Ann PUBLISHER
Journal of Education Sciences (Edusci)
e-ISSN : 3047-2296
p-ISSN : 3032-7393

Analysis of the Level of Difficulty and Differentiating Power of the Final Semester Assessment Made by Teachers of Class XI High School Chemistry Subjects in Cirebon in the Odd Semester of the 2022/2023 School Year

Dewi Nurdiyanti[1], Tania Nurlia Hermawati[2], Mutiara Dwi Cahyani[3]
[1]University of Muhammadiyah Cirebon, West Java, Indonesia. Email: dewinurdiyanti@umc.ac.id
[2]University of Muhammadiyah Cirebon, West Java, Indonesia
[3]University of Muhammadiyah Cirebon, West Java, Indonesia
Corresponding Author: Dewi Nurdiyanti (dewinurdiyanti@umc.ac.id)

Abstract. This study aims to determine the quality on the Odd Semester Final Assessment (PAS) of Grade XI Chemistry Subjects in SMAs in the Cirebon Region for the 2022/2023 Academic Year. This research is a quantitative description research with data collection methods using document analysis to obtain data. The sample of this study was grade XI students. Quantitative research was conducted with the help of the ANATES software program version 4.09. the results showed: (1) In terms of Distinguishing Power, the highest value is in school E (77%) and the lowest value is school C (50%). (2) In terms of Level of difficulty there is a proportional comparison in school B with a ratio of 3:5:2.

Keywords: Difficulty Level, Discrimination Power, Item Analysis

INTRODUCTION

Within the field of education, evaluation activities play a crucial function. In order to assess student's learning progress over a predetermined period and their comprehension of the subject matter, teachers typically conduct tests and non-tests as part of their evaluation activities. Teachers require assessment tools in order to complete this task. Teachers must be proficient in the creation and understanding of evaluation instruments. Assessments of student learning outcomes, such as learning outcomes examinations, are a standard evaluation method. Evaluating schools is primarily done to ascertain the extent to which learning objectives and curriculum are met (Monica S et al., 2019).

Test scores provide a means of tracking the evolution of educational quality. Good test results have the least amount of measurement error possible. These tests are conducted for this aim. Random and systematic errors are the categories into which this measurement mistake falls. Unpredictable mistakes in assessing the test sample's content and emotional swings in people—

including the examiners—when the exam response sheet is manually reviewed are the sources of random errors. Despite this, test items that are overly basic or complex can lead to system faults. Although some teachers consistently assign easy tests, others frequently assign excessively challenging assessments. Furthermore, some instructors are costly to grade, while others are giving and generous. Errors in the system originate from these things (Idrus, 2019, p. 931).

Multiple-choice exams are a valuable tool for assessing a range of thinking skills, from basic knowledge recall to higher-level skills like application, analysis, synthesis, and assessment, as outlined in Asrul's (2014) handbook on learning evaluation. The effectiveness of these exams relies on the quality of the question items used. Analyzing these items can provide insights into whether they should be discarded, revised, or reused as is (Arikunto, 2010). The learning outcome assessment points can be evaluated from the perspective of differentiation power and difficulty level.

One measure that can reveal a question's quality is its medium, easy, or too challenging degree of difficulty. In academic terms, a question is considered easy if most students can correctly answer it and difficult if most students cannot. The difficulty score is determined by the percentage of pupils who can correctly answer the question. Questions get more accessible to solve, the more pupils can accurately answer them. The more complex the questions are, the more responses you get from pupils who cannot answer.

Item difficulty indexes can range from 00 to 1.00, with 0.00 representing the lowest and 1.00 representing the highest, according to Wellington, who cited Hendrik et al. (2021). Arikunto (2013) states that a modest degree of difficulty is excellent quality. According to Yuniaria (2021), a decent set of questions that compares manageable difficulty levels is as follows: 3:15.2 or 3:4:3 is considered medium-difficult. If a question is too hard, it can make students quit or feel like they cannot solve it. Conversely, if a question is too easy, it will not inspire students to be creative. Complex questions can be altered or amended, but poorly included questions cannot be used again in subsequent exams.

One measure of a question's differentiated power is its ability to discern between students who possess high ability and those who are less bright or low. The discrimination index recognizes a negative sign (-) but not the difficulty index. The discrimination index is the sole metric that displays the degree of the distinguishing power, and it is abbreviated as D. Like 1.00, only the difference. Thus there are three points on the differentiating power, namely:

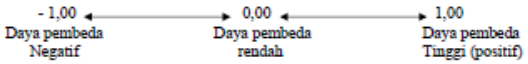

Figure 1. Differentiating Power Index

Because a question cannot be discriminated against, it is not good if intelligent students successfully answer it. Likewise, a question lacks discriminating power if all students—clever or not—cannot respond accurately. Requirements that only brilliant pupils can accurately answer are considered good.

This study intends to evaluate the Final Semester Assessment (PAS) questions at SMA Cirebon to create question items with a level of difficulty in line with the projection at the time of question preparation and with a distinguishing power that can separate students who are smart from those who are less intelligent.

METHOD

This work employed quantitative descriptive research as the research methodology. According to Martono (2010), quantitative research is a study that collects data in numerical values and processes and analyzes these values to derive scientific information. The research was implemented in February 2023 at five high schools located in Cirebon.

The study's population comprises all pupils enrolled in grade XI classes for the 2022–2023 academic year in high schools across the Cirebon Region and defined by the technique of random PLE sampling. Finding a basic, random sample while considering the opportunities available to every member of the research population is known as the random sampling technique. To determine the degree of difficulty and unique selling proposition of each question, the Odd End of Semester Assessment (PAS) for the 2022–2023 academic year served as the data source for this study. The documentation included in the PAS included questions, answer keys, and student answer sheets.

Difficulty Level

A question that falls in between easy and complex is a good one. Students' efforts to tackle problems that are too simple are not prompted by problems that are too easy. On the other hand, because the question is too challenging for them, pupils who find it too tricky will give up and not be motivated to try again. A measure of a problem's ease and difficulty is called the Difficulty Index, according to Asrul (2014). There exists a range of 0.00 to 1.0 for the difficulty index. This

difficulty index indicates the degree of difficulty of the question. An issue is too complex if the difficulty index for the question is 0.0, and it is too easy if the index is 1.0. The formula for finding P is:

$$P = \frac{B}{JS}$$

Information:
P : Difficulty index
B : The number of students who answered the question correctly
JS : Total number of test taker learners

Table 1. Difficulty Index

Magnitude P	Interpretasi
Less than 0,3	Difficult
0,30-0,70	Keep
More than 0,70	Easy

Differentiating Power

The number that shows the magnitude of the discriminating power is called discrimination abbreviated as D. Like the difficulty index, this discrimination index (discriminating power) ranges from 0.00 to 1.00 only the difference is that the difficulty index does not recognize negative signs. A negative sign on the discrimination index is used if something "reversed" indicates the quality of the tester, namely smart children are called less smart and less intelligent children are called smart (Asrul et al, 2014: 153). The formula for determining the discrimination index is:

$$D = \frac{B_A}{J_A} - \frac{B_B}{J_B} = P_A - P_B$$

A : Number of test takers
JA : The number of group participants is high
JB : The number of participants in the group is low
BA : The number of participants in the high group answered the correct questions
BB : The number of participants in the low group answered the correct question

Table 2. Differentiating power index

Differentiating Power Index	Interpretasi
>0,70 – 1,00	Excellent
>0,40 – 0,70	Good
>0,20 - 0,40	Keep
0,00 – 0,20	Bad
ID < 0.00 (Negative)	Discarded/Replaced

DISCUSSION

Finding out the quality of the question items used by class XI chemistry teachers in the Odd End of Semester Assessment (PAS) for the 2022–2023 school year using descriptive quantitative methods was the goal of the research on the analysis of the Final Semester Assessment (PAS) questions for grade XI high school chemistry subjects in Regency and City of Cirebon. The difficulty and distinguishing power of five Final Semester Assessment (PAS) question papers was examined. This means that the researchers assign codes to each of the following: code A for the first inquiry, code B for the second, code C for the second, code D for the second, and code E for the second.

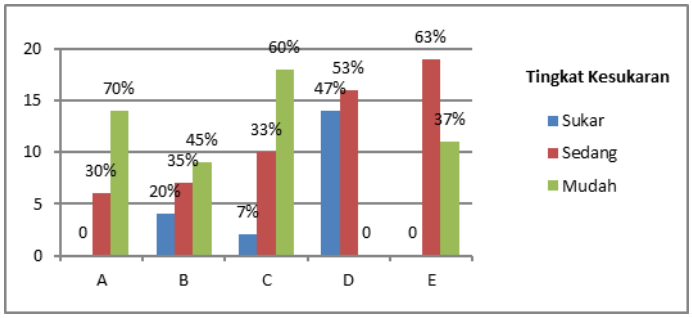

Figure 1. Multiple Choice Question Item Difficulty Diagram

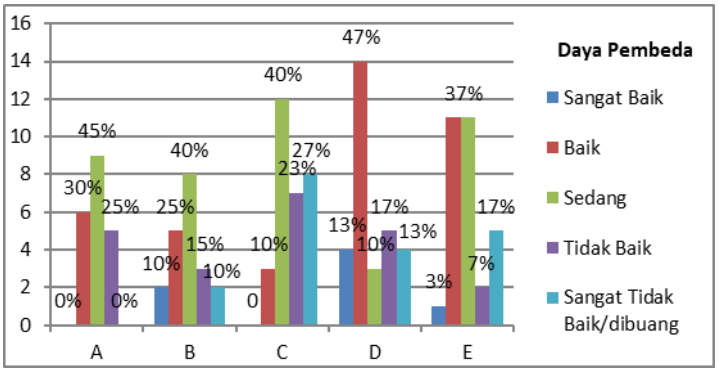

Figure 4.3 Differentiating Power Diagram of Multiple Choice Question Items

Item difficulty indexes can range from 00 to 1.00, with 0.00 representing the lowest and 1.00 representing the highest, according to Wherington, who cited Hendrik et al. (2021). Arikunto (2013) states that a modest degree of difficulty is excellent quality. School B has a comparatively easy, medium, and challenging percentage of 20%, 35%, 45%, or 3: 5: 2. This is the closest to achieving the proportionate difficulty level. That being said, schools A, D, and E have a difficulty that is not proportionate, and none of the indications exist. Yuniaria's (2021)

analysis mentions that a good collection of questions would compare easy, medium, and high-difficulty levels, such as 3:5:2 and 3:4:3.

Easy, medium, and challenging difficulty levels are already proportionally compared in School A, claims Yuniara (2021), to ensure that the exam questions are accurate and fair. A comparison of easy, medium, and challenging questions in school D is 0: 6, 4: 4, 2, with a problematic category of 14 questions (47%) and a medium category of 16 questions (53%) in the study. Since there is no easy category difficulty test, it is evident that this is not proportional. There must be nine simple questions for the test to be proportionate. This means that five medium and four hard questions must be removed from the test for the items to meet the easy criteria.

Like school E, which has a ratio of 3.4:6, 6:0 for easy, medium, and challenging problems, there are no complicated questions. The comparison of difficulty levels has yet to be done in compliance with the rules, which makes the learning outcome test inappropriate. Changes in the test question's proportion are necessary to obtain an appropriate comparison; specifically, nine categories of difficult question items must be added, and seven medium-category questions and two easy-category questions must be subtracted. Interviews with teachers claim that kids utilize their mobile phones to complete the questions. While answering the questions, students can view the online replies. According to recommendations, the exam questions' difficulty level needs to be proportionate, as demonstrated by the research conducted by Laksmi et al. (2021). The study compares manageable, medium, and challenging difficulty levels with a ratio of 2: 4: 1. Arifin (2017: 266) explains that a question can be considered good if its difficulty level is balanced, meaning it is neither too easy nor too challenging.

As can be shown from the analysis's findings, difficulty levels in schools A, B, C, D, and E are all comparable. Schools A and E have 75% and 24%, respectively, that meet the excellent distinguishing criteria. School A has a discrimination index with a medium category for the medium category question items, which are 1, 2, 3, 5, 10, 15, 17, and 20. Questions can still be employed, given their ability to differentiate between the upper and lower groups. A discrimination index of > 0.3 is also present on the excellent category question items, which include questions 4, 9, 11, 12, 18, and 19. Differentiating between the upper and lower groups in the question point is an excellent idea. Either good or medium difficulty characterizes the differentiating capability in the category. To enable maintenance and repurposing of the question items. According to Bagiyono (2017), a differentiating power with a discrimination index 1.0 is

considered good. This finding is consistent with that. However, if D is more than 0.3, the inquiry can be considered practical or valuable.

School C scored 50% on 15 questions and has the lowest percentage of differentiating power. There are fifteen questions, nine of which are not good and six relatively poor. A negative discrimination index indicates that questions 8, 13, 14, 15, 21, and 30 are naughty, and questions 1, 7, 10, 11, 16, 23, 24, 27 and 28 are terrible. Questions with a discrimination index D less than two indicate that all upper and lower groups can correctly answer the question items. As a result, the question is considered to have no discriminating power and is, therefore, wrong according to the discriminating power. Compared to the discrimination index with a negative value, which indicates that the category is inferior, the question items with a negative value are comparatively easy for the lower group while being challenging for the top group. Put otherwise, the reply from the lower group was more accurate than the top group's. The question has an inverse distinguishing power, as the question points show. In other words, from the perspective of the differentiating power, the matter could be better. According to this study, the 60 questions had an average negative discrimination score, meaning 11 shallow and 34 low questions could not be employed. This is consistent with the findings of Ina et al. (2021).

A second study by Yelit (2018) revealed that the question items in the excellent category included:

- As many as 30% or nine questions that were classified as having sufficient discriminating power,
- 20% or six questions that were classified as having sufficient discriminating power and
- 20% or six questions that were classified as very good.

Only talented or intelligent students can accurately answer questions with good discriminating power. The more a problem is capable of differentiating between the upper and inferior groups, the higher its discriminating ability on that question.

According to Rusdiana (2014:177), if the question item fails to differentiate between students in the upper and lower groups, there are several reasons to suspect it. These include an incorrect answer key, multiple correct answers to the same question, a malfunctioning deceiver, material that needs to be more complex or provided, and an unclear measure of competency.

Tests that are too easy or too difficult can inhibit the discriminating power's ability to discern between pupils with high knowledge abilities and those without, according to Surapranata in Prabayanti (2018). Every school that has determined its percentage of

discriminating power has been able to separate pupils who are capable of understanding the content, who are proficient in it, or who have high ability from those who are not proficient in it. Because they are unable to discriminate between students who are less intelligent or who need to comprehend the subject matter, question items that fall into the category of evil and ugly distinguishing power should not be used or replaced. As for question items, those that fall into the categories of medium, reasonable, and very good can be employed again since they can differentiate between groups of intelligent and less intelligent pupils.

The link between Discriminating Power and Difficulty

The question item's ability to discriminate directly correlates to its difficulty level. According to Bagiyono (2017), question items with a difficulty index value of 1 signify correct answers from test takers, while question items with a value of 0 indicate incorrect answers. These question items have low discriminating power and should not be avoided in the next exam. Calculating the distinguishing power of these question items will result in a value of D = 0.

CONCLUSION

From the results of the analysis of the Odd End of Semester Assessment (PAS) class XI questions in SMA throughout the Cirebon Region in the subject of Chemistry, the following conclusions can be drawn:

1. In terms of question types, School A has 0 (difficult question items), 6 (30 medium question items), and 14 (easy question items), comprising 70% of the total. There are four questions in School B's tough question category (20%), seven in the medium question category (35%), and nine in the easy question category (45%). Of the total question types in School C, 2, or 7%, are challenging, 10, or 33%, are medium, and eighteen, or 60%, are accessible. In School D, there are 14 questions (or 47%) that are difficult, 16 questions (or 53%) that are medium, and 0 questions (or 0%) that are easy. There are zero (0%) challenging question item categories in School E, nineteen (63 %) medium question item categories, and eleven (37%) easy question item categories in School E.
2. School A stands out from the competition with a good category score of 30%, a medium category score of 45%, and a lousy category score of 25%. About Distinguishing Power, School B scores 10% in the outstanding category, 25% in the excellent category, 40% in the medium category, and 10% in the naughty category. An outstanding category of 0%, a suitable category of 10%, a medium category of 40%, a terrible category of 23%, and a naughty category of 27% are all represented in School C's Distinguishing Power. About

Distinguishing Power, School D scores well in the outstanding category (13%), good in the category (47%), medium in the category (10%), poorly in the category (13%), and very poorly in the category (17%). An outstanding category of 3%, a suitable category of 37%, a medium category of 37%, a lousy category of 7%, and a naughty category of 17% are all represented in School E's discriminating power.

BIBLIOGRAPHY

Arifin, Z. (2012). *Penenlitian Pendidikan Metode Dan Paradigma Baru.* Bandung: Remaja Rosda Karya

Arikunto, S. (2014). *Evaluasi Pendidikan.* Jakarta: Bumi Aks

Asrul, Ananda, R., & Rosnita. (2015). *Evaluasi Pembelajaran.* Bandung: Citapustaka Media.

Bagiyono. (2017). Analisis Tingkat Kesukaran Dan Daya Pembeda Butir Soal Ujian Pelatihan Radiografi Tingkat I. *Widyanuklida, Vol. 16 No. 1*, 1 - 12

Fatimah, L. U., & Alfath, K. (2018). Analisis Kesukaran Soal, Daya Pembeda Dan Fungsi Distraktor. *Jurnal Komunikasi Dan Pendidikan Islam, Volume 8, Nomor 2*, 37-63

L, I. (2019). Evaluasi Dalam Proses Pembelajaran. *Adaara*, 920-923

Martono, N. (2010). *Metode Penelitian Kuantitatif.* Jakarta: Rajawali Pers

Monica, S., & Sudarman, Y. (2019). Analisis Butir Soal Ujian Tengan Semester Ganjil Kelas Vii Di Smpn 29 Sijunjung. *E-Jurnal Sendratasik*, 1-5

Prastika, Y. D. (2021). Pengaruh Validitas, Reliabilitas Dan Tingkat Kesukaran Terhadap Kualitas Butir Soal Menggunakan Software Anates Di Smkn 3 Bangkalan. *Stkip Pgri Bangkalan*, 1-10

Sukardi. (2011). *Metodologi Penelitian Pendidikan Kompetensi Dan Praktiknya.* Jakarta: Pt Bumi Aksara