Journal of Educational Research and Evaluation
Volume 8, Issue 1, 2024, pp. 183-194
P-ISSN: 2597-422x E-ISSN: 2549-2675
Open Access: https://doi.org/10.23887/jere.v8i1.67895

Using English and Indonesian in Increasing Students' Understanding and Knowledge in Science Lessons

Henny Sanulita[1*]

[1] Universitas Tanjungpura Pontianak, Kalimantan Barat, Indonesia

ARTICLE INFO

Article history:
Received November 25, 2023
Accepted February 10, 2024
Available online February 25, 2024

Kata Kunci :
Inggris, Indonesia, kesetaraan Prestasi, Efektivitas

Keywords:
England, Indonesia, Equality of Achievement, Effectiveness

ABSTRAK

Kementerian Pendidikan Indonesia telah memperkenalkan dasar pengajaran Sains dan Matematika dalam Bahasa Inggris. Namun hasil belajar yang diperoleh siswa pada pembelajaran IPA masih lebih unggul bila menggunakan bahasa Indonesia dibandingkan dengan bahasa Inggris. Padahal bahasa Inggris mendominasi semua sumber pembelajaran. Hal ini menjadi hal yang mendesak untuk diteliti, karena adanya kesenjangan antara teori, ekspektasi, dan fakta di lapangan. Tujuannya adalah untuk menganalisis dampak bahasa terhadap prestasi siswa pada mata pelajaran sains dengan mengkorelasikan dua tes dengan bahasa berbeda (Inggris dan Indonesia) dalam satu skala pengukuran. Metode penelitian menggunakan survei. Subjek penelitian adalah sekolah menengah pertama yang berjumlah 1000 orang. Teknik pengumpulan data menggunakan instrumen tes dan teknik analisis menggunakan Test Analysis Program (TAP). Hasil dan Temuan menunjukan, perubahan kebijakan bahasa, Lembaga Ujian Indonesia dan Majelis Ujian Indonesia menggunakan ujian bilingual untuk proses pembelajaran dan menilai mata Pelajaran untuk semua tingkat jauh lebih baik. Ditemukan bahwa prestasi siswa pada tes versi bahasa Indonesia masih lebih baik dibandingkan dengan tes versi bahasa Inggris. Implikasi pada proses pembelajaran di sekolah sepenuhnya menggunakan bahasa Indonesia dalam melaksanakan materi yang diajarkan dibandingkan menggunakan bahasa Inggris.

ABSTRACT

The Indonesian Ministry of Education has introduced the basis for teaching Science and Mathematics in English. However, the learning outcomes obtained by students in science learning are still superior when using Indonesian compared to English. Even though English dominates all learning sources. This is an urgent matter to research because there is a gap between theory, expectations, and facts in the field. The aim is to analyze the impact of language on student achievement in science subjects by correlating two tests with different languages (English and Indonesian) on one measurement scale. The research method uses a survey. The research subjects were junior high schools, numbering 1000 people. Data collection techniques use test instruments and analysis techniques use the Test Analysis Program (TAP). Results and Findings show that changes in language policy, the Indonesian Examinations Institute, and the Indonesian Examinations Council use bilingual examinations for the learning process and assess subjects for all levels much better. It was found that student achievement on the Indonesian version of the test was still better than the English version of the test. The implications for the learning process in schools are fully using Indonesian in implementing the material taught compared to using English.

1. INTRODUCTION

The Indonesian Ministry of Education introduced the teaching of Mathematics and Science in English in 2003 (Zein et al., 2020). All primary and special school students in their first year are required to study Mathematics and Science in English (Sandilos et al., 2020). Students in National Primary Schools (SRK) study Science and Mathematics in English, while students in Chinese National Type Schools (SJKC) use Mandarin and English (Hu & McGeown, 2020). In secondary school, Year One students study Mathematics and Science in English (Mohamed et al., 2020). Meanwhile, Lower Sixth Form students study Mathematics, Biology, Chemistry, Physics and computing in English (WİNARNO et al., 2020). With this change in language policy, the Indonesian Financial Audit Institute (LPM) and the Indonesian Examinations Council introduced bilingual tests (i.e. English and Indonesian) to assess subjects at all levels of public examinations (Roslan & Chen, 2022). The reason LPM held bilingual tests during the transition period (2003-2007) was so that students could understand the question requirements well. Students are given the option to answer in English or Indonesian, or both. According to the Director General of Education, Datuk Ahmad Sipon, the use of bilingualism in Mathematics and Science subject questions was held until 2007. Starting in 2008, the

*Corresponding author.
E-mail addresses: henny.sanulita@fkip.untan.ac.id (Henny Sanulita)

national Mathematics and Science examinations were proposed to be prepared only in English (Tatto et al., 2020; Nurtanto et al., 2020; Zayyadi et al., 2020).

The use of English as the language of instruction for Science and Mathematics subjects has raised concerns for educators, parents and students. Students must learn Science and Mathematics concepts in a language they do not yet speak. In addition, the Science and Mathematics teachers on duty have been trained to teach these subjects in Indonesian (Scherzinger & Brahm, 2023; Vidergor & Ben-Amram, 2020). Some parents and educators do not welcome the use of English as a medium for teaching Mathematics and Science because they consider learning through the mother tongue to be more effective (Vidergor & Ben-Amram, 2020). Apart from Science and Mathematics subjects, other subjects such as History, Geography and Life Skills are still taught in Indonesian (Kurniati et al., 2022). The Indonesian Ministry of Education will decide not to use bilingualism in the Science and Mathematics public examinations at the end of 2008 (Zein et al., 2020). When Science and Mathematics are assessed in English, validity issues need to be addressed. Therefore, to find out whether students' performance in Science subjects is influenced by language factors, science tests were prepared in two different languages but the same content was given to candidates to enable comparisons. Comparison of student performance based on science tests administered in different languages may not provide accurate and fair information. Research has shown that tests in different languages may not demonstrate psychometric equivalence (Kałamała et al., 2020, Khalil et al., 2020, Liu & Oga-Baldwin, 2022). Therefore, it is important to determine the equivalence of two tests in different languages before comparing student performance. Without identifying the equivalence of tests in different languages, it is invalid to determine the causes of differences in achievement, whether due to language differences or actual student abilities.

To analyze test questions, two types of statistical questions are commonly used (Bucherie et al., 2022; Muka et al., 2020). The first type is known as classic statistical questions, namely the level of difficulty of the questions and the discrimination index. However, this type of index has a weakness in that its value depends on the population. The results of the analysis change when the study groups are different due to differences in the knowledge and skill levels of the samples (Bucherie et al., 2022). The second type of item statistics results from the calibration of the Item Response model which contains item difficulty statistics, measurement error statistics, and item suitability statistics (Giusti et al., 2020). Additionally, they also state that item fit statistics provide an estimate of how well an item fits the model's expectations, i.e. knowledgeable candidates have a higher probability of providing the correct answer. Comparison of tests based on Item Response Theory is based on comparing parameters that are estimated separately and then placed on a common scale (Jin et al., 2020). Item Function Difference Analysis (DIF) is usually used during the adaptation process to identify items that have different functions between groups that use different languages (MacIntyre et al., 2020). DIF is said to exist if samples from different language groups have different probabilities of answering a question correctly after taking into account overall ability. Learning and assessing English Language Learners using English has limited student achievement in mathematics. (Liu & Oga-Baldwin, 2022; Sandilos et al., 2020).Thus, language proficiency influences the reliability and validity of a test.

Other research suggests using the mother tongue as the testing language to overcome this problem (Alakarash & Razak, 2020; Widodo, 2021). However other research argues that translating test questions from English to their mother tongue does not provide benefits for students if the medium of instruction is English (Curle et al., 2020). Items in the mother tongue have confused students' learning concepts in English because they may not be familiar with the terminology used (Fang & Liu, 2020; Friantary & Martina, 2018). Therefore, this research examines the influence of science tests in Indonesian, which is the main medium of teaching, on science achievement. This research consists of two stages: the first stage examines the equivalence of two science tests in different languages. Equivalence is determined by the discrimination index, difficulty index, reliability, and item DIF. Typically, two tests that measure the same content also measure the same construct. However, in this study, the use of different languages may lead to differences in the constructs being measured. The second stage then compares the performance of students given science tests in different languages. In general, people in Indonesia pay attention to exam performance, especially general exams. This is because exams play an important role in the assessment system in Indonesia. Apart from general exams, exams and exams at the school level such as semester exams are also emphasized by schools and parents. However, science subject exams at the school level are carried out only in English and are not bilingual like in general exams. Thus, student achievement at the school level may not be comparable to general exam performance. Even though science subjects are taught in English, students who are weak in English tend to give answers in Indonesian. Students may not be able to demonstrate true science accomplishments if they answer in a language they have not fully mastered.

This research is very urgent to research because there are differences between theory, expectations, and reality in the field. The theory says that the use of English in explaining the learning

process is interesting for teachers and students with the hope that science and mathematics lessons will also use English as the language of instruction. However, the reality in the field is that educators, parents, and students are worried that the use of English in the learning process has an impact on their knowledge and learning outcomes when applied to science and mathematics subjects. The existence of this gap makes it urgent to immediately investigate this research with the general aim of finding out the influence of language on student learning achievement in science subjects by correlating two different language tests (English and Indonesian) on a general measurement scale. Meanwhile, the specific aim is to find out the difference in learning achievement between students who answer science questions in English and students who answer science questions in Indonesian. To find out the results of tests that use different languages in terms of level of difficulty, reliability, and differences in science achievement between students who answer questions using two languages.

2. METHODS

The research method used in this research is a quantitative method with a survey approach (Strijker et al., 2020). The research subjects were 1000 middle school students. The samples were selected from ten schools involved in the research; five are National High Schools and the other five are National High Schools. Selected schools have carried out Final Academic Year Examinations by measuring Class One and Class Two students. According to the TIMSS grade 8 criteria, the students involved must have received a minimum of eight years of formal education. Therefore, the sample of this study only consisted of second-grade students. The number of samples in this research was 1000 students. Each student answers only one version of the science test. A total of 500 students answered the English version of the items and another 500 students answered the Indonesian version of the items. The sample selected for this research consisted of students who had the minimum ability to read and understand English and Indonesian based on classification by the class teacher. The data collection procedure for this research was carried out a week after the end-of-year exams, namely at the end of August and early November 2023. Thus, students were considered to have studied relevant science topics. When Indonesia takes the TIMSS class 8 exam in 2023, the exam will also be held at the end of October and early November 2023.

Data collection technique. This research uses science tests in two different language versions, namely English and Indonesian. Both versions of the test contain multiple-choice objective questions. The English version of the test questions were selected from the Third International Mathematics and Sciences Study (TIMSS) questions and are by the Science syllabus for Indonesia. There are a total of 146 questions issued by TIMSS. A science teacher has helped identify 40 questions that are by the international science syllabus in Indonesia. Two experienced science teachers who have taught science subjects are the test examiners and assessors of the selected questions. They match the items to the international Science syllabus. Subjects deemed not by the syllabus will be deleted. The two science teachers grouped the items into eleven headings, five for Stage One and six for Stage Two. Stage one topics include; (i) Cells as the Basic Unit of Life, (ii) Matter, (iii) Diversity of Resources on Earth, (iv) Air Around Us, and (v) Energy Sources. Topics for stage Two are (i) Heat, (ii) The World Through Our Senses, (iii) Nutrition, (iv) Biodiversity, (v) Water and Its Solutions, and (vi) Simple Machines. Based on the test specification table constructed, they identified 10 questions that were not suitable for the Form One and Form Two science titles. The final science exam consists of only 30 questions. To determine the validity of the test content, three science teachers were asked to assess the suitability of the test items on science syllabus topics using a Likert scale, namely from very suitable (5) to not suitable (1) to determine the suitability of the measuring items. appropriate topic. If the mean of an item is less than 2.5, then the item is not suitable for testing that topic. The evaluation results of the three science teachers showed that all questions were on the topics listed. After identifying items for the English version, the items were then translated and adapted to the Indonesian version. Translation is carried out simultaneously by two science teachers and a language teacher who is fluent in English and Indonesian. The three translators then discussed and agreed to test the Indonesian translation version. Exams are held on different days for different schools according to school activities and permits. One week is used to carry out exams in all schools. Even the Lower Secondary Assessment (PMR) test has 40 questions and a time allocation of one hour. Test time distribution is based on the number of items in each test version. The total time to answer each version of the test is 45 minutes because the test only has 30 items. Different versions of the test, namely the English version of the test and the Indonesian version of the test, are given to students according to their number position in the class. Students with odd numbers answered the English version of the test and students with even numbers answered the Indonesian version of the test. Thus researchers can collect data from the same two groups randomly. When carrying out the test, researchers and science teachers at school read out instructions for the number of questions in the test and the time given. Students answer on the answer sheet provided with the question

paper. With explanations from teachers and researchers, problems such as students not having time to answer all the questions and omissions such as not answering questions on the last page can be avoided. After the period is up, all question papers and answer papers are collected again.

Data analysis techniques by obtaining discrimination index values, difficulty index, and then the reliability of the two tests using the SPSS Version 25.0 tool (Legrand et al., 2022; Yu et al., 2021). Item differences were identified using logistic regression and Rasch models. A comparison of the achievements of students who answered the English and Indonesian tests was carried out using descriptive statistics and distribution curves. To determine whether a question can perform well in a test, a discrimination index value has been determined. An index value of 0.40 or more is a very good item, 0.30 to 0.39 is a good item but can be improved, 0.20 to 0.29 is an item that is moderate and needs to be modified, and an item whose discrimination index value is below 0.19 is excellent. weak and requires a lot of modification or deletion (Kundu et al., 2020). The difficulty level of a question is the percentage of students taking a test who answered a question correctly. The higher the percentage of a question answered correctly, the easier the question is. The higher the difficulty index, the easier the question is (Wu et al., 2020). To determine the level of difficulty of a question, figures of speech are calculated by dividing the number of students who answered the question correctly by the number of students who answered the question (Madilo et al., 2020). In this study, the KR-20 (Kuder-Richardson 20) was used to determine the reliability of the questions because the KR-20 is suitable for multiple-choice tests. The KR-20 is used to measure internal consistency reliability or how well the test measures cognitive factors. The index ranges from 0.00 to 1.00. A value close to 0.00 indicates that there are still many unknown factors and are not factors that need to be measured. Conversely, if the value is close to 1.00, it means there is one factor being measured. Before identification, the suitability of the data to the Rasch Model is first determined. In this study, the mean squared value was in the range of 0.70-1.30. However, in this study, priority was given to the appropriate item content to decide whether the item should be excluded or not. Next, this research identifies the root mean square value In Abazov et al., (2016) states that items with $t < 1.96$ are items that show differences in item function at the $p < 0.05$ level. A comparison of student achievement in different language tests, namely English and Indonesian, was carried out by looking at the distribution of scores. This is demonstrated by using bar graphs, using z-scores for line graphs, and descriptive statistical analysis.

3. RESULT AND DISCUSSION

Results

Analysis of the discrimination index is shown in Table 1. Two items (item 6 and item 9) for the Indonesian version and one item for the English version (item 9) have a discrimination index of less than 0.20. Therefore, overall the discrimination index of both versions of the test has a satisfactory and good discrimination index based on the criteria proposed the discrimination index value is equal to or exceeds 0.20.

Table 1. Comparison of Discrimination Indices

Items	Science Test in Indonesian language	Science Test in English	Difference
1	0.44	0.41	0.03
2	0.31	0.49	−0.18
3	0.34	0.43	−0.09
4	0.56	0.58	−0.02
5	0.51	0.34	0.17
6	0.15*	0.33	−0.18
7	0.38	0.26	0.12
8	0.38	0.55	−0.17
9	0.09*	0.04*	0.05
10	0.58	0.49	0.09
11	0.34	0.42	−0.08
12	0.61	0.60	0.01
13	0.28	0.37	−0.09
14	0.28	0.33	−0.05
15	0.66	0.63	0.03
16	0.44	0.42	0.02
17	0.55	0.59	−0.04
18	0.46	0.51	−0.05
19	0.62	0.59	0.03

Items	Science Test in Indonesian language	Science Test in English	Difference
20	0.31	0.48	−0.17
21	0.47	0.35	0.12
22	0.29	0.44	−0.15
23	0.41	0.56	−0.15
24	0.65	0.59	0.06
25	0.61	0.63	−0.02
26	0.48	0.66	−0.18
27	0.22	0.20	0.02
28	0.66	0.62	0.04
29	0.38	0.43	−0.05
30	0.58	0.57	0.01

* Discrimination index less than 0.20.

A negative value occurs if the discrimination index for Indonesian language items is lower than the discrimination index for English items. The results of the analysis of the item difficulty index are presented in Table 2.

Table 2. Comparison of Item Difficulty Levels

Items	Science Test in Indonesian language	Science Test in English	Difference
1	0.75	0.70	0.05
2	0.64	0.58	0.06
3	0.84	0.81	0.03
4	0.65	0.61	0.04
5	0.53	0.49	0.04
6	0.92	0.88	0.04
7	0.74	0.71	0.03
8	0.55	0.48	0.07
9*	0.18*	0.18*	0
10	0.53	0.50	0.03
11	0.87	0.83	0.04
12	0.55	0.51	0.04
13*	0.73	0.73	0
14	0.82	0.76	0.06
15	0.57	0.58	−0.01
16	0.46	0.45	0.01
17	0.67	0.63	0.04
18	0.42	0.40	0.02
19	0.35	0.36	−0.01
20	0.73	0.66	0.07
21	0.54	0.49	0.05
22*	0.54	0.54	0
23	0.65	0.60	0.05
24	0.70	0.72	−0.02
25*	0.61	0.61	0
26	0.79	0.65	0.14
27	0.50	0.48	0.02
28*	0.68	0.68	0
29	0.86	0.84	0.02
30	0.40	0.35	0.05

Questions of the same level of difficulty for both test versions. A negative value occurs if the difficulty level of the Indonesian language items is lower than the difficulty level of the English items. The differences in item discrimination indices between the two versions are also shown in Table 1. The differences were found to be insignificant. The largest difference is 0.18, namely item 2, item 6, and item 26. 10 items show a difference in the discrimination index exceeding 0.10, namely item 2, item 5, item 6, item 7, item 8, item 20, item 21, item 22, item 23, and item 26. Seventeen items have a discrimination index difference of less than 0.10. Item 9 has the lowest difficulty index for the Indonesian version and the English

version, with a value of 0.18. This question is the most difficult because only 18% of students answered this question correctly. Meanwhile, the easiest item is item 6 for the Indonesian and English versions, which means the difficulty index is 0.92 for the Indonesian version and 0.88 for the English version. The nine items in the Indonesian language version have a difficulty index value between 0.7–0.9, namely item 1, item 3, item 7, item 11, item 13, item 14, item 20, item 26, and item 29. For the language version England, 8 questions have a difficulty index value in the range of 0.7–0.9. These items are item 3, item 6, item 7, item 11, item 13, item 14, item 24, and item 29. Items with a difficulty index value are classified as easy items. Most of the questions, namely 19 items for the Indonesian version and 21 items for the English version, are at a satisfactory level of difficulty, namely between 0.2 and 0.7. A comparison of the difficulty index values for Indonesian and English language items in the third row of Table 2 shows that the difference in the difficulty index for items in both versions is less than 0.1, except for item 26 with a difference of 0.14. Furthermore, five items have the same difficulty index for both language versions, namely item 9, item 13, item 22, item 25, and item 28. Two items in the Indonesian version have a smaller difficulty index than the items in the English version. namely item 19 and item 24. These items seem to be more difficult in Indonesia. However, the 23 items in the Indonesian version had a higher difficulty index value than the items in the English version. This shows that most of the questions are easier in Indonesian. The overall reliability of the Indonesian version of the test is 0.82 and 0.85 for the English version. Reliability comparison showed in Table 3. Differences in item function (DIF) using WINSTEPS software in Table 4.

Table 3. Reliability Comparison

Items	KR-20 Indonesia if the goods are issued	English KR-20 if goods are issued
1	0.814	0.841
2	0.82	0.842
3	0.815	0.838
4	0.811	0.838
5	0.815	0.844
6	0.819	0.84
7	0.817	0.845+
8	0.818	0.84
9	0.824+	0.850+
10	0.813	0.841
11	0.814	0.838
12	0.812	0.838
13	0.82	0.843
14	0.818	0.842
15	0.81	0.837
16	0.817	0.843
17	0.812	0.837
18	0.816	0.841
19	0.811	0.838
20	0.818	0.841
21	0.816	0.844
22	0.822+	0.843
23	0.816	0.84
24	0.808	0.836
25	0.811	0.836
26	0.811	0.835
27	0.824+	0.849+
28	0.808	0.836
29	0.813	0.838
30	0.813	0.839

Table 4. Differences in Item Function (DIF) using WINSTEPS Software

Items	DIF measurement	DIFS.E.	DIF measurement	DIF S. E	DIF *Contrast*	*Joint* S.E.	t	d.f
	English language			Indonesian language		The difference		

Items	DIF measurement	DIFS.E.	DIF measurement	DIF S. E	DIF *Contrast*	*Joint* S.E.	t	d.f
1	−0.54	0.11	−0.65	0.11	0.11	0.16	0.72	995
2	0.11	0.1	−0.02	0.1	0.14	0.14	0.96	995
3	−1.23	0.12	−1.32	0.13	0.09	0.18	0.52	995
4	−0.04	0.1	−0.07	0.1	0.02	0.15	0.17	995
5	0.58	0.1	0.56	0.1	0.02	0.14	0.13	995
6	−1.87	0.15	−2.13	0.17	0.26	0.22	1.16	995
7	−0.62	0.11	−0.58	0.11	−0.04	0.16	−0.28	994
8	0.61	0.1	0.45	0.1	0.16	0.14	1.12	995
9	2.42	0.13	2.54	0.13	−0.12	0.18	−0.69	994
10	0.53	0.1	0.54	0.1	−0.01	0.14	−0.09	994
11	−1.4	0.13	−1.54	0.14	0.14	0.19	0.74	995
12	0.47	0.1	0.44	0.1	0.03	0.14	0.19	995
13	−0.71	0.11	−0.53	0.11	−0.18	0.16	1.15	995
14	−0.88	0.11	−1.1	0.12	0.22	0.17	1.3	995
15	0.14	0.1	0.35	0.1	−0.22	0.14	1.54	995
16	0.77	0.1	0.91	0.1	−0.14	0.14	0.95	995
17	−0.16	0.1	−0.16	0.1	0	0.15	0.03	995
18	1.04	0.1	1.09	0.1	−0.05	0.14	0.34	995
19	1.24	0.1	1.43	0.1	−0.18	0.15	1.26	994
20	−0.31	0.11	−0.54	0.11	0.23	0.15	1.5	995
21	0.57	0.1	0.51	0.1	0.06	0.14	0.41	995
22	0.32	0.1	0.51	0.1	−0.19	0.14	1.37	995
23	0.01	0.1	−0.06	0.1	0.07	0.15	0.46	995
24	−0.65	0.11	−0.33	0.11	−0.32	0.15	2.0*	995
25	−0.03	0.1	0.15	0.1	−0.18	0.14	1.27	995
26	−0.27	0.1	−0.88	0.12	0.61	0.16	3.8*	995
27	0.61	0.1	0.68	0.1	−0.07	0.14	0.49	995
28	−0.42	0.11	−0.21	0.1	−0.22	0.15	1.45	995
29	−1.52	0.13	−1.47	0.14	−0.06	0.19	−0.3	995
30	1.31	0.1	1.19	0.1	0.12	0.15	0.8	995

Positive DIF indicates that English items are more difficult. * indicates DIF. The reliability of the Indonesian version of the test can be increased by removing item 9, item 22 and item 27. For the English version of the test, reliability increases when item 7, item 9 and item 27 are removed. Item 9 and item 27 are common items that can increase the reliability of both versions of the test in different languages. The table shows only two items that show DIF based on the criteria of t 1.96 at the p level 0.05. The items marked as DIF are item 24 and item 26 because their respective t values are –2.08 and 3.87. In addition, the difference between the item difficulty parameters calibrated with WINSTEPS shows that 14 items are more difficult to answer in Indonesian (the DIF difference is negative), namely items 7, 9, 10, 13, 15, 16, 18, 19, 22 , 24, 25, 27, 28 and 29. For the remaining 16 items that have a positive DIF difference, the items are more difficult in English. Based on this table, the conclusion that can be obtained is that the two versions of this test are quite equivalent because only two items show DIF.

Comparison of Student Achievement in English and Indonesian Version Tests. Comparison of student achievement in tests in different languages, namely English and Indonesian, can be shown by using descriptive statistics, bar graphs and z-score distribution. Comparison of student achievement with descriptive statistics showed in Table 5.

Table 5. Comparison of Student Achievement with Descriptive Statistics

	Indonesian version test	Test English version
Minimum Score	4 (13.30%)	2 (6.70%)
Maximum Score	30 (100.00%)	30 (100.00%)
Median Score	19 (63.30%)	18 (60.00%)
Mean Score	18.746 (62.50%)	17.838 (59.50%)
Standard Deviation	5.473	5.946
Variant	29.949	35.36
Mean Discrimination Index	0.435	0.463

	Indonesian version test	Test English version
Mean Item Difficulty	0.625	0.595
Skewness	−0.292	−0.325
Kr-20	0.820	0.845
Number of Students	500	500

Based on descriptive statistics, it is known that the minimum scores for the English version and the Indonesian version are 2 and 4 respectively. Meanwhile, the maximum score for both versions is the same, namely 30. The median score and average score for the Indonesian version are higher than the Indonesian version. English. The median score for the Indonesian version is 19 and the English version is 18. Meanwhile, the average score for the Indonesian version is 18,746, slightly higher than the English version, namely 17,893. Overall, the performance of students who answered the Indonesian version of the test was better than students who answered the English version. The standard deviation and variance values for the English version are greater than the Indonesian version. This corresponds to the larger skewness values for the English version of the test. However, because the negative value is relatively close to zero, student achievement in both versions of the test tends to be normally distributed. However, a comparison between the average difficulty of the questions proves that the Indonesian version of the test is easier (p = 0.625) compared to the English version (p = 0.595). For discrimination purposes, the English version of the test is a better test. Table 6 shows the distribution of student scores which have been converted into z-scores and z-scores (normalized). In general, the number of students who got a positive z-score was greater among students who took the Indonesian version of the test compared to students who took the English version of the test. Comparison in tabular form does not show significant differences. To provide a more comprehensive comparison, a line graph is depicted based on the z-score distribution as in Figure 1. Overall, the line graph shows that the results of the Indonesian version of the test are better than the English version of the test. Between a z score of −0.8 to 0.5, more students who answered the Indonesian version of the test scored in this range compared to students who answered the English version of the test. Distribution of student scores on the Indonesian version of tests and exams English version Indonesian language group English language group showed in Table 6.

Table 6. Distribution of Student Scores on the Indonesian Version of Tests and Exams English Version Indonesian Language Group English Language Group

Skor	Skor-z	%	Skor-z	Skor-z	%	Skor-z Normalized
1	−2.51	0.6	−2.65	−2.16	2.4	−2.07
2	−2.33	1.6	−2.29	−1.99	3.4	−1.89
3	−2.15	2.8	−2.01	−1.82	5.6	−1.69
4	−1.96	4	−1.82	−1.65	7.4	−1.51
5	−1.78	5.6	−1.66	−1.49	9	−1.39
6	−1.6	7.8	−1.49	−1.32	13.6	−1.21
7	−1.42	10.2	−1.34	−1.15	17.4	−1.01
8	−1.23	14.2	−1.16	−0.98	21.2	−0.86
9	−1.05	19	−0.97	−0.81	24.6	−0.74
10	−0.87	23.6	−0.79	−0.65	30	−0.6
11	−0.68	30.2	−0.61	−0.48	34.6	−0.45
12	−0.5	33.8	−0.46	−0.31	39.8	−0.32
13	−0.32	38.2	−0.35	−0.14	44.2	−0.2
14	−0.14	45.6	−0.2	0.03	51	−0.01
15	0.05	53	−0.01	0.2	56.2	0.09
16	0.23	60	0.16	0.36	61.6	0.22
17	0.41	65.6	0.32	0.53	67.8	0.37
18	0.59	70.6	0.47	0.7	74.8	0.56
19	0.78	77.2	0.64	0.87	80.8	0.76
20	0.96	83.8	0.85	1.04	86	0.97
21	1.14	89	1.09	1.2	90.8	1.19
22	1.33	93.6	1.35	1.37	95.8	1.49
23	1.51	96.6	1.65	1.54	97.4	1.82
24	1.69	99	2.01	1.71	99	2.09
25	1.87	99.6	2.45	1.88	99.8	2.51
26	2.06	100	2.87	2.05	100	3.09

Discussion

The findings from the research show that in terms of the discrimination index, the English version and the Indonesian version of the questions do not show very significant differences. These findings are in line with the research (Cardoso et al., 2023; Padarian et al., 2020). Data analysis in this study found that there were only three items that showed a discrimination index difference greater than 0.1, namely 0.18. Apart from that, it was found that almost all the questions had good discrimination characteristics, namely above 0.2, except for two questions in the Indonesian version and one question in the English version. Questions with low discrimination are the results of questions that are in the too-easy category (item 6) or questions that are in the too-difficult category (item 9). Meanwhile, for the comparison of the level of difficulty on the test, almost all questions have a difference of less than 0.1, except for item 26 which has the largest difference with a value of 0.14. This item is easier to find in the English version. Parameter calibration using the Rasch Model also supports the finding that item 26 is easier in the English version. The most difficult item, item 9, was identified using classical statistics and the Rasch Model. As a result, item 9 also has the lowest discrimination index, namely 0.18. It was discovered that when researchers studied this problem together with two science teachers, they thought that the cause of low student achievement was that students did not know the language that causes acid rain. In terms of reliability, the English version of the items is quite equivalent to the Indonesian version of the items. The reliability of the Indonesian version of items is 0.82 while the reliability of the English version of items is 0.845. Reliability values are relatively high (above 0.80) and do not change significantly when items are removed from any version of the test.

The results of the DIF analysis only identified two items whose functions were different when given in two different languages. Items that have DIF are item 24 and item 26. Item 24 and item 26 show DIF because the calculated t value is 2.08 and 3.87 respectively for $t > 1.96$ at the $p < 0.05$ level. However, if stricter criteria are selected with $t > 2.58$ at the $p < 0.01$ level then only one item shows DIF, namely item 26. This finding is in line with the logistic regression analysis which detected only item 26 as a DIF item. It was found in logistic regression that it was the same as the Rasch Model results if the criteria used were at the $p < 0.01$ level. After testing item 26, it was discovered that the difficulty level of item 26 in the Indonesian version was much higher than the difficulty level in the English version. The existing DIF may be because the English version of question 26 is difficult for students to understand compared to the Indonesian version. By comparing student performance on both versions of the test, it was found that overall there was no significant difference between the use of Indonesian and the use of English. However, if we look at the frequency of each score, student achievement on the Indonesian version of the test is better than the English version of the test. From a descriptive comparison of student achievement, the minimum score for the Indonesian version of the test is 4 compared to the English version of the test, namely 2. The average score for the Indonesian version of the test is higher, namely 18,746 compared to the English version of the test, namely 17,838. This finding is in line with previous research findings that the use of Indonesian in learning and testing is still better than the learning and testing process using English (Susanto et al., 2020; Bashori et al., 2021).

The implication of this research is to provide an overview and confidence for educators to use Indonesian in teaching science and mathematics lessons to students in secondary schools. This finding has implications for the learning process in schools that have begun to fully use Indonesian in implementing the material taught, even though the books used during the learning process are books in English. Science teachers and mathematics teachers in secondary schools have translated English into Indonesian before giving it to students in class. Almost all of the learning plans and materials prepared by the teachers come from English books, this can be seen from the references used by the teachers. However, when given to students the material and lesson plans are already in Indonesian. This helps students and does not create new difficulties for students in studying at school and studying at home using the material provided by the teacher to students (Hernawati, 2016; Mawardini & Ningsih, 2022; Rohmawati & Kristanto, 2018). Another implication of this research is to confirm that using Indonesian is a mother tongue that can be accepted by all students at school. The findings of this research convince educators, parents the community, and students that in their daily lives using Indonesian is still much better than using English.

The weakness of this research is that the research only uses tests in science material and does not use tests in other materials. Another weakness of this research is that the research was only conducted in secondary schools and the subjects were chosen randomly. In this case, it could be that the sampling sample is too small and a larger sample is needed to confirm further findings. Suggestions from this research for further research are to conduct research at three levels, subjects consisting of elementary school, middle school, and high school. This research is recommended so that future research does not create bias. The next suggestion is that the research be carried out by conducting qualitative research involving more sources to be interviewed. So that the primary data found can strengthen arguments and convince educators, parents the community, and students to use Indonesian in the teaching and learning process.

4. CONCLUSION

The conclusion of this research is the results and research findings which show that the English version of the questions is equivalent to the English version of the questions based on analysis of discrimination ability, level of difficulty, reliability, and differences in question function. After determining the equivalence of the two items in the test, a comparison of achievement for different language versions of the test can be determined. It was found that the comparison showed that overall the difference in student scores on the two versions of the test was not significant. However, the facts found were that the research results showed that student achievement on the Indonesian language version of the test was still better than the English version of the test. These findings imply that Science assessment should continue in bilinguals for certain groups of students or certain schools only. Such accommodations can increase the validity of the test and ensure that the test provides correct information about students' abilities in science, especially students who are weak in English. However, the use of bilingualism may be more beneficial for certain groups or for students who have had an early English language background, namely those who are more fluent in English than Indonesian.

5. REFERENCES

Abazov, V. M., Abbott, B., Acharya, B. S., Adams, M., Adams, T., Agnew, J. P., Alexeev, G. D., Alkhazov, G., Alton, A., Askew, A., Atkins, S., Augsten, K., Aushev, V., Aushev, Y., Avila, C., Badaud, F., Bagby, L., Baldin, B., Bandurin, D. V., … Zivkovic, L. (2016). Measurement of spin correlation between top and antitop quarks produced in pp- collisions at √s = 1.96 TeV. *Physics Letters, Section B: Nuclear, Elementary Particle and High-Energy Physics*, *757*(10 June 2016), 199–206. https://doi.org/10.1016/j.physletb.2016.03.053.

Alakarash, H., & Razak, N. (2020). The Asian ESP Journal. *The Asian ESP Journal*, *16*(4), 6–21.

Bashori, M., van Hout, R., Strik, H., & Cucchiarini, C. (2021). Effects of ASR-based websites on EFL learners' vocabulary, speaking anxiety, and language enjoyment. *System*, *99*(July 2021), 102496.1-16. https://doi.org/10.1016/j.system.2021.102496.

Bucherie, A., Hultquist, C., Adamo, S., Neely, C., Ayala, F., Bazo, J., & Kruczkiewicz, A. (2022). A comparison of social vulnerability indices specific to flooding in Ecuador: principal component analysis (PCA) and expert knowledge. *International Journal of Disaster Risk Reduction*, *73*(October 2021), 102897.1-21. https://doi.org/10.1016/j.ijdrr.2022.102897.

Cardoso, A. S., Bryukhova, S., Renna, F., Reino, L., Xu, C., Xiao, Z., Correia, R., Di Minin, E., Ribeiro, J., & Vaz, A. S. (2023). Detecting wildlife trafficking in images from online platforms: A test case using deep learning with pangolin images. *Biological Conservation*, *279*(December 2022), 109905.1-9. https://doi.org/10.1016/j.biocon.2023.109905.

Curle, S., Yuksel, D., Soruç, A., & Altay, M. (2020). Predictors of English Medium Instruction academic success: English proficiency versus first language medium. In *System* (Vol 95). Elsevier Ltd. https://doi.org/10.1016/j.system.2020.102378.

Fang, F., & Liu, Y. (2020). 'Using all English is not always meaningful': Stakeholders' perspectives on the use of and attitudes towards translanguaging at a Chinese university. *Lingua*, *247*(November 2020), 102959.1-18. https://doi.org/10.1016/j.lingua.2020.102959.

Friantary, H., & Martina, F. (2018). Evaluasi Implementasi Penilaian Hasil Belajar Berdasarkan Kurikulum 2013 oleh Guru Bahasa Inggris dan Bahasa Indonesia di MTS Ja-Alhaq Kota Bengkulu. *Silampari Bisa: Jurnal Penelitian Pendidikan Bahasa Indonesia, Daerah, dan Asing*, *1*(2), 76–95. https://doi.org/10.31540/silamparibisa.v1i2.202.

Giusti, E. M., Jonkman, A., Manzoni, G. M., Castelnuovo, G., Terwee, C. B., Roorda, L. D., & Chiarotto, A. (2020). Proposal for Improvement of the Hospital Anxiety and Depression Scale for the Assessment of Emotional Distress in Patients With Chronic Musculoskeletal Pain: A Bifactor and Item Response Theory Analysis. *Journal of Pain*, *21*(3–4), 375–389. https://doi.org/10.1016/j.jpain.2019.08.003.

Hernawati, F. (2016). Pengembangan Perangkat Pembelajaran Matematika Dengan Pendekatan Pmri Berorientasi Pada Kemampuan Representasi Matematis. *Jurnal Riset Pendidikan Matematika*, *3*(1), 34. https://doi.org/10.21831/jrpm.v3i1.9685.

Hu, X., & McGeown, S. (2020). Exploring the relationship between foreign language motivation and achievement among primary school students learning English in China. *System*, *89*(April 2020), 102199.1-10. https://doi.org/10.1016/j.system.2020.102199.

Jin, K. Y., Reichert, F., Cagasan, L. P., de la Torre, J., & Law, N. (2020). Measuring digital literacy across three age cohorts: Exploring test dimensionality and performance differences. *Computers and Education*, *157*(November 2020), 103968.1-55. https://doi.org/10.1016/j.compedu.2020.103968.

Kałamała, P., Szewczyk, J., Chuderski, A., Senderecka, M., & Wodniecka, Z. (2020). Patterns of bilingual

language use and response inhibition: A test of the adaptive control hypothesis. *Cognition*, *204*(June), 104373.1-14. https://doi.org/10.1016/j.cognition.2020.104373

Khalil, H., Al-Shorman, A., Alghwiri, A. A., Abdo, N., El-Salem, K., Shalabi, S., & Aburub, A. (2020). Cross cultural adaptation and psychometric evaluation of an Arabic version of the modified fatigue impact scale in people with multiple sclerosis. *Multiple Sclerosis and Related Disorders*, *39*(April 2020), 101878.1-31. https://doi.org/10.1016/j.msard.2019.101878.

Kundu, S., Ughade, J. M., Sherke, A. R., Kanwar, Y., Tiwari, S., Jatwar, R., Gurudiwan, R., & Kundu, S. G. (2020). Impact Measurement on Medical Faculty for Adhering to Appropriate Guidelines in Framing Effective Multiple-Choice Questions for Item Analysis. *Journal of Medical Education*, *19*(1), 1–15. https://doi.org/10.5812/jme.103482.

Kurniati, E., Suwono, H., Ibrohim, I., Suryadi, A., & Saefi, M. (2022). International Scientific Collaboration and Research Topics on STEM Education: A Systematic Review. *Eurasia Journal of Mathematics, Science and Technology Education*, *18*(4), 1-14. https://doi.org/10.29333/ejmste/11903.

Legrand, N., Nikolova, N., Correa, C., Brændholt, M., Stuckert, A., Kildahl, N., Vejlø, M., Fardo, F., & Allen, M. (2022). The heart rate discrimination task: A psychophysical method to estimate the accuracy and precision of interoceptive beliefs. *Biological Psychology*, *168*(November 2021), 1-14. https://doi.org/10.1016/j.biopsycho.2021.108239.

Liu, M., & Oga-Baldwin, W. L. Q. (2022). Motivational profiles of learners of multiple foreign languages: A self-determination theory perspective. *System*, *106*(November 2021), 102762.1-17. https://doi.org/10.1016/j.system.2022.102762.

MacIntyre, P. D., Gregersen, T., & Mercer, S. (2020). Language teachers' coping strategies during the Covid-19 conversion to online teaching: Correlations with stress, wellbeing and negative emotions. *System*, *94*(November 2020), 102352.1-13. https://doi.org/10.1016/j.system.2020.102352.

Madilo, F. K., Owusu-Kwarteng, J., Parry-Hanson Kunadu, A., & Tano-Debrah, K. (2020). Self-reported use and understanding of food label information among tertiary education students in Ghana. *Food Control*, *108*(August 2019), 106841.1-6. https://doi.org/10.1016/j.foodcont.2019.106841.

Mawardini, I. D., & Ningsih, S. S. (2022). Pembelajaran Matematika Kelas IV Madrasah Ibtidaiyah Masa Pandemi Covid – 19. *Jurnal Basicedu*. https://doi.org/10.31004/basicedu.v6i2.2426.

Mohamed, R., Ghazali, M., & Samsudin, M. A. (2020). A Systematic Review on Mathematical Language Learning Using PRISMA in Scopus Database. *Eurasia Journal of Mathematics, Science and Technology Education*, *16*(8), 1-12. https://doi.org/https://doi.org/10.29333/ejmste/8300.

Muka, T., Glisic, M., Milic, J., Verhoog, S., Bohlius, J., Bramer, W., Chowdhury, R., & Franco, O. H. (2020). A 24-step guide on how to design, conduct, and successfully publish a systematic review and meta-analysis in medical research. *European Journal of Epidemiology*, *35*(1), 49–60. https://doi.org/10.1007/s10654-019-00576-5.

Nurtanto, M., Pardjono, P., Widarto, W., & Ramdani, S. D. (2020). The effect of STEM-EDP in professional learning on automotive engineering competence in vocational high school. *Journal for the Education of Gifted Young Scientists*, *8*(2), 633–649. https://doi.org/10.17478/JEGYS.645047.

Padarian, J., Minasny, B., & McBratney, A. B. (2020). Machine learning and soil sciences: A review aided by machine learning tools. *Soil*, *6*(1), 35–52. https://doi.org/10.5194/soil-6-35-2020.

Rohmawati, E., & Kristanto, V. H. (2018). Pengembangan Media Pembelajaran Menggunakan Geogebra Pada Sub Pokok Bahasan Garis Singgung Persekutuan Dua Lingkaran. *PYTHAGORAS: Jurnal Program Studi Pendidikan Matematika*, *7*(1), 78–88. https://doi.org/10.33373/pythagoras.v7i1.1186.

Roslan, M. H. bin, & Chen, C. J. (2022). Educational Data Mining for Student Performance Prediction: A Systematic Literature Review (2015-2021). *International Journal of Emerging Technologies in Learning*, *17*(5), 147–179. https://doi.org/10.3991/ijet.v17i05.27685.

Sandilos, L. E., Baroody, A. E., Rimm-Kaufman, S. E., & Merritt, E. G. (2020). English learners' achievement in mathematics and science: Examining the role of self-efficacy. *Journal of School Psychology*, *79*(February 2020.), 1–15. https://doi.org/10.1016/j.jsp.2020.02.002.

Scherzinger, L., & Brahm, T. (2023). A systematic review of bilingual education teachers' competences. *Educational Research Review*, *39*(March), 100531.1-23. https://doi.org/10.1016/j.edurev.2023.100531.

Strijker, D., Bosworth, G., & Bouter, G. (2020). Research methods in rural studies: Qualitative, quantitative and mixed methods. *Journal of Rural Studies*, *78*(June 2018), 262–270. https://doi.org/10.1016/j.jrurstud.2020.06.007.

Susanto, J., Zheng, X., Liu, Y., & Wang, C. (2020). The impacts of climate variables and climate-related extreme events on island country's tourism: Evidence from Indonesia. *Journal of Cleaner Production*, *276*(2), 124204.1-9. https://doi.org/10.1016/j.jclepro.2020.124204.

Tatto, M. T., Rodriguez, M. C., & Reckase, M. (2020). Early career mathematics teachers: Concepts, methods,

and strategies for comparative international research. *Teaching and Teacher Education*, *96*(November 2020), 103118.1-18. https://doi.org/10.1016/j.tate.2020.103118.

Vidergor, H. E., & Ben-Amram, P. (2020). Khan academy effectiveness: The case of math secondary students' perceptions. *Computers and Education*, *157*(July), 103985.1-12. https://doi.org/10.1016/j.compedu.2020.103985.

Widodo, G. (2021). Penggunaan Bahasa Ibu Sebagai Alat Komunikasi Pengantar Bahasa Indonesia Di Sekolah Dasar. *Jurnal Ilmiah Edukasia*, *1*(1). https://doi.org/10.26877/jie.v1i1.7960.

Winarno, N., Rusdiana, D., Samsudin, A., Susilowati, E., Ahmad, N., & Afifah, R. M. A. (2020). The steps of the Engineering Design Process (EDP) in science education: A systematic literature review. *Journal for the Education of Gifted Young Scientists*, *8*(4), 1345–1360. https://doi.org/10.17478/jegys.766201.

Wu, Z., He, T., Mao, C., & Huang, C. (2020). Exam paper generation based on performance prediction of student group. *Information Sciences*, *532*(September 2020), 72–90. https://doi.org/10.1016/j.ins.2020.04.043.

Yu, S. Y., Suh, E. E., Kim, Y. M., Nguyen, T. A. P., Badamdorj, O., Seok, Y., Jang, S., & Ahn, J. (2021). Tablet PC-based competency evaluation for nursing students in three Asian countries: Cross-sectional comparative study. *Nurse Education in Practice*, *57*(Februari 2021), 103230.1-6. https://doi.org/10.1016/j.nepr.2021.103230.

Zayyadi, M., Nusantara, T., Hidayanto, E., Made, I., & Dijah, C. S. A. (2020). *Content and Pedagogical Knowledge of Prospective Teachers in Mathematics Learning : Commognitive*. *8*(March), 515–532.

Zein, S., Sukyadi, D., Hamied, F. A., & Lengkanawati, N. S. (2020). English language education in Indonesia: A review of research (2011–2019). *Language Teaching*, *53*(4), 491–523. https://doi.org/10.1017/S0261444820000208.