455 |
| 80 | 3.9 | 2449 | 1179 | 2456 | 2462 |
| 81 | 4.0 | 2456 | 1182 | 2463 | 2468 |