Independent Registration Exam Question Item Analysis for the State Islamic Higher Education New Students

By

Moh. Sahlan
UIN KHAS Jember, Indonesia
Email: mmohsahlan7@gmail.com

Luluk Mauli Diana
UIN KHAS Jember, Indonesia
Email: lu2kdiana@gmail.com

Asnawan
University of Al Falah Assunniyyah Jember, Indonesia
Email: asnawan@inaifas.ac.id

Abstract

This study aims to analyze the quality of the registration exam questions for new students at the State Islamic University of Kiai Haji Achmad Siddiq (UIN KHAS) Jember. This research approach is descriptive quantitative evaluative involving 300 prospective new students. Data analysis item of this question utilizes the *Anates* application program. The results show that the *first* of 87 questions have 76 items (87%) in the difficult category, 11 items (13%) in the medium category, and 0% in the easy category; *second*, from the 87 *item* questions, they contain 12.5% with very bad discriminating power, 63% bad, 21% sufficient, 0% good and very good; *third*, the alternative answers as distractors functioned well, but there were 6 question items whose distractors were less effective. It can be concluded that the items compilation are in a poor quality, therefore the revisions are needed if the items tend to be reused, while questions with good categories can be archived in the registration exam question bank.

Keywords: Anates Application, Independent Registration Exam, Question Item Analysis

I. Introduction

Entering a State Islamic Higher Education, a prospective student must compete to be accepted by taking three routes. *First*, through the academic achievement path under the coordination of the National Committee of the Ministry of Religion. *Second*, through the registration exam organized by the Ministry of Religion in the form of a written test in paper form or using a computer or a combination of the two, religious competence is carried out jointly under the coordination of the national committee. *The third* is through the selection process carried out by each PTKIN. If prospective students do not accept through the SPAN-PTKIN and UM-PTKIN pathways, then prospective students still have the opportunity to take part in the selection through the Independent Registration Exam, whose implementation system is regulated by each university (PMA Number 17 of 2017).

UIN KHAS Jember in carrying out self-selection to prospective students through written tests and interview tests. The written exam was held after the interview exam in the form of a reading and writing test of the Qur'an, and the practice of worship and religion. Each prospective student is required to take the exam to determine whether to be accepted as a student or not.

In carrying out the written test, the committee uses an online-based test with the term *computer-based test (CBT)*, participants are given questions that have been prepared by the committee in the form of multiple-choice as many as 87 questions, in which there are Academic Potential Tests, Islamic Insights, and Languages. The test takers are given 90 minutes to work on these questions, with the hope that prospective students have the knowledge and skills needed to attend lectures at UIN KHAS Jember.

Selection tests are usually related to educational decisions that must be made to accept or reject those who are interested in entering a study program. Considering that the selection of new student admissions at UIN KHAS Jember is carried out every year, it is necessary to have a question bank. Especially the question bank which is used to select new students, has a very big advantage for the development of the test, especially if the test is carried out periodically.

Milman & Arter (1984) revealed that the question bank will be very useful if one of the following conditions: (1) there is no ready-to-use test, (2) the administration of the test requires more than one test kit, (3) the bank system questions allow relatively experienced people to create high-quality tests. Items that can be entered into the question bank are items of high quality, namely items that are accepted (passed) based on the results of the analysis that has been carried out (Dit PMU Dirjen Dikdasmen: 2000).

In the test item analysis guide it is stated that the benefits of analysis include: (1) determining which tests are not functioning well; (2) improving the quality of test items including the level of difficulty, discriminatory power, and distractors, as well as improve learning through test doubts and certain skills that make students find it difficult (Aiken, 1994 in the Ministry of National Education, 2008: 2). Item analysis was carried out to test the level of feasibility of each item based on the level of difficulty and discriminating power of the question because not all item items should be considered worthy of use. Determining the revision of a question item is not solely based on the magnitude of the index of difficulty level and differentiating power of the questions, but also on the distribution of the frequency distribution of the answers provided, in other words, it is necessary to analyze the effectiveness of distractor items for each item of the question.

The question items compiled from year to year that are used for the independent registration exam have never been analyzed, nor have they been tested before being used, so their quality cannot be identified. This is supported by an initial study conducted by researchers that of the 1600 prospective students who took the independent registration examination, the highest pure score obtained was 50.00, only 10 percent of the number who took part in the selection. For example, if the committee uses a standard reference approach, the standard value is set to at least 60, and no prospective students will certainly be accepted. The question is whether the questions prepared by the lecturer are of good quality, not good or the ability of the prospective participants who take part in the selection is not good, so this is where it is necessary to analyze the item questions. By analyzing item items, accurate information will be obtained about the extent of validity, difficulty index, discriminating power, and reliability of each of the test items. In other words, whether the item has met the criteria for a good question or not. Making academic test results the main element in determining the acceptance of prospective new students, certainly has an impact on the provision of quality questions. However, so far the questions that have been made have only been collected and stored, and comprehensive analysis has never been carried out.

II. Literature Review

This study compares the accuracy of admission systems at the State Islamic Higher Education Institution (PTKIN) in predicting student achievement. Samples were drawn from five PTKIN in Indonesia. To measure students' academic achievement, GPA data at the end of semester 2 was used. It was taken from student admission channels as selected in 2015, namely SPAN-PTKIN, UM-PTKIN, and Mandiri. Academic data was also taken on the number of prospective students who register after admission through the three channels. Descriptive statistical analysis and inferential statistical analysis techniques were used. ANOVA was applied to examine the differences of academic achievement by students received through the three channels, utilising SPSS. The results proved that the prediction of students' achievement rates on PTKIN students, received through the SPAN-PTKIN channel, is higher and more effective than those received through the UM-PTKIN and Mandiri channels. Further, the Mandiri channel has the most effective registration of prospective students compared to SPAN-PTKIN and UM-PTKIN.

In relation to the results of the study, Stemler has used data on First Year Grade Point Average (FYGPA) as an indicator of academic achievement in relation to the entrance exam for higher education. In fact, until now, the size of academic success of FYGPA is still maintained. (Stemler, 2012: 5–17).

Hence, the results of research conducted by Koretz et al. show that the SAT and HSGPA (High School Grade-Point Average) test scores are still the basis for considering student academic achievement (2014). Some previous literatures uses certain pre-college characteristics as predictor variables to foretell the success of college student studies. Yorke points out that pre-college factors influence the success of college student studies, such as the academic achievement index, and efforts to improve student academic achievement (1998: 189–201). Previous studies often test the variables of pre-college characteristics as predictor variables that influence the success of first year college students. There are three pre-college characteristics, namely the student's background, self-perception as to ability, and achievement orientation and motivation (Bauer & Liang, 2003: 277–290)

As identified by Terenzini, there are 6 (six) pre-college characteristic factors that can be used as predictor variables to measure the success of student study in the first year: (1) the score of school achievement; (2) gender; (3) scholastic talent test scores (SAT); (4) ethnicity; (5) family care education; and (6) family income levels (1984: 178–194). A pre-college characteristic, the student's background may be the value of academic achievement in school such as the score of the school exam. Previous studies have shown that the score of school exams, which students acquire before entering college, significantly predicts the academic achievement of college students (Daugherty & Lane, 1999: 355–362). In fact, the academic achievement of the school (SMA), such as the score of the final exam (national exam), student report cards, etc. can be used as variables to predict the success of college students. (Terenzini, 178–194 1984)

The research of Evans et al. also consistently shows that GPA for undergraduate students (UGPA) is a good predictor of the success of graduate students. The following studies give examples: (1) a combination of GRE and GPA from graduate programs is a strong predictor of the academic success of both postgraduate and doctoral students; (2) UGPA is the most important and significant predictor of overall academic performance (Evans et al., 2007: 544– 567); (3) UGPA is the most valid predictor and has the most significant relationship to student success (Omizo et al, 1997: 947–953); (4) GRE and UGPA are generally valid predictors of the first year S2 GPA and graduation GPA, of graduate programs (Kuncel et al, 2001: 162–181).

Based on those studies, the quality of the registration exam questions for new students at the State Islamic University of Kiai Haji Achmad Siddiq (UIN KHAS) Jember. This research approach is descriptive quantitative evaluative involving 300 prospective new students. Data analysis item of this question utilizes the *Anates* application program. The results show that the *first* of 87 questions have 76 items (87%) in the difficult category, 11 items (13%) in the medium category, and 0% in the easy category; *second*, from the 87 *item* questions, they contain 12.5% with very bad discriminating power, 63% bad, 21% sufficient, 0% good and very good.

III. Methodology

This research approach uses a quantitative, descriptive type and evaluation design intending to know the quality of the items for the Independent Registration Exam. The subjects of this study were 300 prospective students of UIN KHAS Jember. Data was collected through the documentation contained in the Computer-based *Test* (CBT) application in the form of questions, answer keys, and answers from prospective students. The data collected were analyzed using the Anates, to know the discriminatory power, level of difficulty, and the effectiveness of the distractor. The following is an item difficulty index (Sahlan, 2015).

Table 1. *Item Question Difficulty Index*

No. P Score	Interpretation
1 0,00 - 0,30	Difficult
2 0,31 - 0,70	Medium
3 0,71 - 1,00	Easy

One of the requirements for a good test instrument must look at the power of discrimination (discrimination). The distinguishing power of a test item aims to find differences between test-takers who have high abilities and those who have low abilities (Sahlan, 2015).

Table 2. *Item Question of Distinguishing Power Index Criteria*

DP Score	Interpretation
Negative Sign	Worst
0.00 - 0.20	Bad
0.21 - 0.40	Sufficient
0.41 - 0.70	Good
0.71 - 1.00	Best

In addition to having different power requirements, you must also pay attention to the level of functioning of the distractor. Each type of multiple choices test has one question and several alternative answers. A good distractor is how much the wrong choice can deceive the test takers who do not understand the available answer keys. The more test takers choose a distractor, the distractor can play its function well (Iskandar & Rizal, 2017). A good distractor is an alternative answer that is avoided by test takers who are intelligent and chosen by participants who are less intelligent (weak). Or in other words, a good item is an alternative to distracting answers that the test takers choose evenly. On the other hand, items with bad distractors were not chosen evenly. The distractor works well when the distractor is at least 5%

of test-takers or more chosen by low-ability groups (Daryanto, 2012: 193).

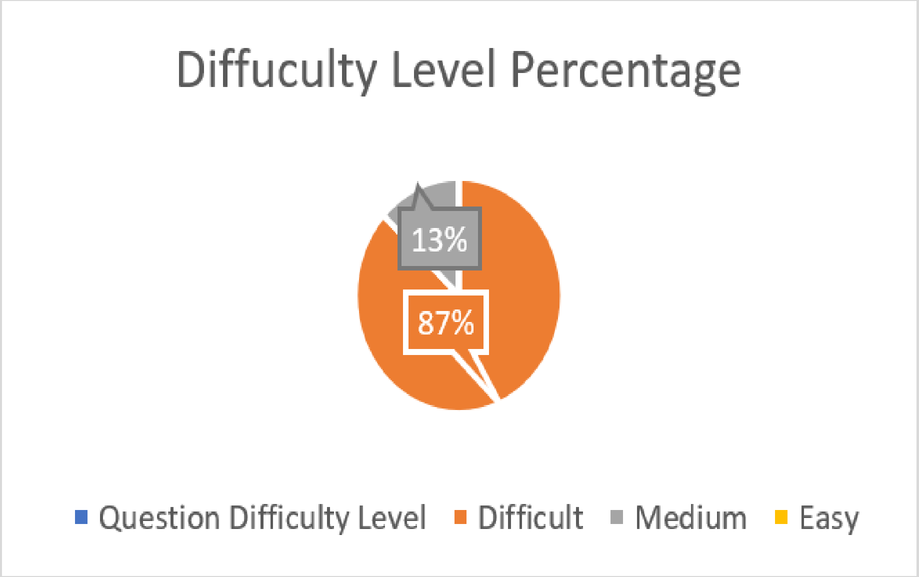
IV. Results and Discussion

Test Difficulty Level Analysis

The item difficulty level is the proportion of how many test participants choose the correct answer to each item so that it can be seen whether the item is classified as easy, medium, or difficult.

Table 3. *Exam Questions of Independent Registration Difficulty Level*

No	Criteria	Questions Number	Total	Percentages
1	Difficult	1, 2, 3, 4, 5, 6, 8, 9, 10, 11, 12, 13, 15, 16, 17, 18, 19, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 45, 46, 47, 48, 49, 50, 52, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 83, 84, 85, 86, 87	76	87%
2	Medium	7, 14,20, 31, 44, 51, 53, 54, 65, 73, 82	11	13%
3	Easy	-	-	0%



Picture 1. *Independent Registration Exam Questions Difficulty Level Percentage*

In table 3 and Picture 1, information is obtained from 87 multiple choice tests, there are 76 items (87%) with a difficulty level of questions in the difficult category (0.00 – 0.30), 11 items (13%) in the medium category (0 .31 – 70), 0 (0%) in the easy category (0.71-1.00).

From the description above, it can be explained that the Independent Registration Exam questions for UIN KHAS Jember which were prepared by the lecturers have a poor quality of difficulty level because to obtain good results, the proportion between the levels of difficulty should be normally distributed. As said by Arifin (2009), these proportions can be determined: *first*, 25% difficult questions, 50% medium questions, and 25% easy questions, or *secondly*, 20% difficult questions, 60% medium questions, and 20% easy questions; or *Third*, 15% difficult questions, 70% medium questions, and 15% easy questions. This depends on the decision of the question-making team meeting or following the technical instructions for writing questions.

Several factors are suspected to be the cause that the questions prepared by the lecturers with the quality level of difficulty are not good when viewed from the proportion of the distribution of difficult, medium, and easy difficulty levels: *First*, the determination of the material being tested is not taken from the Senior High School curriculum but submitted it is entirely up to the lecturers who are in charge of the courses so that the standard of material can be used in higher education courses, and *secondly*, the lecturers who make the questions do not understand the proportion of the distribution of the level of difficulty of the questions as mentioned above.

Based on the theory that the questions are very difficult or very easy, it does not mean that they should not be used. This depends on what the question is used for. If it is used for the selection of registration exams where there are many participants, while the quota is taken a little, then the proportion of difficult questions is higher. On the other hand, if there are few registrants or examinees, then we choose questions that are in the easy category. In addition, difficult questions will increase the enthusiasm for learning for high-ability participants, while very easy questions will increase motivation for low-ability participants (Arikunto, 2005).

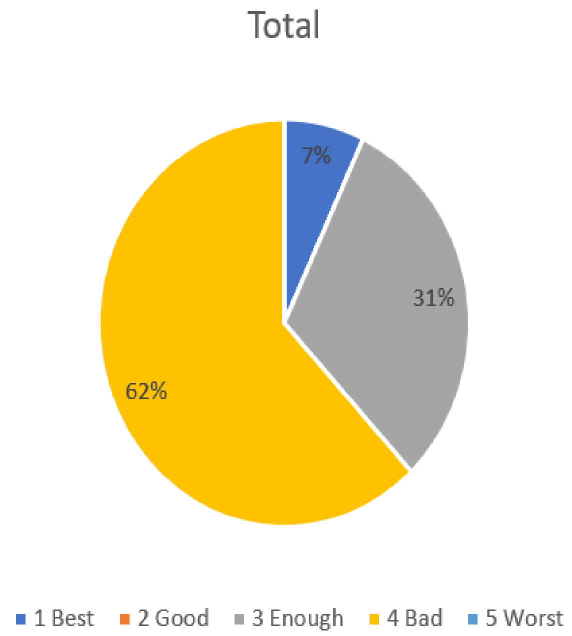
Concerning the level of difficulty of this question, the question writer must pay attention to several things, namely: (a) items whose category is very difficult or very easy, it is possible that they will not provide useful information for the majority of test-takers. Therefore, questions like this can be distributed among the answer choices that do not meet the criteria; (b) if the questions are classified as very difficult or very easy, but each distribution of answers as a distractor on the question shows answers that are evenly distributed, logical and have negative discriminating power (except for the key), then the questions are still eligible to be used; (c) if the criteria for the item which are very difficult or very easy, but have distinguishing power and effectiveness of distractors are met, then the item can be used and accepted as an alternative to be stored in the question bank and can be used for the next test; (d) if the questions that are classified as very difficult or very easy, the distinguishing power and effectiveness of the distractors do not meet the predetermined provisions, then the item needs to be revised and tested again (Arifin, 2009: 279) so that the question can be entered into in the question bank.

Level Analysis of Difference Power

Distinguishing power or often called discriminatory power is a question that can distinguish prospective participants who have intelligent abilities and prospective participants who are less intelligent. The results of calculations based on the interpretation category are less than 0.20 in the bad category, 0.20 - 0.40 in the sufficient category, 0.40 - 0.70 in the good category, 0.70 - 0.100 in the very good category, while the with a negative sign (-) the category is very bad.

Table 4. *Question Item Power Difference*

No.	Criteria	Item Questions Number	Total	Percentages
1	Best	-	0	0%
2	Good	-	0	0%
3	Sufficient	1, 9, 22, 24, 25, 26, 31, 32, 34, 37, 40, 41, 48, 49, 53, 54, 59, 60, 61, 62, 64, 67, 70, 79, 81, 82, 86,	27	31%
4	Bad	2, 3, 4, 5, 6, 7, 8, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 23, 27, 28, 29, 30, 33, 35, 38, 39, 42, 43, 44, 45, 46, 47, 50,51, 57, 58, 63, 65, 68, 69, 71, 72, 74, 76,77, 78, 80, 83, 84, 85, 87	54	62%
5	Worst	36, 55, 56, 66, 73, 75,	6	7%



Picture 2. *The Percentage of Exam Questions Registration Distinguishing Power*

Based on table 4 and figure 2, information is obtained that from 87 multiple-choice items there are 6 items (7%) with very poor discriminating power, 54 items (62%) have poor discriminating power; 27 items (31%) have sufficient discrepancy, while good and very good 0 items (0%). So, in general, the level of discriminatory power of questions is categorized as bad. The results of this study are not following the theory which says that one of the analyzes that must be carried out to determine the quality of a good item is to analyze the level of discriminatory power of the questions.

Suharsimi (2002: 211) states that the discriminatory power of the questions is the ability of the questions being tested to distinguish between high-skilled candidates and low-skilled candidates. If the questions can be answered correctly by all potential participants, both those with low abilities and those with high abilities, then the question can be categorized as poor because the item does not have distinguishing power. Likewise, if the questions that cannot be answered by the candidates with high and low abilities are also included in the category of questions that are not good because the questions also do not have distinguishing power, while the questions in the category of good distinguishing power are questions that can be answered correctly by the prospective participants. only the high-ability test, while on the low-ability many answered incorrectly.

Based on the explanation above, the results of this study can be said that the items prepared are questions that are of less quality. This means that the question has not been able to distinguish between high-skilled candidates and low-skilled candidates, because on average 19.3% are classified as bad. Thus, it can be concluded that the question of the Independent Registration Examination of UIN KHAS Jember cannot be said to be a good measuring tool because this question has not been able to carry out its function as a distinguishing power. Therefore, this problem needs to be improved by increasing the power of discrimination. Questions with poor and very bad discriminating power should be discarded or revised, and questions with sufficient discriminating power can be entered into the question bank.

Based on the results of the analysis of the items for the Independent Registration Exam using the *Anates* 4.0 application, from a total of 87 items of distractor level questions, it can be shown in the following table.

Table 5. *Questions Distribution Based on Distractor Alternatives*

Distractor		Question	Total	Percentages
Best	Alternative 3	1, 2, 3, 4, 5, 6, 9, 11, 13, 15, 16, 17, 18, 19, 21, 22, 26, 27, 28, 32, 33, 34, 35, 38, 40, 42, 44, 45, 46, 48, 57, 59, 60, 61, 63, 64, 65, 66, 67, 69, 70, 71, 75, 76, 77, 82, 83, 84, 87	49	56,32%
Good	Alternative 1	7, 8, 15, 20, 23, 24, 25, 37, 39, 43, 47, 49, 50, 51, 52, 56, 58, 62, 68, 72, 74,	22	25,29%
	Alternative 2	10, 14, 29, 30, 32, 36, 41, 53, 54, 55, 79,81	12	13,79%
Less than Good	Alternative 1	12, 78, 80, 85	4	4,60%
Not Good	-	-	-	-

The distribution pattern of answers can be obtained by counting the number of prospective examinees who choose alternative answers A, B, C, D, or who do not choose an alternative answer at all, which is commonly called omit. The results obtained from the CBT application system for the answer sheets of prospective examinees are known that not all prospective students answer all the questions (omit). From the pattern of distribution of answers, it can be seen whether the distractor can function properly or not. A well-functioning distractor can meet at least 5% of all test takers selected. For all prospective students who take the independent registration exam as many as 300 prospective examinees as samples.

The results showed that 49 items (56.32%) had three distractor alternatives that functioned very well, 22 items (25.29%) had one good distractor alternative and two good distractor alternatives 12 items (13.79%), and one alternative 4 items are not good (4.60%), while those who have bad distractors do not exist (0%). In general, it can be said that the quality of the distractor alternative is very good.

An alternative distractor can be said to be effective when there are 5% selected participants for the registration exam. This is following the opinion of Sudijono (2012) that the distractor has functioned properly if the alternative answer as a distractor has been selected by at least 5% of all test participants. Question items whose alternative distractors work well can be reused for the next year's registration exam.

The distractors that did not function properly in this study amounted to 4 answer choices, namely items 12, 78, 80, and 85. This means that the distractors in the item questions do not have a high interest in the registration exam participants who do not understand the concept or lack mastery of the material. exam so that they choose the correct answer. According to Purwanto (2009), the purpose of the distractor is to mislead the test-takers so that they do not choose the answer key. Distractors attract the attention of test-takers who do not understand the subject matter to choose it. For the distractor to function properly, the distractor must be created and arranged as closely as possible with the correct answer key. Distractors who cannot carry out their functions properly because they are too conspicuous and understood by all test takers as distractors are recommended to be revised.

V. Conclusions and Recommendations

The analysis of items for the independent registration exam at UIN KHAS Jember involving 300 prospective students who took part in the selection shows that the quality of the questions is not good and is not suitable for use in the next year's independent registration exam, so revisions are needed if the question items are to be collected in the question bank and reused. This is evidenced from the analysis carried out using the *Anates* there are 76 item questions out of 87 items or 87% are in the difficult category, 11 item items or 13% are in the medium category, and the easy category is 0. The level of distinguishing power is 63% in the category bad. However, the answer's alternative distractor has worked fine. This can be shown by looking at the distribution pattern of the alternative answers of the participants for the registration exam; only 6 question items whose distractors are not functioning properly. Therefore, it is recommended that before compiling questions, it is necessary to have a workshop or training in preparing quality questions based on *Higher Order Thinking Skills* (HOTS). With good quality questions, prospective students who receive at UIN KHAS Jember are qualified. In addition, questions that have been made before being used should be tested so that their quality can be known.

VI. Limitations

The research that has been done over some time only focused on the State Islamic Jember. This research is also concerning itself Arabic learning. While in the present moment the learning process is conducted as offline and online learning. This research is also limited to the Islamic Universities only because non-Islamic Universities.

VII. Acknowledgments

This research would never be able finished without some contributions from five universities. one universities allowed us to conduct research in their places to dig more data and find solutions.

References

Anastasi, A & Urbina, S. (1997). *Psichologicsl Testing and Measurement.* Boston: Allyn and Bacon.

Arifin, Zainal. (2009). *Evaluasi Pembelajaran: Prinsip, Teknik, Prosedur.* Jakarta: Departemen Agama RI.

Abdul Muhid, Burhanuddin, Amiruddin, Bunyamin M Yapid , Suci Ayu Kurniah Putri. (2019). A Comparison of Admission Systems, in Predicting Students' Academic Achievement in State Islamic Higher Education Institutions (PTKIN*), International Journal of Innovation, Creativity and Change. www.ijicc.net Volume 9*, Issue 11.

Burhan Nurgiyantoro. (2001). *Penilaian Dalam Pengajaran Bahasa Dan Sastra*, (Yogyakarta: BPPE).

Bauer, K.W. and Liang, Q., (2003). The effect of personality and precollege characteristics on firstyear activities and academic performance. *Journal of College Student Development, 44*, 3, 277–290.

Depdiknas. (2008). *Panduan Analisis Item Soal. Jakarta: Direktorat Pembinaan SMP.*

Dirjen Belmawa. (2019) *Panduan Penyusunan Kurikulum Pendidikan Tinggi Di Era Industri 4.0.* Jakarta: Dirjen Belmawa Kemenristek Dikti.

Gregory, R. (2007). *Psychological testing: history, principles, and applications (5th ed.)*. New York: Pearson Education Group, Inc.

Gronlund, N.E & Waugh C.K. (1995), *Assessment of Student Achievement.* New Jersey.

Haryati, Mimin. (2007). *Model dan Teknik Penilaian pada Tingkat Satuan Pendidikan.* Jakarta: Gaung Persada Press.

Iskandar, A., & Rizal, M. (2017). Analisis kualitas soal di perguruan tinggi berbasis aplikasi TAP. *Jurnal Penelitian dan Evaluasi Pendidikan, 21* (2), 12–23.

Kemendikbud. (2015). *Panduan Penilaian Untuk Sekolah Menengah Atas.* Jakarta: Dirjen Dikdasmen.

Kemendikbud. (2017). *Panduan Penilaian Oleh Pendidik dan Satuan Pendidikan Untuk Sekolah Menengah Pertama.* Jakarta: Dirjen Dikdasmen.

Kemenristekdikti (2018). *Kementerian Riset Teknologi dan Pendidikan Tinggi Republik Indonesia. SNMPTNSBMPTN 2018 Diikuti 85 PTN. https://ristekdikti.go.id/snmptnsbmptn-2018-diikuti-85ptn/#Diakses Tanggal 1 Maret 2018.*

Keputusan Dirjen Pendis Nomor 3751 Tahun 2018 tentang *Petunjuk Teknis Penilaian Hasil Belajar pada Madrasah Aliyah*, Jakarta: Kementerian Agama.

Kusairi & Suprananto. (2012). *Pengukuran dan Penilaian Pendidikan.* Yogyakarta: Graha Ilmu.

Kuncel, N.R., Hezlett, S.A. and Ones, D.S., (2001). A comprehensive meta analysis of the predictive validity of the graduate record examinations: Implications for graduate selection and performances. *Psycholigical Bulettin, 127* (1), 162–181.

Leclercq, Dieudonne A. &. Bruno, J. E. (1992). Proceedings of the NATO Advanced Research Workshop on *Item Banking: Interactive Testing and Self-Assessment*, held in Ltege, Belgium, October, 27-31.

Lembaran Negara, Peraturan Menteri Agama Nomor 17 Tahun 2017 tentang Perubahan Atas Peraturan Menteri Agama Nomor 74 Tahun 2015 tentang Penerimaan Mahasiswa Baru Program Sarjana Pada Perguruan Tinggi Keagamaan Islam Negeri.

Linn, R.L & Gronlund, N.E, (1995). *Measurement and Assessment Teaching*, Englewood Cliffs, NJ: Prentice-Hall.

Mardapi, Djemari. (2008). *Teknik Penyusunan Instrumen Tes dan Nontes.* Yogjakarta: MItra Cendikia.

Mehrens, William A. & Lehmann, Irvin J. (1991). *Measurement and Evaluation In Education And Psychology*. Amerika: Primed in the United States of America.

Millman, J & Arter, J.A. (1984). Issues In Item Banking. *Journal of Educational Measurement, 21* (4), Winter 1984, Pp. 315-330.

Mustahdi. (2019). *Modul Penyusunan Soal HOTS PAI*. Jakarta: Direktorat Pembinaan Sekolah Menengah Atas.

Nitko, A. J., & Brookhart, S. M. (2007). *Educational assessment of students*. New Jersey: Pearson Education.

Panduan Pengembangan dan Manajamen Sistem Bank Soal, Dit PMU Dirjen Dikdasmen: 2000.

Ratnawulan, Elis dan Rusdiana. (2015). *Evaluasi Belajar*. Bandung: CV Pustaka Setia.

Safari. (2008). *Analisis Item Soal, Asosiasi Pengawas Sekolah Indonesia.* Depdiknas, Jakarta: CV Purnama.

Sahlan, Moh. (2015). *Evaluasi Pembelajaran*. Jember: STAIN Press.

Suharsimi, A. (2011) *Dasar-dasar Evaluasi Pendidikan.* Jakarta: PT. Bumi Aksara.

Surapranata, S. (2004). *Analisis, Validitas, Reliabilitas dan Interpretasi Hasil Tes: Implementasi Kurikulum 2004.* Bandung: Remaja Rosdakarya.

Zainul, A. dan Nasution, N. (2001). *Penilaian Hasil Belajar*. Jakarta: PAU-PPAI UT.

Stemler, S.E. (2012). What should university admissions tests predict?. *Educational Psychologist, 47* (1), 5–17.

Terenzini, P.T., Theophilides, C. and Lorang, W., (1984). Influences on students' perception of their personal development during the first three years of college. *Researching Higher Education, 21*, 178–194.