Journal of
Data Acquisition and Processing

ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.7632339

THE PUNISHMENT SCORE MODEL TO THE MATHEMATICS LEARNING OUTCOMES OF HIGH SCHOOL STUDENTS IN JAKARTA

Erdawaty Kamaruddin
Universitas Negeri Jakarta
erda_kamaruddin@unj.ac.id

Ivan Hanafi
Universitas Negeri Jakarta
ivan.hanafi@unj.ac.id

Ibnu Salman
National Research and
Innovation Agency
ibnu009@brin.go.id

Lisa Dwi Ningtyas
Universitas Negeri Jakarta Indonesia
lisadwiningtyas5@gmail.com

Abstract

The Punishment score is a scoring model with a penalty: how to get a score on a multiple-choice test by subtracting the score for wrong answers. Penalties are given to educate students, so they only guess the answers to questions they understand, and these guesses can produce wrong and correct answers. To improve the quality of learning, it is necessary to prevent students from guessing answers, including scoring with a penalty. Based on these assumptions, the problem arises 'How is the application of the punishment score model to the mathematics learning outcomes of high school students in Jakarta?" This study aims to find empirical evidence about applying the punishment score model to the mathematics learning outcomes of high school students in Jakarta. The study targets are: (1) to determine the most appropriate number of options on a multiple choice test using the punishment score model and (2) to form an honest character in students. This study uses a quasi-experimental method. The research instrument was a multiple-choice test in Mathematics. There are 30 test items. Data were obtained from a three-option multiple-choice test in the first group and a five-option multiple-choice test in the second group with a punishment score model. Furthermore, the average score fairness index was calculated for each group using the Donlon and Fischer method. The higher the fairness index, the more good the student scores. The results showed that the fairness index of the scores on the five-option multiple-choice test was higher than the

three-option multiple-choice test. So the five-option multiple choice test in the punishment score model is more appropriate to use to improve the quality of learning. In implementing the research, the Donlon and Fischer method can be used to analyze the results of mental measurements by calculating the item difficulty level on the delta scale.

Keywords: punishment score, fairness index score, penalty, multiple choice test

1. Introduction

Assessment is the main element in the learning process. Assessment does evaluate not only students but also all components of the learning process. This is the purpose of the assessment: to determine student learning progress, the effectiveness and efficiency of the various learning components used, and the follow-up of learning (Wilson & Roscoe, 2020). There are many assessment models in the world of education, and there are skills assessments, attitude assessments, and assessments of students' thinking abilities (Öztürk et al., 2020). All of these assessment models are expected to provide an overview of the actual performance of the object being assessed, be it students, teachers, the learning process, or other learning components, because, based on the assessment results, a decision-making process will be carried out.

Assessment is obtained from the results of measurements using measuring instruments in the form of tests and non-tests. The most widely used test today is the multiple-choice test (Milligan, 2020; Seo et al., 2020; van Nguyen et al., 2020). The result of the measurement is a score. Scoring in multiple-choice tests highly depends on the scoring model used, and different scoring models will impact the score obtained. If the correct score model applies, students will speculate on answering by guessing the items they do not master in the hope that the answers are correct (Ossai & Enwefa, 2020; Schlichting, 2022). This happens because the correct score model will only give value to the correct answer without considering the students' guessed or wrong answers. As a result, the values obtained by students are inaccurate or not to the actual abilities of students.

The previous research by Jabbari and Johnson stated that there was a relationship between punishment scores and course-taking in mathematics subjects on long-term student trajectories (Jabbari & Johnson, 2020). In addition, there is research conducted by Dynes et al., which looked for a relationship between punishment scores on a person's behavioral, emotional, and cognitive problems (Dynes et al., 2020). Furthermore, Twardawski, Hilbig, and Thielmann researched the relationship between punishment goals and the possibility of a student misbehaving, such as guessing the answers to questions (Twardawski et al., 2020).

In connection with this thought, it is necessary to study the possibility of scoring deviations in students, including through an analysis of the application of the punishment score model to the assessment process. Furthermore, it is necessary to study how to detect these score deviations, which can be done by calculating the fairness index of scores for each student using the Donlon and Fischer methods. Thus, the purpose of this scientific study is to obtain empirical data about the results of the analysis of the application of the punishment score model to high school students mathematics learning outcomes.

2. Theory Study

2.1 Multiple Choice Test

According to Gronlund, the test is a comprehensive, systematic, and objective evaluation procedure, the results of which can be used by educators as a basis for decision-making in the learning process (Gronlund, 1993). Written tests can be in the form of description and objective tests. Kubiszyn divides objective tests into four forms: true-false tests, matching tests, multiplechoice tests, and short-answer or complete tests (Kubiszyn & Borich, 2007). An objective test is a test in which all the information needed to answer the test is available. The items contain possible answers that students must choose. Popham calls it the selected response test (Popham, 1981). In general, objective tests are presented in multiple choices. Reynolds, Livingston, and Willson revealed, “Multiple choice items are by far the most popular of the selected-response items. They have gained this degree of popularity because they can be used in a variety of content areas and can assess both simple and complex learning outcomes” (Cecil et al., 2009). According to Suryadibrata, objective tests consisting of multiple-choice questions have been known in Indonesia since the 1960s (Sumadi, 1995). Thorndike put forward several provisions in compiling multiple-choice tests, namely: the language used is easy to understand, there is one correct answer for the alternative answers, the items contain material that has been studied, the items are single, not bound to each other, and are stated in sentences that clear (Thorndike, 1997). Multiple choice tests provide more than one wrong answer. In a multiplechoice test, the greater the number of options provided, the smaller the chance of the correct answer being selected by speculation. One of the weaknesses of using multiple-choice tests is the opportunity to speculate or choose available answers randomly (Yu & Wu, 2020). On the other hand, some advantages are that it can be scored easily and quickly, has high objectivity, can be used to measure various cognitive levels, and can cover a wide range of material in a test.

The number of options on the multiple choice test is the number of alternative correct answer choices given in the multiple choice questions. The number of options in the multiple choice test can indirectly affect the scoring results. The more options provided, the smaller the chance of students answering the test item correctly. Likewise, the greater the number of options provided, the more attention the students will take in taking the test.

The items on the multiple-choice test can be analyzed empirically to test their validity and reliability so that the quality of the items can be accounted for. Correcting student answers and scoring can be done easily and quickly and have high objectivity. Items can measure various cognitive levels of students and can cover a wide range of material in one test. This form is appropriate for large-scale exams where the results, such as a large-scale national exam, must be announced immediately. The description shows that compiling a multiple-choice test is not easy; it requires seriousness, precision, high precision, high costs, and a relatively long time, especially in finding distractors in answer choices that are equivalent to the answer key so that they can function appropriately as distractors. Furthermore, minimize opportunities for students to provide answers by guessing.

Multiple choice tests are the most popular type of test because they are widely used in schools and are often used as a selection tool. The advantages of multiple-choice tests include being more practical, can be used for a large enough number of students, can be scored easily and quickly, having high objectivity, can be used to measure various cognitive levels, and can cover a wide range of material in a test. However, this type of test is often considered to be the cause of the low quality of education in Indonesia. The weaknesses of multiple-choice tests are that they cannot be used to evaluate logical and systematic thinking processes, reasoning and analytical skills, writing skills, and expressing ideas through good language skills. There is a tendency to guess in answering multiple-choice items because all the answers are already available, and students can only choose.

These statements emphasize that measurement problems related to multiple-choice tests are issues that need to be studied and developed. Furthermore, the assessment will focus on measuring learning outcomes in multiple-choice tests using the punishment score model.

2.2 The Punishment Score Model

The punishment score model is a way of obtaining a score on a multiple-choice test by giving a penalty in the form of a reduced score on items with wrong answers (Ostrosky et al., 2022). The penalty amount depends on the number of options in the multiple choice test. The punishment score model is expected to minimize the guess factor in the correct score model by clearing out items that cannot be answered (Jacobsen, 2020). As it is known, in the correct score model, the score obtained on the multiple choice test is calculated by adding the scores on the test items with the correct answers only without taking into account the test items with wrong answers or the test items that students did not answer (Babatimehin et al., 2022; LiconaChávez et al., 2020; Riggs et al., 2020). The correct score model assumes that all test items have the same weight, so the correct score model formula only calculates the correct answers. As a result, the correct score model provides an opportunity for students who do not know at all about certain test items to be able to answer these test items by guessing. If the guess is correct and repeatedly occurs on several test items, the student will get a final score higher than his actual ability. That is, the student's score is improper because the correct answer score he gets is not based on his knowledge and ability but because of his guess, which happens to be right. Theoretically, in the correct score model for students with the same abilities but different luck factors, the final scores of students who guess will be higher than the final scores of students who do not guess, and a model like this will reduce the value of item validity, instrument reliability, item differentiability. , and the level of difficulty of the items (Bateson & Dardick, 2020; Collignon et al., 2020; Sahin & Colvin, 2020; Sandeep Prakash & Dattatraya Hanumantrao, 2020).

2.3 The Donlon and Fischer method

An unreasonable score is a score that does not match the student's ability even though all the test items are good. Score irregularities occur when students succeed in answering difficult items and need help to answer easy items. Score unreasonably is very likely to occur in multiple-choice tests, and scoring unreasonably can be detected by calculating an index known as the score fairness index. The fairness index of the score is a number that shows how much the score obtained by the student can describe the student's actual ability. The higher the fairness index, the more good the student's score is. The fairness index of the score can be calculated using the classical measurement theory approach and the modern measurement theory approach using the item responsiveness theory. This study calculated the fairness index of scores using the Donlon and Fischer method through student biserial correlation. The formula used is as follows:

$$\rho = \frac{\mu_\Delta - \mu_\Delta p}{\sigma_\Delta}\,\frac{}{y}$$

Formula description:

ρ = fairness index

μ_Δ = mean item difficulty on the delta scale for items done by the test takers

μ_Δ = mean item difficulty on the delta scale for items correctly answered by the i-th test takers

σ_Δ = standard deviation

p = the proportion of correct answers to the items worked

y = probability on the standard normal probability distribution at point divided by P_{it}

The application of different scoring models will impact the scores obtained by each student (Grissom et al., 2015), because students will consider the possibility of answering or not answering test items that they do not master. If the correct score model applies, students will answer test items they do not master by guessing. Conversely, if the punishment score model applies, students will not answer test items they do not master.

Based on this description, the punishment score model is expected to minimize the guess factor in the correct score model so that the scores obtained by students are by their actual abilities.

3. Research Methods

This study uses a quasi-experimental method. The research instrument was a multiplechoice math test with a punishment score model. The first group of respondents was given a three-option multiple-choice test, and the second group was given a five-option multiple-choice test. Next, the average score fairness index was calculated using the Donlon and Fischer methods for each group and compared to whichever score had a higher level of fairness. This study uses a comparative design: comparing the average scores obtained by students, as shown in Table 1.

Table 1 Comparative Design of Three Options and Five Options Multiple Choice Tests on the Punishment Score Model

Scoring Model	Multiple Choice Test	
	Three Options	Five Options
Punishment Score		
	Fairness index score (μPS3)	Fairness index score (μPS5)

This research was conducted in DKI Jakarta with high school students in class XI IPA as respondents. Data collection used research instruments carried out directly by students to obtain primary data. The research instrument was a multiple choice test with three and five options for Mathematics class XI IPA, totaling thirty items. After compiling the items, a content validation process is carried out through the expert match technique by calculating the percentage of items that match the indicators based on expert judgment. The next step is to perform empirical validation in the form of field testing of the instrument. The validity of the items was tested using the point biserial correlation coefficient between the item scores and the total score of the test, as follows:

$$r_{pbis\,(i)} = \frac{\bar{x}_i - \bar{x}_t}{sd}\sqrt{\frac{p_i}{q_i}}$$

Dragon suggests that the test items are declared valid if the correlation coefficient value is 0.20 or more. After the validity test, the instrument reliability test was carried out using the internal consistency coefficient reliability formula (Kuder-Richardson 20) (Naga, 2012). According to Nitko, a research instrument is declared reliable if it has a high-reliability coefficient (Nitko & Brookhart, 2014), i.e., 0.90 or more. Meanwhile, Guilford wrote the formula as follows:(Guilford, 1982)

N □ 2 □ □ p q
□KR □20 □ A i i
N □1 □ A2

Research activities starting from preparation to data analysis, lasted for approximately seven months, with details of activities as follows: (1) preparation of research proposals, (2) preparation of instruments, (3) validation of instruments, (4) field data collection; (5) data processing, (6) data analysis, (7) article writing, and (8) research report writing.

The population in this study were high school students in DKI Jakarta, spread across five regions: West Jakarta, Central Jakarta, South Jakarta, East Jakarta, and North Jakarta. While the research sample consisted of class XI public high school students in DKI Jakarta. The sampling technique used was purposive random sampling. A sample of seven hundred high school students in DKI Jakarta was obtained. The seven hundred respondents were divided into two groups, each group numbering three hundred and fifty students. The first group was given a multiple-choice math test instrument with three options, and the second group was given the same instrument but with five options. This number is by Gable's recommendation, which stipulates that the sample size is 5 to 10 times the number of items (Gable, 1986). Crocker and Algina also stated that the number of samples is five times the number of items (Crocker & Algina, 1986). Furthermore, Naga explained that in statistics, the criterion of 0.05 is often used as the limit between large and small sample sizes (Naga, 2012). A sample size less than 0.05 of the population size is categorized as a small sample. Meanwhile, Mueller added that there is no definite limit on the number of respondents in item analysis. However, the measurement results will be more stable if 100 respondents are used instead of 10 (Mueller, 1986). This opinion was reinforced by Tabachnick, who suggested that a sample size that was good enough to obtain reliable results was at least 200 respondents (Tabachnick & Fidell, 1989).

4. Research Results and Discussion

4.1. Description of Research Results Data

The research data was obtained from a multiple-choice test instrument in mathematics for class XI IPA, given to two groups of students, with 350 students in each group. The first group was given a three-option multiple-choice test, and the second group was given a five-option multiple-choice test with a punishment score model. Thus, 700 data were obtained and divided into two groups for processing.

Based on the data description analysis results, the index of the reasonableness of student scores in the five-option multiple-choice test group was higher than students in the three-option multiple-choice test group with the punishment score model. A description of the research data as a whole is summarized in Table 4.1.1 and Table 4.1.2 below:

Table 4.1.1 Summary of Student Score Characteristics in the Three-Option and Five-Option Multiple Choice Tests with the Punishment Score Model.

Group	Minimum	Maximum	Average	Median	Modus	Varians	Stand. Dev
SPS3	13	26	20.0200	19.9187	19.1111	7.1772	2.6790
SPS5	13	26	20.0743	20.0846	19.2166	7.0833	2.6614

Table 4.1.2 Summary of the Characteristics of the Fairness Index of Student Scores on the Test Multiple Choice Three Options and Five Options with Punishment Model score

Group	Minimum	Maximum	Average	Median	Modus	Varians	Stand. Dev
PS3	0,3168	0,9952	0,7500	0,7495	0,7362	0,0206	0,1435
PS5	0,4745	0,9976	0,7712	0,7679	0,9549	0,0188	0,1367

4.1.1 Data Description of the Three-Option Multiple Choice Test in the Punishment Model score

Based on the student scores obtained, fundamental statistical values are calculated, summarized in Table 4.1.1 and Table 4.1.2. The relative position of the average student score in the three-option multiple-choice test with the punishment score model is shown in Figure 4.1.1 From this figure. It can be seen that the position of the average score is above the median

(μ_{PS3} > median).

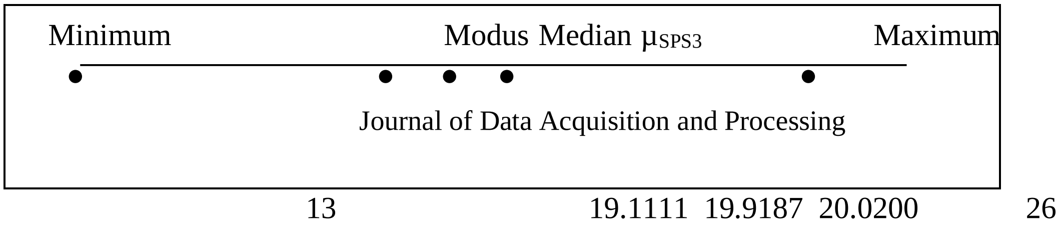

Figure 4.1.1 Relative Position of Students' Average Scores in Multiple Choice Tests Three Options with the Punishment Score Model

The relative position of the fairness index of the average student score in the threeoption multiple-choice test with the punishment score model is shown in Figure 4.1.2. From this figure, it can be seen that the position of the average score is above the median (μ_{PS3} > median).

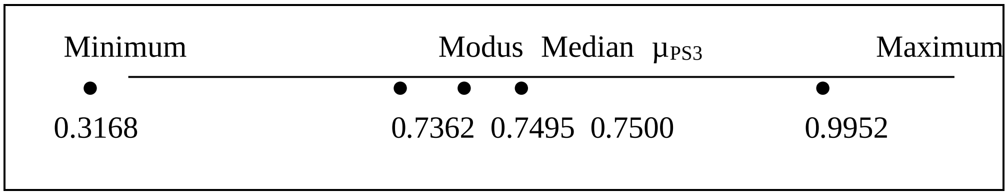

Figure 4.1.2 Relative Position of the Fairness Index of Average Scores of Students on a Three-Option Multiple Choice Test with the Punishment Model score

The relative position of the average student score in the five-option multiple choice test with the punishment score model is shown in Figure 4.1.3. From this figure, it can be seen that the position of the average score is below the median (μ_{SPS5} < median).

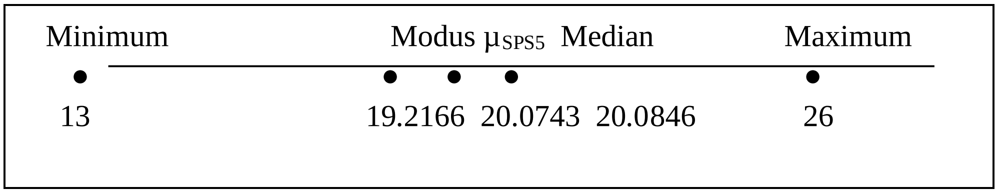

Figure 4.1.3 Relative Position of Students' Average Scores in the Multiple Choice Test of Five Option with Model Punishment Score

The relative position of the fairness index of the average student score in the five-option multiple-choice test with the punishment score model is shown in Figure 4.1.4. From this figure, it can be seen that the position of the average score is above the median (μ_{PS5} > median).

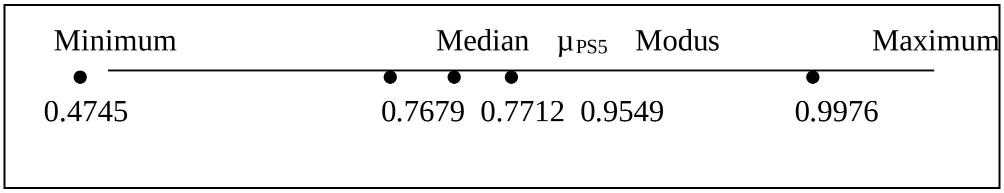

Figure 4.1.4 Relative Position of the Fairness Index of Average Scores of Students on the Test Multiple Choice Five Options with Punishment Score Model

4.2. Results of Testing Requirements Analysis

Before testing the hypothesis, the analysis requirements test is first carried out in the form of a normality test and a variance homogeneity test, as follows:

4.2.1 Normality Test

The normality test determines whether all data groups come from populations with typical distribution characteristics. The normality test in this study was carried out using the Lilliefors test. The summary of the normality test results for all groups of students studied in this study is shown in Table 4.2.1 below:

Table 4.2.1 Normality Test Results for All Respondent Groups

Group	Value		Conlusion
	L_0	$L_{tab\ (0,01)}$	
PS3	0.0548	0.0551	Normal
PS5	0.0549	0.0551	Normal

In Table 4.2.1, it can be seen that all groups of students who were tested for normality gave scores L_0, which is smaller than the value L_{tab} at the significance level α = 0.01. Thus it can be concluded that the index of the reasonableness of scores for all groups of students comes from a normally distributed population so that normality requirements can be fulfilled.

4.2.2 Variance Homogeneity Test

The variance homogeneity test determines whether all data groups tested in this study come from populations with homogeneous variances. The homogeneity test of variance in question is the variance similarity test of all research data groups using the Bartlett test.

Tests were carried out on two groups of sample data: (1) the three-option multiplechoice test group on the punishment score model and (2) the five-option multiple-choice test group on the punishment score model. From the calculation of the variance homogeneity test for the fairness index of multiple choice test scores for all groups of research data, a value was obtained $\square^2_{hit} = 0.0665$ while value $\square^2_{tab} = 6.635$ at the significance level $\alpha = 0.01$. Because of value $\square^2_{hit} \leq \square^2_{tab}$ hence the hypothesis H_0 is accepted. Thus it can be concluded that the sample comes from a homogeneous population. After testing the analysis requirements are met, the research study can proceed to test the research hypothesis.

4.3 Research Hypothesis Testing Results

The research hypothesis: "In the punishment score model, the fairness index of the five-option multiple-choice test scores is higher than the fairness index of the three-option multiple-choice test scores based on the Donlon and Fischer method."

From the testing results, the hypothesis obtained the value t_{hit} = 2. Value $t_{tab} = t_{(0.95)(698)}$ = 1.645 at the significance level $\alpha = 0.05$ and $t_{(0.99)(698)}$ = 2.326 at the significance level $\alpha = 0.01$. If compared, the value is obtained $t_h$at is bigger than the value t_{tab} at the significance level $\alpha = 0.05$, so it was decided to reject the null hypothesis (H_0). Thus it can be said that in the punishment score model, the fairness index of the five-option multiple-choice test scores is higher than the fairness index of the three-option multiple-choice test scores.

Because the index of fairness of the five-option multiple-choice test scores is higher than the index of fairness of the three-option multiple-choice test scores, it can be concluded that in the punishment score model, the five-option multiple-choice test produces a more reasonable score than the three-option multiple-choice test based on the Donlon method. And Fisher.

To show how much the sample mean differs from the null hypothesis mean (H_0), Cohen's d effect size was used. The results of the calculation of the magnitude of the effect size of the two groups of respondents are shown in Table 4.3.1 below:

Table 4.3.1 Effect Size of the Fairness Index of Three Option Multiple Choice Test Scores and the Fairness Index of Five Option Multiple Choice Test Scores in the Punishment Score Model.

Respondent Group	Effect Size
PS3	5,2301
PS5	5,6333

Table 4.3.1 shows that the effect size of the fairness index score is a five-option multiple-choice test (d_{PS5} □ 5,6333) more significant than the fairness index effect size of the three-option multiple-choice test scores (d_{PS3} □ 5,2301) on the punishment score model. This indicates that a five-option multiple-choice test in the punishment score model is more appropriate than a three-option multiple-choice test.

5. Conclusion

Based on this discussion, an empirical generalization is obtained that the number of options in the multiple-choice test influences the level of the fairness index of the scores. The five-option multiple-choice test tends to produce a higher score fairness index than the threeoption multiple-choice test in the punishment score model, using the Donlon and Fischer method.

6. Acknowledgement

The inappropriateness of math scores should be minimized by administering a five-option multiple-choice test using the punishment score model. Based on the study, the test produces a high fairness index score so that it has a reasonable level of student score acquisition and can describe students' actual abilities.

7. References

Babatimehin, T., Chidubem Deborah, A., & Oluseyi Peter, A. (2022). Comparative Analysis of Number Correct and Item Pattern Scoring Methods for Waec and Neco Chemistry Items Among Senior Secondary Students in Osun State. The Universal Academic Research Journal, 4(1), 32–39. https://doi.org/10.17220/tuara.2022.01.04

Bateson, A., & Dardick, W. (2020). A Comparison of the Two-Option Versus the Four-Option Multiple-Choice Item: A Case for Fewer Distractors. Personnel Assessment and Decisions, 6(3). https://doi.org/10.25035/pad.2020.03.005

Cecil, R., Ronald, B. R., & Victor, W. (2009). Measurement and Assessment in Education. Pearson Education Inc.

Collignon, S. E., Chacko, J., & Wydick Martin, M. (2020). An Alternative Multiple-Choice Question Format to Guide Feedback Using Student Self-Assessment of Knowledge. Decision Sciences Journal of Innovative Education, 18(3), 456–480. https://doi.org/10.1111/dsji.12213

Crocker, L., & Algina, J. (1986). Introduction to Classical and Modern Test Theory. Holt, Rinehart & Winston, Inc.

Dynes, M., Knox, M., Hunter, K., Srivathsal, Y., & Caldwell, I. (2020). Impact of education about physical punishment of children on the attitudes of future physicians. Children's Health Care, 49(2), 218–231. https://doi.org/10.1080/02739615.2019.1678472

Gable, R. K. (1986). Instrument Development in The Affective Domain. In Instrument Development in the Affective Domain. Kluwer-Nijhoff Publishing. https://doi.org/10.1007/978-94-015-7259-0_6

Grissom, J. A., Kalogrides, D., & Loeb, S. (2015). Using Student Test Scores to Measure Principal Performance. Educational Evaluation and Policy Analysis, 37(1), 3–28. https://doi.org/10.3102/0162373714523831

Gronlund, N. E. (1993). How to Make Achievement Tests and Assessments. In Allyn and Bacon. Allyn and Bacon.

Guilford, J. P. (1982). Psychometric Methods. In Tata-McGraw Hill. Tata-McGraw Hill. https://doi.org/10.1177/014662168500900316

Jabbari, J., & Johnson, O. (2020). Veering off track in U.S. high schools? Redirecting student trajectories by disrupting punishment and math course-taking tracks. Children and Youth Services Review, 109 (February), 2020–2023. https://doi.org/10.1016/j.childyouth.2019.104734

Jacobsen, W. C. (2020). School punishment and interpersonal exclusion: Rejection, withdrawal, and separation from friends. Criminology, 58(1), 35–69. https://doi.org/10.1111/1745-9125.12227

Kubiszyn, T., & Borich, G. (2007). Educational Testing and Measurement Classroom Practice. John Wiley & Sons, Inc.

Licona-Chávez, A. L., Montiel Boehringer, P. K., & Velázquez-Liaño, L. R. (2020). Quality assessment of a multiple choice test through psychometric properties. MedEdPublish, 9, 91. https://doi.org/10.15694/mep.2020.000091.1

Milligan, J. A. (2020). What Is the Value of Synchronous Engagement in Small Remote Organic Chemistry Classes? Analysis of Multiple-Choice Polling Data from the COVIDImpacted Spring Semester of 2020. Journal of Chemical Education, 97(9), 3206–3210. https://doi.org/10.1021/acs.jchemed.0c00686

Mueller, D. P. (1986). Measuring Social Attitude. In Teachers College, Columbia University. Teachers College, Columbia University. https://doi.org/10.1016/S0271-7123(80)80034-4 Naga, D. S. (2012). Teori sekor pada pengukuran mental. PT Nagarani Citrayasa.

Nitko, A. J., & Brookhart, S. M. (2014). Educational Assessment of Students Sixth Edition. In Pearson New International Edition.

Ossai, P. A. U., & Enwefa, C. (2020). Undergraduate students' achievement in GST 101 using corrected scores and uncorrected scores. Journal of Educational and Social Research, 10(2), 126–133. https://doi.org/10.36941/jesr-2020-0032

Ostrosky, B. D., Reeve, K. F., Day-Watkins, J., Vladescu, J. C., Reeve, S. A., & Kerth, D. M. (2022). Comparing Group-Contingency and Individualized Equivalence-Based Instruction to a PowerPoint Lecture to Establish Equivalence Classes of Reinforcement and Punishment Procedures with College Students. The Psychological Record, 72, 1–32. https://doi.org/10.1007/s40732-021-00495-6

Öztürk, M., Akkan, Y., & Kaplan, A. (2020). Reading comprehension, Mathematics selfefficacy perception, and Mathematics attitude as correlates of students' non-routine Mathematics problem-solving skills in Turkey. International Journal of Mathematical Education in Science and Technology, 51(7), 1042–1058. https://doi.org/10.1080/0020739X.2019.1648893

Popham, W. J. (1981). Modern Educational Measurement. In Prentice Hall. Prentice Hall.

Riggs, C. D., Kang, S., & Rennie, O. (2020). Positive impact of multiple-choice question authoring and regular quiz participation on student learning. CBE Life Sciences Education, 19(2). https://doi.org/10.1187/cbe.19-09-0189

Sahin, F., & Colvin, K. F. (2020). Enhancing response time thresholds with response behaviors for detecting disengaged examinees. Large-Scale Assessments in Education, 8(1), 1–24. https://doi.org/10.1186/s40536-020-00082-1

Sandeep Prakash, N., & Dattatraya Hanumantrao, N. (2020). Effect of surprise test and instruction for negative marking on item analysis in Pharmacology. IP International Journal of Comprehensive and Advanced Pharmacology, 5(1), 19–21. https://doi.org/10.18231/j.ijcaap.2020.005

Schlichting, D. (2022). Corrected Score Functions under Additive Berkson Error. LMU Munich.

Seo, Y., Lee, K., Clavera, I., Kurutach, T., Shin, J., & Abbeel, P. (2020). Trajectory-wise multiple choice learning for dynamics generalization in reinforcement learning. Advances in Neural Information Processing Systems, 2020-Decem(NeurIPS).

Sumadi, S. (1995). Penggunaan Bentuk Soal Pilihan Ganda dalam Ujian. Bulletin Pengujian dan Penilaian, 15.

Tabachnick, B. G., & Fidell, L. S. (1989). Using Multivariate Statistics. In Harper-Collins Publisher, Inc. Harper-Collins Publisher, Inc.

Thorndike, R. M. (1997). Measurement and evaluation in psychology and education. In Prentice Hall. Prentice Hall.

Twardawski, M., Hilbig, B. E., & Thielmann, I. (2020). Punishment goals in classroom interventions: An attributional approach. Journal of Experimental Psychology: Applied, 26(1), 61–72. https://doi.org/10.1037/xap0000223 van Nguyen, K., Tran, K. V., Luu, S. T., Nguyen, A. G. T., & Nguyen, N. L. T. (2020). Enhancing lexical-based approach with external knowledge for Vietnamese multiplechoice machine reading comprehension. IEEE Access, 8, 201404–201417. https://doi.org/10.1109/ACCESS.2020.3035701

Wilson, J., & Roscoe, R. D. (2020). Automated Writing Evaluation and Feedback: Multiple Metrics of Efficacy. Journal of Educational Computing Research, 58(1), 87–125. https://doi.org/10.1177/0735633119830764

Yu, F. Y., & Wu, W. S. (2020). Effects of student-generated feedback corresponding to answers to online student-generated questions on learning: What, why, and how? Computers and Education, 145, 103723. https://doi.org/10.1016/j.compedu.2019.103723