International Journal of Evaluation and Research in Education (IJERE)
Vol. 10, No. 4, December 2021, pp. 1334~1343
ISSN: 2252-8822, DOI: 10.11591/ijere.v10i4.21412

Detection of potential errors in measurement results of madrasa admission instruments in Indonesia

Ahmad Yusuf[1], Kusaeri Kusaeri[2], Ahmad Hidayatullah[3], Dian Candra Rini Novitasari[4], Ahmad Hanif Asyhar[5]
[1,2,4,5]UIN Sunan Ampel Surabaya, East Java, Indonesia
[3]Kementerian Agama Republik Indonesia, Jakarta, Indonesia

Article Info

Article history:

Received Dec 29, 2020
Revised Aug 9, 2021
Accepted Sep 12, 2021

Keywords:

Admission
Madrasa
Measurement
Potential errors
Religious education

ABSTRACT

Madrasa (Islamic boarding school) in Indonesia have a strategic role in character building. At present madrasa education is still considered second class education. Besides, to improve the quality of madrasas can be started by improving the quality of the student national admission to all madrasas in Indonesia. This study aimed to trace the potential errors in the measurement results of Students National Admission of Madrasah Aliyah Negeri (SNPDB MAN-IC) 2020. Tracing was carried out on two aspects: i) Equality between test sets used based on evidence of test responses; and ii) Further tests on equality between question sets based on evidence of relationship between variables, taking into account the origin of the participating schools (MTs/JHS) and the origin of the participating regions (West, Central and East of Indonesia). This study involved 13,115 participants in 23 MAN-ICs throughout Indonesia in 2020. The materials tested comprised learning potential and academic ability (Mathematics, Natural Sciences, Social Studies, English, Arabic, and Islamic Religious Education). The study used achievement test with mathematics as a sample of test subjects. Based on the test response evidence, it was found that seven of the 15 questions were thought to have an indication of inequality between item sets. The results of tracing the evidence between variables indicated that it was the participants' origin of institutions that influenced the inequality between item sets. On the other hand, regional origin did not affect the inequality between item sets because the majority of participants came from the western region of Indonesia.

Corresponding Author:

Ahmad Yusuf
Department of Information System
UIN Sunan Ampel Surabaya
East Java, 60237 Indonesia
Email: ahmadyusuf@uinsby.ac.id

1. INTRODUCTION

Madrasa (Islamic boarding school) has a strategic role for character building in Indonesia. Currently, madrasa education is still considered second class education in Indonesia. This happens because the administration of madrasas still faces a number of major problems, such as management issues and the poor quality of madrasa education [1], [2]. Madrasa studies on a national and international scale tend to involve three research groups. First, studies that focus on the historical review and education system of madrasas [3]-[7]. These researchers have an interest in examining the madrasa curriculum because madrasas are considered as a place to sow the seeds of radical Islam for young Muslims. Second, groups that have a

concern for madrasa achievements and dynamics [8]-[11]. Third, studies that concentrate on the problems of madrasa management due to various limitations they have. How each madrasa transforms to minimize its limitations is also their spotlight [9], [12]-[16].

The three research groups, when examined closely, focus their research more on traditional madrasas, so that what appears in their eyes is only the negative sides or weaknesses of madrasas. The superiority of madrasas has not been portrayed at all, even though Indonesia has Madrasah Aliyah Negeri Insan Cendekia (MAN-IC) which has a myriad of achievements, both in national and international levels, and are not inferior to public schools. These madrasas are spread in 23 provinces [16]. MAN-ICs are here to answer the demands of the Indonesian people who are worried about the impact of modernization that puts forward the spirit of competition instead of collaboration, mutual cooperation or community. The spirit of competition has caused people to experience alienation so that eventually they want to find their identity through religion [17].

MAN-IC's Education System tries to balance school culture with Islamic social-culture through activities in the dormitory [18]. School culture is designed to build academic traditions to explore mathematics, science and other subjects, while Islamic social culture is to build Islamic social-cultural character among students. People's need for fulfilling the balance of school culture and Islamic social-culture has resulted in high public interest in sending children to MAN-IC. If in 2016, 6,973 students took part in the selection of MAN-IC, then in 2020 17,344 students competed for a quota of 2,400 seats.

With the tight competition to enter MAN-ICs, the annual selection of new students certainly has important and strategic significance because it has implications for the learning process that will be conducted at MAN-ICs and the quality of the future graduates [19]. In addition, the competition also affects the sense of fairness from the selection system used [20]. The last point is important and needs to be emphasized because fairness in student selection has not received much attention.

An unfair and discriminatory selection system will be very detrimental or beneficial to certain groups of participants due to irrelevant factors; the influence also occurs systematically [21]. In a context like this, prospective students who actually have good potential and deserve to be accepted are very likely to fail and not to be accepted because the selection system used is not fair. This condition can certainly have bad consequences [22]. Not only on prospective students or parents of the prospective students, this unfairness also affects other people with an interest (social consequences of test selection). This can be seen when potential participants who are not accepted feel that their self-esteem is threatened in the eyes of their peers, or even feel hopeless with their future [21]. Therefore, how MAN-ICs select new students always attracts attention, especially regarding the fairness of the selection process.

Fairness in the context of student selection exams has been interpreted in various ways, depending on the point of view and aspects used. Fairness has a long history, with definitions that have evolved over time [23]. Educational Testing Services defines fairness as the extent to which conclusions made from the selection results are valid for all groups of participants who participate in the selection [24]. Some researchers [25], [26] are of the view that a selection instrument can be called fair if it does not benefit certain groups of participants or provides equal opportunities to all participants in demonstrating abilities optimally. Other experts also say that fairness has nothing to do with the quality of the test, but rather refers to the way the test results are used [27]. Referring to these opinions, it can be explicitly stated that fairness is directly related to the validity of the selection test.

Current assessment theories divide validity into five categories of validity evidence, namely evidence based on question content, evidence based on test responses, evidence based on internal structure, evidence based on relationships between variables, and evidence based on test consistency [28]-[30]. This paper is directed at two types of validity evidence, namely evidence based on test responses and evidence based on relationships between variables.

From the two types of validity evidence, it is expected that intact information can be obtained regarding the selection instrument used by MAN-ICs. Thus, the data generated to be used as a basis for making decisions on acceptance or rejection of participants is appropriate and fair for them [31]. The right selection decision as a result of a fair selection instrument is when prospective participants who are accepted at MAN-ICs indeed prove successful in showing high ability, while participants who are rejected, if given the opportunity to study at MAN-ICs, are not able to show sufficient ability [21].

The selection of prospective new students in MAN-ICs cannot be done in one place with one test, but in many places with more than two exam sessions because the number of participants is large and spread throughout Indonesia. For this reason, a parallel test instrument was developed to reduce the leakage of questions used in the sessions. In the construction of these parallel tests, indicators and levels of difficulty between sets developed are pursued to be equivalent [28], [32], because they measure the same construct.

The use of parallel instruments is very likely to trigger the potential error between groups in the measurement process [33]. The potential error not only affect the measurement results and selection

decisions, but also reduce fairness of test [31]. In addition, no measurement yields perfect accuracy [34]. Therefore, it is important to know how the potential error might occur in a measurement activity for selection purposes. This paper will focus on examining the equality of the question sets used in the National Admission of MAN-IC Students (SNPDB MAN-IC). If any inequality between sets is found, further tracing of the origin of schools and regions will be carried out. In this way it is expected to detect whether there are potential errors in the measurement results of the instruments used on the SNPDB MAN-IC. This result is very important to be used as a foothold for the SNPDB MAN-IC Committee of the Ministry of Religion in the coming years.

2. RESEARCH METHOD

The object studied in this study was a parallel question instrument of the selection of madrasa admissions that focused on mathematics. From the parallel instrument, the level of equality between sets will be examined. The analysis was carried out based on the response of the participants' answers to each instrument in the form of the level of truth and the distribution of sets to participants based on their school and region origins.

This study employed a quantitative approach using descriptive analysis that focuses on explaining the extent to which the SNPDB MAN-IC score results provide a picture of the actual state of the participants who took part in the selection [35]. From this study, the validity of the selection instruments used in the SNPDB MAN-IC will be revealed, especially the validity of evidence on the fairness of selection through analysis of test response evidence and relationships between variables in the distribution of test sets to participants [28]. Analysis of the evidence of the test responses was carried out in order to see the level of fairness of each question from the level of truth of the participants between the sets worked on. Meanwhile, the analysis of relations between variables in the distribution of test sets was carried out in order to see the factors that influenced the distribution of participants' answers.

2.1. Data collection

The data used in the study were SNPDB MAN-IC participant data for the academic year of 2020/2021 (n=13,115). The data were taken from the SNPDB MAN-IC Committee of the Ministry of Religion Affair of the Republic of Indonesia. When the participants took part in the selection, they were in the second semester of grade IX of the middle school and the average age was 15 years. The data were divided into two components, namely data from the selection results and participant profile data.

The data components of the selection results comprised the answers of the participants, the working time of the participants, and the status of the answers of the participants (true or false) contained in each question for each participant. Attributes contained in the participant profile data component comprised date of birth, gender, school origin, middle school address, province of school origin, and type of school origin. The province of school origin referred to one out of 34 provinces in Indonesia which were mapped into three regions namely Western Indonesia, Central Indonesia, and Eastern Indonesia. The division of regions was done based on time zones in Indonesia. Attributes of type of school origin consisted of Public Islamic Junior High School, namely Madrasah Tsanawiyah (MTs), Private MTs, Public Junior High Schools (JHS), and Private JHS.

Participant profile data were collected through survey forms filled out by the participants when registering for the SNPDB 2020/2021 selection system, while participant response data were obtained when the participants took the test. The test was a computer-based test [36], so that participants' answers could be recorded in a database system developed by the SNPDB MAN-IC Committee of the Indonesian Ministry of Religion.

2.2. Research instrument

The instrument used to obtain the data components of the selection results was in the form of a multiple choice (4-answer) written test. The instrument is achievement test that used in SNPDB. This test was developed by a team of national question compilers from the Directorate of Islamic school center assessment at the Indonesian (MORA). This test contents validated by experts consist of mathematics teachers, lecturers and education evaluation experts. The materials tested in SNPDB 2020/2021 were divided into two groups: the learning potential test and the academic ability test. The learning potential test consisted of five sub-subjects: numerical, verbal, spatial, and reasoning. Academic ability tests comprised sub-tests: mathematics, science, social studies, English, Arabic, and Islamic Education.

The test sample used in this study was a mathematics test (consisting of two question sets each of which consisted of 15 items). The sub-subject of the mathematics test was chosen as a sample with the consideration that mathematics is one of the leading subjects in MAN IC, so the score of the results of this test will affect student performance when they study at MAN ICs later. In addition, parents in Indonesia pay more attention to their children's mathematics learning outcomes than to other subjects [37]. Besides, critical

thinking ability is very important in mathematics learning [38]-[40]. Self-confidence in mathematical abilities is also an important factor that can influence student learning and performance in school [41].

2.3. Data analysis

Descriptive statistics (mean, frequency, and percentage) of released variables were calculated to answer the corresponding research problem. Subsequent analysis was carried out to test the equality between sets by considering the origin of the school and region. At the initial stage, the percentage of correct responses of each question q called correctness index (CI) on each set s is calculated for each type of institution or region with the formula [42].

$$CI\ (s,q) = \frac{(number\ of\ correct\ response\ of\ item\ q\ on\ set\ s)}{(number\ of\ response\ of\ item\ q\ on\ set\ s\ )} \quad (1)$$

The response value used to calculate the difference in each question q between set 1 and set 2 in the institution l and region w by using the relative difference formula.

$$\text{relative}\ difference\ (q) = \frac{(CI(1,q)-CI(2,q))}{((CI(1,q)+CI(2,q))/2)} \quad (2)$$

3. RESULTS AND DISCUSSION

Out of the 13,115 test takers, 4,756 participants (36.26%) were male and 8,359 participants (63.74%) were female. The composition of participants based on gender is very important to be presented in the context of the selection of MAN ICs because the education of MAN ICs is dormitory-based. On a dormitory basis, the quota for accepted male and female students is limited according to their capacity. Generally, the quota comparisons for male and female students are equal or balanced in 23 MAN ICs throughout Indonesia. Because the composition of female participants is more, then the level of competition in this group is certainly tighter compared to the group of male participants.

From the composition of the origin of the test participants' institutions or schools, 7,584 participants (57.82%) came from MTs and 5,531 participants (42.10%) came from JHS. When examined further in terms of public or private schools, 5,585 participants (42.58%) of the SNPDB MAN IC came from Public MTs and 1,999 participants (15.24%) came from private MTs. In JHS group, 2,466 participants (18.80%) were from Public JHS and 3,065 participants (23.37%) were from private JHS. The composition of the origin of the school is also important in this context because the Ministry of Religion provides a quota of 60% for MTs students and 40% for JHS students. More allocation for MTs students is reasonable because MTs are under the auspices of the Ministry of Religion. Thus, if more participants are found from JHS, it can be ensured that the competition in this group is tighter than the group of MTs students.

Most participants came from the West Indonesia region, i.e. 9,627 (76%), followed by the participants from Central Indonesia, i.e. 2,790 (22%). Thus, the fewest participants came from Eastern Indonesia, i.e. 301 participants (2%). The composition of the origin of region also has a significant role as does the origin of schools as described previously. Furthermore, the number of answers recorded on each item in each set for the mathematics sub-test items is presented in Table 1. Based on Table 1, the number of answers recorded on each item in each set did not appear to be the same. That was caused by the process of duplicating the item set constructed in parallel on each item.

Table 1. Distribution of item response on each set

Items	Set 1		Set 2	
	N	P	N	P
1	5890	45%	7225	55%
2	8591	66%	4523	34%
3	8097	62%	5016	38%
4	8248	63%	4863	37%
5	11048	84%	2060	16%
6	5348	41%	7759	59%
7	5458	42%	7649	58%
8	5288	40%	7816	60%
9	6226	48%	6878	52%
10	5811	44%	7291	56%
11	6302	48%	6800	52%
12	7148	55%	5951	45%
13	5042	38%	8056	62%
14	8156	62%	4937	38%
15	5211	40%	7882	60%

Based on Table 1, a review of the evidence of the test response was then conducted because one of the important things in selection fairness is equality in the questions the participants are working on [43], [44]. Recent study also showed that students' assessment result was influenced by the design of tasks and students' experiences especially in mathematics subject [45]. Therefore, participant responses were analyzed based on correctness index, which were compared with the set of questions worked out by participants as shown in Figure 1.

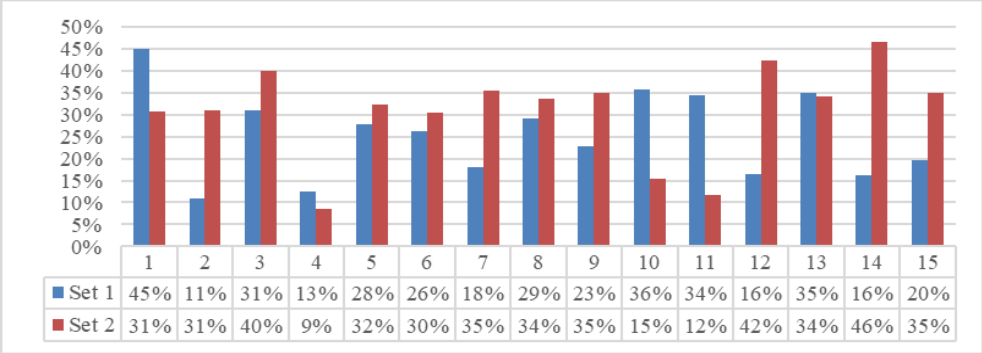

Figure 1. Correction index of each item on Set 1 and Set 2

Figure 1 shows several gaps in the level of correct answers in Set 1 and Set 2. The smallest difference in the level of correct answers is seen in the problem number 13 (1%), while the biggest difference is seen in the problem number 12 (26%). The average difference in the level of correct answers between Set 1 and Set 2 is 14%. There are seven question items having an above-the-average difference in the level of correct answers, namely the items number 2, 7, 10, 11, 12, 14, and 15, with the biggest difference of 30%. This finding should be suspected as an indication of inequality between Set 1 and 2 in terms of the level of difficulty of the questions or the ability of test takers. Therefore, referring to the evidence of this test response, it is necessary to analyze the relationship between the participant profile variables and the level of correct answers (in terms of origin of school and origin of region) in each set. From this analysis, we can identify the causes of why the seven items have a high difference in the level of correct answers.

3.1. Evidence of relationship between participant's response and origin of institution

The first variable that was hypothesized to be a factor that also influenced the results was the type of origin of educational institutions [46]. Research on the quality of madrasas has also been carried out by previous studies [47]–[50] in term of *Kompetisi Sains Madrasah* (Madrasa Sciences Competition). There were four types of origin of schools, namely Public MTs, Private MTs, Public JHS, and Private JHS. Table 2 shows the level of correct answers of participants based on the origin of the educational institution.

Table 2. Correctness index on each item based on origin of school

Items	Origin of school				
	Public MTs	Private MTs	Public JHS	Private JHS	Total
2	19%	18%	18%	15%	18%
7	28%	29%	25%	31%	28%
10	24%	24%	27%	23%	24%
11	22%	20%	24%	23%	23%
12	27%	28%	27%	30%	28%
14	27%	28%	28%	27%	28%
15	28%	29%	30%	29%	29%

From Table 2, the information was obtained that in general the percentage of correct answer levels of the participants from each type of educational institution did not differ too significantly. A significant difference occurred in the item number three, where the participants from private MTs had a much smaller level of correct answers compared to the others. The data also showed that the level of correct answers of the participants from JHS was better compared to that of the participants from MTs.

The achievement of the participants from JHS that was better than that of those from MTS was strengthened by the previous research which showed better learning achievement of JHS students compared to other types of origin of institutions [10], [51]. In addition, other studies also described the problems experienced by schools and madrasas [52]. The study explained that most madrasas in Indonesia were private, while public schools were mostly public. With the status as a state school, funding and control in the teaching and learning process have been largely fulfilled. The problem with public school students is they lack of religious knowledge compared to students from madrasas. On the other hand, with the addition of religious materials, the burden on madrasa students is more compared to students from schools [5]. This can cause the ability of madrasa students in general subjects such as mathematics to be less optimal as shown in the results of the study.

Further examination of the seven questions having a high degree of correct answer difference (the items number 2, 7, 10, 11, 12, 14, and 15) based on the participants' origin of institutions is described in Table 3. Table 3 indicates the unequal proportion of participants' origin of institution on the items number 2, 7, 10, 12, 14, 15, especially the participants from Public MTS and Public JHS. For the questions number 7 and 10, the percentage of the participants from Public JHS is greater, and the percentage of the participants from Public MTs in Set 1 is smaller than that in Set 2. On the other hand, for the questions number 2 and 12, the percentage of Public JHS participants in Set 1 is smaller than that in Set 2. Besides, for the questions number 14 and 15, the percentage of Public JHS participants is smaller, and the percentage of Public MTs participants in Set 1 is greater than that in Set 2. Only the question number 11 has a relatively equal distribution between Set 1 and Set 1.

Table 3. Percentage of response based on origin of school on each set

Items	Set 1				Set 2			
	Public MTs	Private MTs	Public JHS	Private JHS	Public MTs	Private MTs	Public JHS	Private JHS
2	43.0%	15.4%	16.8%	24.8%	41.8%	14.9%	22.6%	20.6%
7	38.8%	12.9%	27.5%	20.9%	45.3%	17.0%	12.6%	25.1%
10	39.5%	13.6%	26.1%	20.7%	45.0%	16.5%	13.0%	25.4%
11	42.3%	14.8%	19.2%	23.8%	42.9%	15.7%	18.5%	23.0%
12	43.6%	15.9%	17.6%	22.9%	41.4%	14.5%	20.2%	23.9%
14	44.1%	15.8%	16.2%	23.9%	40.2%	14.2%	23.1%	22.5%
15	45.2%	16.5%	14.6%	23.8%	40.9%	14.4%	21.6%	23.1%

Based on the findings on Table 3, the difference in the level of difficulty between sets compared to the difference in the percentage of examinees based on the type of institution is shown in Table 2. The difference is relatively positive if the percentage in Set 1 is greater than that in Set 2, and it will be negative if otherwise. The biggest difference is in the problem number 7, with a value of 73.91 in JHS, which means that in Set 1 the percentage of JHS participants is greater than that in Set 2.

Based on Table 4, in terms of the origin of institution, the questions number 2, 7, 10, 14 and 15 have a substantial difference in value, especially in JHS. The problems number 11 and 12 have a high relative difference in the level of correct answers (89.66 and 95.65), but based on the origin of institution, the relative difference does not have a high value. This indicates that there are other factors that affect participants' responses between sets. Meanwhile, the question number five has a relatively high difference in public JHS but has a small difference in the level of correctness. This shows that the origin of institution does not really affect the participants' answers between sets for the question number five.

Table 4. Relative difference based on origin of school

Items	Public MTs	Private MTs	Public JHS	Private JHS	All
2	2.73	3.44	-29.70	18.39	-95.24
7	-15.39	-27.46	73.91	-18.52	-64.15
10	-13.01	-19.30	67.16	-20.46	82.35
11	-1.40	-5.63	3.47	3.42	95.65
12	5.27	8.66	-13.76	-4.05	-89.66
14	9.21	10.65	-34.95	5.99	-96.77
15	9.92	13.44	-38.87	2.90	-54.55

The data on Table 4 show that the Public JHS participants tend to be able to solve problems better, the same as the previous findings. In the data, in Set 2 there are seven problems with more distribution on public JHS than that in Set 1 which has a higher level of correctness in Set 2 (both of which are negative). Whereas in Set 1, there are two questions with more distribution on public JHS than that in Set 2 which has a

higher level of correctness (both of which are positive). The problem number 7 has a different pattern; the participants from Public MTs are more in Set 2, resulting in a higher level of correctness in Set 2.

By looking at the relationship between variables, the correlation between the relative differences in each institution is calculated with the relative differences in the level of correctness using Pearson correlation. Based on the results of calculation of correlations and critical values, the Pearson correlation table shows the influence, especially on the relative difference in the number of participants between sets from Public MTs (r=-0.455) and Public JHS (r=0.453) on the relative difference in test results between sets. Therefore, it can be concluded that the distribution of participants' origin of institution between question sets has an effect on the level of correct answers.

3.2. Evidence of relationship between participant response and regional origin

The variable of participants' origin of area was also hypothesized to have an effect on participants' responses. Research on the relationship between participants' responses and regional origin has been widely studied by academics [53], [54]. Participants' origin of area is divided into three: western, central and eastern regions. Table 5 shows the level of correctness of the participants' responses based on these regions.

Table 5 shows in general the correctness level of each origin region did not differ too significantly. Significant differences occurred in number three, with the participants from the West having a higher percentage of correctness than those in the other parts (West 37%, Central 27%, East 18%). Table 5 displays the level of correctness of the participants from the West was better compared to that from the other regions.

Table 5. Correctness index on each item based on origin of region

Items	Region			Total
	West	Central	East	
2	17%	19%	16%	18%
7	28%	29%	29%	28%
10	24%	25%	28%	24%
11	23%	22%	17%	23%
12	29%	26%	31%	28%
14	28%	28%	24%	28%
15	29%	30%	27%	29%

When examined further based on the participants' region, the difference in the level of correct answers between the seven questions (the items number 2, 7, 10, 11, 12, 14, and 15) is shown in Table 6. In Table 6, it appears that the participants from the West were more dominant than those from the Central and East, both in Set 1 and Set 2. The correctness level between Set 1 and Set 2 was relatively similar between the participants from West, Central and East regions.

Based Table 6, in terms of origin of region, the difference in the level of difficulty between sets, when compared with the difference in correctness level among the test participants, is in accordance with (1). Based on the response value, the difference between the questions in Set 1 and Set 2 is calculated, each according to (2). The difference in the distribution of participant composition between Set 1 and Set 2 is shown in Table 7.

Table 7 shows that high differences exist in the eastern and central regions. This happened because the majority of participants came from the western region, so the eastern and central regions had small values which resulted in a more striking relative difference. The distribution of participants who were mostly from the western region but produced a varied relative difference in the level of correctness between Set 1 and Set 2 indicates that the distribution of regional origin does not significantly affect the level of correctness between sets.

Table 6. Percentage of response based on origin of region on each set

Items	Set 1			Set 2		
	West	Central	East	West	Central	East
2	75.3%	21.8%	2.8%	76.4%	22.1%	1.5%
7	77.3%	21.1%	1.6%	74.5%	22.6%	2.9%
10	76.9%	21.6%	1.5%	74.7%	22.2%	3.1%
11	75.5%	21.9%	2.5%	75.8%	22.0%	2.2%
12	75.1%	22.7%	2.2%	76.4%	21.1%	2.5%
14	74.0%	22.7%	3.2%	78.4%	20.6%	0.9%
15	73.9%	23.4%	2.7%	76.9%	21.0%	2.1%

Table 7. Relative difference based on origin of region

Items	West	Central	East
2	-1.42	17.87	80.85
7	3.75	-98.14	-144.62
10	2.96	-84.53	-141.72
11	-0.33	-64.60	-64.30
12	-1.75	-30.40	-54.02
14	-5.82	9.85	112.78
15	-3.94	-95.41	-99.01

To see the relationship between variables, the correlation between the relative differences in each institution and the relative differences in the level of correctness was calculated using Pearson correlation. The correlation of the West (r=0.364), Central (r=-0.236), and East (r=-0.134) regions has a relatively low value. The results of the calculation of correlation and Pearson correlation critical values show that the relative difference in the number of participants between sets based on the regions has no effect on the relative difference in test results between sets. This shows that the distribution of the participants' origin of area between sets on parallel questions has no effect on the difference in the level of correctness.

4. CONCLUSION

From the evidence of the test responses, it was found that seven out of the 15 questions were thought to have an indication of inequality between test sets of mathematics, although in terms of content and construct, the seven items were arranged in parallel. The results of testing the evidence relations between variables found that the type of participants' origin of institutions (MTs/JHS) was the cause of unequal test responses. The finding showed that the level of correct answers of the participants from JHS was better compared to that from MTs. An unbalanced proportion between the number of participants from MTs and JHS responding to the items caused a higher level of correctness in certain items in certain sets. Other results showed that participants' origin of regions did not affect the inequality between sets.

Future research recommendations are how to construct tests that can minimize errors not only in the scope of madrasas but also for the scope of Indonesian schools in general. Other methods can also be used as a comparison of the most appropriate method for determining the measurement results of Madrasa Admission Instruments in Indonesia. This research finding is very important to answer the initial concern that the instrument used for the SNPDB MAN-IC is inappropriate. Thus, these findings can be used as a foothold for the SNPDB MAN-IC Committee of the Indonesian Ministry of Religious Affairs in the coming years.

ACKNOWLEDGEMENTS

This research was supported by Kementerian Agama Republik Indonesia especially Direktorat Kurikulum, Sarana, Kelembagaan dan Kesiswaan Madrasah and UIN Sunan Ampel Surabaya.

REFERENCES

[1] F. Alawiyah, "Madrasa Education in Indonesia," (in Indonesia), *Aspirasi: Jurnal Masalah-Masalah Sosial*, vol. 5, no. 1, pp. 51–58, 2014. [Online]. Available: https://jurnal.dpr.go.id/index.php/aspirasi/article/view/449.

[2] K. Kusaeri, "2nd International Conference on Statistics, Mathematics, Teaching, and Research 2017," *J. Phys. Conf. Ser.*, vol. 1028, Jun. 2018, doi: 10.1088/1742-6596/1028/1/011001.

[3] S. Al-Hasani, A. Ismail, B. Kazeemkayode, and D. Elega, "Creating a Practicing Muslim: A Study of Qawmi Madrasah in Bangladesh," *Br. J. Educ. Soc. Behav. Sci.*, vol. 20, no. 3, pp. 1–9, May 2017, doi: 10.9734/BJESBS/2017/30910.

[4] B. Lintner, "Religious Extremism and Nationalism in Bangladesh," in *Religious Radicalism and Security in South Asia*. Asia-Pacific Center for Security Studies, 2004, pp. 413–436.

[5] R. Lukens-Bull, "Madrasa by any other name: Pondok, pesantren, and islamic schools in Indonesia and larger southeast Asian Region," *J. Indones. Islam*, vol. 4, no. 1, pp. 1-21, 2010, doi: 10.15642/JIIS.2010.4.1.1-21.

[6] K. McClure, "Madrasas and Pakistan's education agenda: Western media misrepresentation and policy recommendations," *International Journal of Educational Development*, vol. 29, no. 4, pp. 334-341, 2009, doi: 10.1016/j.ijedudev.2009.01.003.

[7] N. Nizah, "Dynamics Madrasah Diniyah: A Historical Review," (in Indonesia), *Edukasia J. Penelit. Pendidik. Islam*, vol. 11, no. 1, pp. 181-202, 2016, doi: 10.21043/edukasia.v11i1.810.

[8] M. Ali and Z. A. Furqon, "Madrasah Students' Achievement Study in Indonesia," *Global and Stochastic Analysis*, vol. 3, no. 3, pp. 181-190, 2016.

[9] M. N. Asadullah and N. Chaudhury, "To madrasahs or not to madrasahs: The question and correlates of enrolment in Islamic schools in Bangladesh," *Int. J. Educ. Dev.*, vol. 49, pp. 55-69, May 2016, doi: 10.1016/j.ijedudev.2016.01.005.
[10] K. Kusaeri and A. Ridho, "Learning outcome of mathematics and science: Features of Indonesian madrasah students," *J. Penelit. dan Eval. Pendidik.*, vol. 23, no. 1, pp. 95-105, 2019, doi: 10.21831/pep.v23i1.24881.
[11] A. Muhid, A. Ridho, A. Yusuf, M. Kurjum, M. Thohir, and A. H. Asyhar, "Cognitive and personality test as a predictor of religious education achievement among students of religious program of Islamic schools in Indonesia," *Elem. Educ. Online*, vol. 19, no. 4, pp. 2408-2418, 2020, doi: 10.17051/ilkonline.2020.764424.
[12] N. Ahid, "Madrasah Aliyah Management Problems and Its Solutions," (in Indonesia), *Islam. J. Stud. Keislam.*, vol. 4, no. 2, pp. 336-353, 2014, doi: 10.15642/islamica.2010.4.2.336-353.
[13] R. Hashim, S. A. Rufai, and M. R. M. Nor, "Traditional Islamic education in Asia and Africa: a comparative study of Malaysia's Pondok, Indonesia's Pesantren and Nigeria's traditional madrasah," *World Journal of Islamic History and Civilization*, vol. 1, no. 2, pp. 94-107, 2011.
[14] Z. Musthan, "An Effective Leadership Model for Madrasah Principals in Indonesia," *Pertanika J. Soc. Sci. Humanit.*, vol. 27, no. 1, pp. 133–147, 2019.
[15] N. Rao and M. Hossain, "Confronting Poverty and Educational Inequalities: Madrasas as a Strategy for Contesting Dominant Literacy in Rural Bangladesh," *Int. J. Educ. Dev.*, vol. 31, pp. 623–633, 2011, doi: 10.1016/j.ijedudev.2011.01.012.
[16] Indonesian Ministry of Religion, "Student National Admission of Madrasa," (in Indonesia), 2020. [Online]. Available: https://kemenag.go.id/read/pendaftaran-seleksi-siswa-baru-man-dibuka-11-januari-xkeek.
[17] R. Jackson, "Cameron, 'multiculturalism' and education about religions and beliefs," Ekklesia, May 2011. [Online]. Available: http://old.ekklesia.co.uk/node/14271.
[18] M. Sukarno, "Developing Insan Cendikia High School to Address Modernisation, Democracy, and Globalisation Challenges," (in Indonesia), *Jurnal Kependudukan Indonesia*, vol. 9, no. 2, pp. 117-134, 2014. [Online]. Available: http://ejurnal.kependudukan.lipi.go.id/index.php/jki/article/view/41.
[19] L. Fagioli, "Is the SAT the Root of all Evil? Reviewing the Evidence on Admission Policies and Diversity in Higher Education," *LUX*, vol. 2, no. 1, pp. 1-17, May 2013, doi: 10.5642/lux.201301.15.
[20] Asrijanty, "The Predictive Validity of Scholastic Aptitude Test and Achievement Test as Selection Criteria for University Entrance," (in Indonesia), *J. Pendidik. dan Kebud.*, vol. 20, no. 4, pp. 515-534, 2014, doi: 10.24832/jpnk.v20i4.163.
[21] S. Azwar, "Selection decisions in high stake exams: psychometric discourse," (in Indonesia), Thesis, Universitas Gadjah Mada, Yogyakarta, 2010.
[22] A. Muhid, A. Yusuf, Kusaeri, D. C. R. Novitasari, A. H. Asyhar, and A. Ridho, "Determining scholastic aptitude test as predictors of academic achievement on students of islamic school in indonesia," *New Educ. Rev.*, vol. 61, pp. 211–221, 2020, doi: 10.15804/tner.2020.61.3.17.
[23] N. J. Dorans, *Fairness in Educational Assessment and Measurement*, 1st ed. Routledge, 2016.
[24] M. Zieky, "Fairness reviews in assessment," in *Handbook of test development*. Routledge, 2011, pp. 373–390.
[25] B. Malau-Aduli, J. Walls, and C. Zimitat, "Validity, Reliability and Equivalence of Parallel Examinations in a University Setting," *Creat. Educ. J.*, vol. 3, no. 6, pp. 923-930, 2012.
[26] P. K. Crane, G. van Belle, and E. B. Larson, "Test bias in a cognitive test: differential item functioning in the CASI," *Stat. Med.*, vol. 23, no. 2, pp. 241–256, 2004, doi: 10.1002/sim.1713.
[27] L. Suskie, *Assessing student learning: A common sense guide*. John Wiley & Sons, 2018.
[28] American Psychological Association, *Standards for educational and psychological testing-American Educational Research Association*. Washington, D.C: American Educational Research Association, 2014.
[29] D. A. Cook, B. Zendejas, S. J. Hamstra, R. Hatala, and R. Brydges, "What counts as validity evidence? Examples and prevalence in a systematic review of simulation-based assessment," *Adv. Heal. Sci. Educ.*, vol. 19, no. 2, pp. 233-250, 2014, doi: 10.1007/s10459-013-9458-4.
[30] T. D. Reeves and G. Marbach-Ad, "Contemporary Test Validity in Theory and Practice: A Primer for Discipline-Based Education Researchers," *CBE—Life Sci. Educ.*, vol. 15, no. 1, Apr. 2016, doi: 10.1187/cbe.15-08-0183.
[31] D. David and N. Guspendri, "Impact of admission type on students' achievement in economy and business school of Batusangkar State Institute of Islamic Studies," *J. Penelit. dan Eval. Pendidik.*, vol. 22, no. 2, pp. 197–207, 2018, doi: 10.21831/pep.v22i2.22367.
[32] W. H. Finch and B. F. French, *Educational and psychological measurement*. New York, NY: Routledge, 2019.
[33] J. Møen and M. Tjelta, "Grading standards, student ability and errors in college admission," *Scand. J. Educ. Res.*, vol. 54, no. 3, pp. 221-237, 2010, doi: 10.1080/00313831003764503.
[34] F. A. Montero, A. M. Heráldez, and C. A. López, "Criteria and instruments for doctoral program admission," *Electron. J. Res. Educ. Psychol.*, vol. 12, no. 3, pp. 853-886, 2014, doi: 10.14204/ejrep.34.13138.
[35] S. Azwar, *Psychological Research Methods*, 2nd ed. (in Indonesia), Pustaka Pelajar, 2017.
[36] A. Z. Abidin, E. Istiyono, N. Fadilah, and W. S. B. Dwandaru, "A computerized adaptive test for measuring the physics critical thinking skills," *International Journal of Evaluation and Research in Education (IJERE)*, vol. 8, no. 3, pp. 376-383, 2019, doi: 10.11591/ijere.v8i3.19642.
[37] S. Lailiyah, T. Nusantara, C. Sa'Dijah, E. B. Irawan, Kusaeri, and A. H. Asyhar, "Structuring students' analogical reasoning in solving algebra problem," *IOP Conference Series: Materials Science and Engineering*, vol. 296, no. 1, 2018, doi: 10.1088/1757-899X/296/1/012029.

[38] K. Kusaeri and A. Aditomo, "Pedagogical Beliefs about Critical Thinking among Indonesian Mathematics Pre-service Teachers," *Int. J. Instr.*, vol. 12, no. 1, pp. 573–590, 2019, doi: 10.29333/iji.2019.12137a.
[39] N. Diana, D. Suryadi, and J. A. Dahlan, "Analysis of Students' Mathematical Connection Abilities in Solving Problem of Circle Material: Transposition Study," *J. Educ. Gift. Young Sci.*, vol. 8, no. 2, pp. 829–842, Jun. 2020, doi: 10.17478/jegys.689673.
[40] S. Syaiful, K. Kamid, and N. Huda, "Identifying of emotional quotient junior high school students in mathematics," International *Journal of Evaluation and Research in Education (IJERE),* vol. 10, no. 1, pp. 202–208, 2021, doi: 10.11591/ijere.v10i1.20681.
[41] H. Z. Chan and M. N. L. Y. Abdullah, "Validity and reliability of the Mathematics Self-Efficacy Questionnaire (MSEQ) on primary school students," *Pertanika J. Soc. Sci. Humanit.*, vol. 26, no. 4, pp. 2161–2177, 2018.
[42] L. Törnqvist, P. Vartia, and Y. O. Vartia, "How Should Relative Changes be Measured?" *Am. Stat.*, vol. 39, no. 1, pp. 43–46, Apr. 1985, doi: 10.1080/00031305.1985.10479385.
[43] G. Camilli, "Test fairness," in R. L. Brennan, Ed., *Educational measurement*, 4th Ed. Westport: American Council on Education & Praeger Publishers, 2006, pp. 221–256.
[44] P. Kline, *A handbook of test construction: Introduction to psychometric design*. London, New York: Methuen, 1986.
[45] A. T. Fatimah and S. Prabawanto, "Mathematical Understanding and Reasoning of Vocational School Students in Agriculture-Based Mathematical Tasks," *J. Educ. Gift. Young Sci.*, vol. 8, no. 2, pp. 701–712, Jun. 2020, doi: 10.17478/jegys.702884.
[46] C. Lubienski and S. T. Lubienski, *Charter, private, public schools and academic achievement [electronic resource]: new evidence from NAEP mathematics data*. National Center for the Study of Privatization in Education, Teachers College, Columbia University, 2006.
[47] A. H. Asyhar, *et al.*, "Graph Degree Linkage Clustering for Identify Student's Performance on Kompetisi Sains Madrasah in Indonesia," in Y.D. Zhang, T. Senjyu, C. So–In, A. Joshi, eds., *Smart Trends in Computing and Communications: Proceedings of SmartCom 2020. Smart Innovation, Systems and Technologies,* vol 182. Springer, Singapore, 2021, doi: 10.1007/978-981-15-5224-3_20.
[48] D. C. R. Novitasari, *et al.*, "Identify Education Quality Based on Islamic Senior High School Data in Kompetisi Sains Madrasah Using Fuzzy C-Means Clustering," in Y.D. Zhang, T. Senjyu, C. So–In, A. Joshi, eds., *Smart Trends in Computing and Communications: Proceedings of SmartCom 2020. Smart Innovation, Systems and Technologies,* vol 182. Springer, Singapore, 2021, doi: 10.1007/978-981-15-5224-3_49.
[49] A. Yusuf, *et al.*, "Identify Elementary Student Distribution Based on Kompetisi Sains Madrasah Data Using Probabilistic Distance Clustering," in Y.D. Zhang, T. Senjyu, C. So–In, A. Joshi, eds., *Smart Trends in Computing and Communications: Proceedings of SmartCom 2020. Smart Innovation, Systems and Technologies,* vol 182. Springer, Singapore, 2021, doi: 10.1007/978-981-15-5224-3_27.
[50] Kusaeri, *et al.*, "Stepwise Iterative Maximum Likelihood Clustering Based on Kompetisi Sains Madrasah' Scores for Identifying Quality of Junior High School Grading Distribution," in Y.D. Zhang, T. Senjyu, C. So–In, A. Joshi, eds., *Smart Trends in Computing and Communications: Proceedings of SmartCom 2020. Smart Innovation, Systems and Technologies,* vol 182. Springer, Singapore, 2021, doi: 10.1007/978-981-15-5224-3_21.
[51] S. Nanang, "PTKIN Admission Test and School Origin in Predicting Student Learning Achievement Using the Dummy Regression Model," (in Indonesia), *Al-Jabar*, vol. 10, no. 2, Apr. 2018. [Online]. Available: http://repository.radenintan.ac.id/6296.
[52] F. A. Yahya, "Pesantren, School and Madrasah Management Problems: Quality and Input-Process-Output Quality Problems," (in Indonesia), *el-Tarbawi*, vol. 8, no. 1, pp. 93–109, 2015, doi: 10.20885/tarbawi.vol8.iss1.art6.
[53] G. N. Marks, "Demographic and socioeconomic inequalities in student achievement over the school career," *Aust. J. Educ.*, vol. 58, no. 3, pp. 223–247, 2014, doi: 10.1177/0004944114537052.
[54] M. H. Davis, S. Capobianco, and L. A. Kraus, "Measuring Conflict-Related Behaviors: Reliability and Validity Evidence Regarding the Conflict Dynamics Profile," *Educ. Psychol. Meas.*, vol. 64, no. 4, pp. 707–731, 2004, doi: 10.1177/0013164404263878.