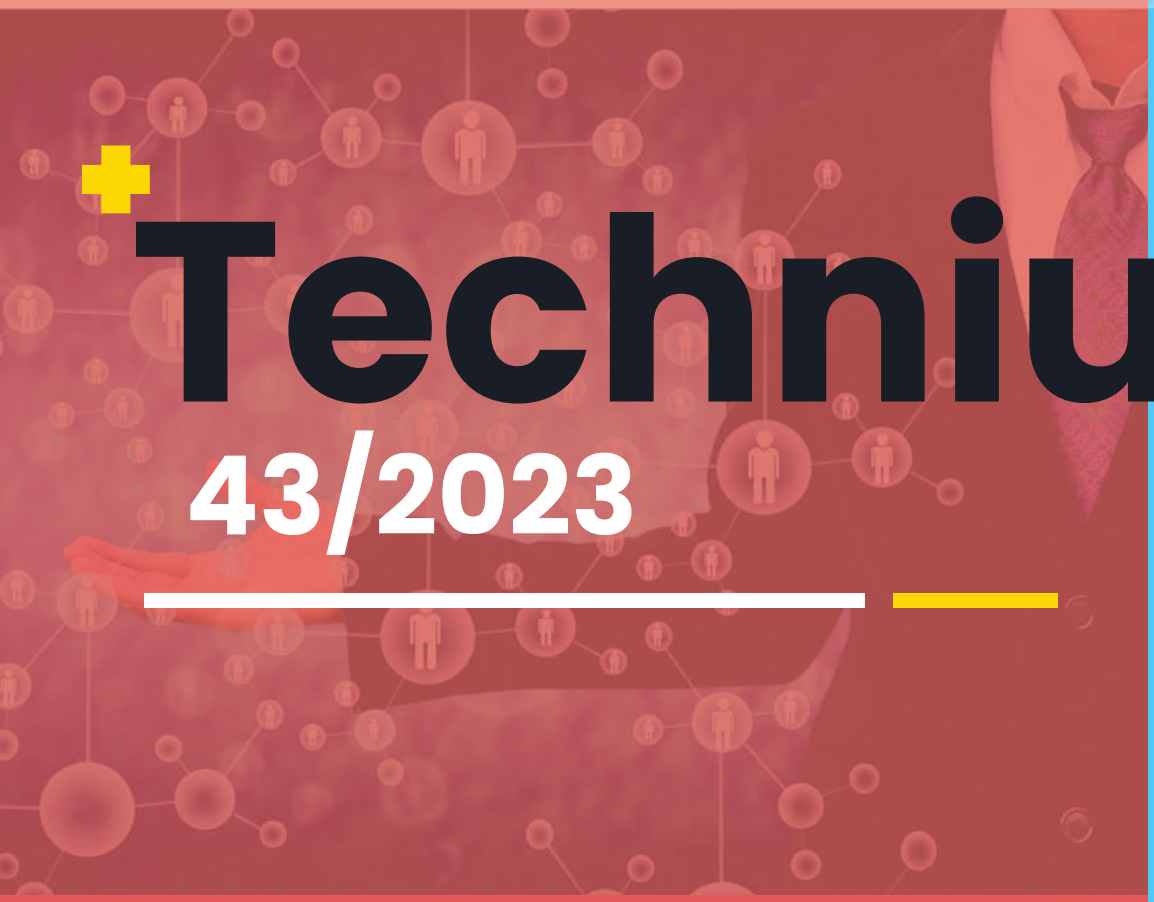
TECHNIUM
SOCIAL SCIENCES JOURNAL



Technium.

43/2023

2023

A new decade for social changes

Technium Social Sciences

Powered by PLUS COMMUNICATION
International Communication & PR

Technium Social Sciences Journal
Vol. 43, 205-212, May, 2023
ISSN: 2668-7798
www.techniumscience.com

Evaluation of Multiple Choice Items in the Telecommunication and Navigation Engineering Department with Validation

BB. Harianto[12], Arie Wardhono[1], Bambang Suprianto[1], Wiwid Suryono[2]

1. Universitas Negeri Surabaya, Surabaya, Indonesia.
2. Politeknik Penerbangan Surabaya, Surabaya, Indonesia

Bambang.19014@mhs.unesa.ac.id

Abstract. Improved multiple-choice questions (MCQs) item analysis is fundamental to be used for further tests. Besides increasing item analysis can also be used to removes items that are trapped in a test. This journal focuses on the quality of the questions on the test. It explores the relationship between the difficulty index (p-value) and the discrimination index (DI) and the efficiency of the distractor (DE). The research will be conducted by 49 6th semester students currently studying Diploma, especially in air navigation engineering in education at Aviation Polytechnic of Surabaya. Fifty multiple-choice questions will be given while a final exam for the transmission media and antenna courses. The value of validity using the Pearson correlation with a significance level of 5% and reliability using Cronbach's Alpha. We find that the difficulty index (p) is 59% with an SD value of 10%, while the DI value is 30 % with an SD value 10% and DE with a value of 25% (SD 22%) has item items with average difficulty and discriminatory power related to functional impairments that must be integrated into the test later to improve the quality evaluation.

Keywords. MCQs, discrimination index, difficulty index, distractor efficiency

1. Introduction

A professional teacher must meet competency standards. According to the Law No. 14 of the Republic of Indonesia on Teachers and Lecturers of 2005, competence is a set of knowledge, skills, and behaviours that teachers or lecturers must possess, internalize and master in the performance of their professional functions, Article 10 paragraph (1) states that "Teacher competencies referred to in Article 8 include pedagogical competencies, personal competencies, social competencies, and professional competencies obtained through professional education". Based on the description above, one of the competencies that must happen What is mastered is the ability to teach. Teaching ability is the ability to understand students' abilities, the design and implementation of learning, the evaluation of learning outcomes, and the ability to develop students to realize their various potentials.

Evaluation is the process of critically checking the program. It involves collecting and analyzing information about the activities, characteristics, and results of the project. Its purpose is to value a program, improve its effectiveness, and provide information for programming

Technium Social Sciences Journal
Vol. 43, 205-212, May, 2023
ISSN: 2668-7798
www.techniumscience.com

decisions. (Harris, 1987). The evaluation process is one of the essential tools in achieving learning objectives. One method that can be used to conduct a learning evaluation process is conducting an exam or exam. A test is a measurement tool in the form of questions, commands, and instructions aimed at the examinees to get these instructions.

Assessment tools or tests are a critical component in learning activities to determine the extent to which students have achieved learning. Therefore an educator or teacher must have the ability to plan, arrange and make assessment tools. However, after completing and giving tests to students, teachers rarely evaluate the assessment tools or test items. Most teachers are only focused on assessing their students without identifying the difficulty level, discrimination features, and functions of the distractors. In fact, according to (Johnson & Kress, 1971), identification of each item of learning outcomes test items is carried out in the hope that it will produce a variety of valuable information, which basically will be feedback to make improvement and refinement of items that have been issued in the learning outcomes test so that in the future, the results of learning compiled or designed by the tester (teachers, lecturers, and others) can carry out its function as a measure of learning outcomes that have high quality.

Identification of each item is made in the hope of finding a variety of information, which is feedback to make improvements and refinement of the items, so that in the future, the results of tests of learning outcomes are arranged or designed by the teacher can measure what he wants to measure which is done continuously and carried out by independent institutions on a regular, comprehensive, transparent and systematic basis, to assess the achievement of national education standards. The item identification activity requires an assessment tool or technique. Educational evaluation tools used to collect data can be either test or non-test. Two forms of tests used in this evaluation must be accountable, meaning that the tests can qualify as a good evaluation tool when viewed from the quality of the items.

Needs analysis of items in the teaching and learning process can be used tests that have been standardized and the teacher's test. Standardized tests have undergone a standardization process, namely the process of validity and reliability. The test is valid and reliable for a particular purpose and a particular group. Tests that the central government has standardized are used in national exams. At the same time, the teacher's test is a test prepared by the teacher himself to evaluate the success of the teaching and learning process. Usually, teacher-made tests are widely used in schools.

Evaluation should be able to encourage students to be more diligent in learning. It can motivate teachers to improve the quality of the learning process further to improve their competencies and student outcomes. In this connection, a teacher is. You can not only teach well but also evaluate well. As part of the study program, assessment activities should be further optimized.

Research (Quaigrain & Arhin, 2017) found that building a multiple-choice test item for the final semester exam takes time and careful selection of content to produce the required test results. Controlling quality is essential for developing tests. As a result, when teachers believe they are lacking, they should undertake project analyses or ask for assistance.

Meanwhile, in (Mehta & Mokhasi, 2014), item analysis was mostly used to make a decent question bank and assess class performance as part of formative assessment. This study concluded that items with average difficulties and high discrimination power with functional impairments should be included in subsequent tests.

Other opinions were expressed by (Abed et al., 2015) on tests that measure the numerical ability of Jordanian university students use the Reaction Theory (IRT) project. Generally speaking, the test of numerical ability for students of educational sciences has good

Technium Social Sciences Journal
Vol. 43, 205-212, May, 2023
ISSN: 2668-7798
www.techniumscience.com

psychometric characteristics. The test can be used as a numerical exploration of ability strengths and weaknesses to support strengths and assess weaknesses by using appropriate strategies and activities to teach digital skills (Slepkov et al., 2021). One of MCQ's longstanding criticisms is that it prioritizes testing low-level facts and classification knowledge over more complex applications, analysis, and synthesis knowledge. It can be assumed that the complexity of the test knowledge of the introductory course is lower than that of the advanced course.

(Sheng & Lou, 2018) This suggested that gamification had a positive impact on motivation, interactions, and performance in engineering lab activities, indicated by the high number of students participating in the GM site. Additionally, the number of responses submitted at the GM website was higher (at the start), and the number of days of differential participation was days higher for students in the GM group.

The Academy of Engineering and Aviation Safety Surabaya, known as Poltekbang Surabaya, is one of the official universities under the Ministry of Transportation of the Republic of Indonesia, which provides vocational and academic-based education in the field of aviation engineering and safety. Considering the broad scope of the discussion of the problem of this study when viewed from the title of the study, the researcher limits the problem to be investigated, including:

1. This research was conducted in the telecommunication and air navigation department study class XA and XB aviation polytechnic of Surabaya with 49 answer sheets.
2. The exam questions analyzed are questions of transmission media and antenna subjects consisting of 50 multiple-choice items.
3. Discus with Aviation Polytechnic Surabaya lecturers, especially those who are competent in the subject of transmission media and antennas, for use in summative tests (semester exams) in the odd semester of the 2019/2020 academic year.

Through item analysis, you can observe the characteristics of the items and improve the test quality (Kim et al., 2016). According to (Lange et al., 1967), item revisions will enable us to identify too difficult or too easy items that cannot distinguish There is meaningless interference between students who study content and students who have no content, or the question has pointless interference. Educators can delete these non-discriminatory items from the item collection, change items, modify instructions to correct misconceptions about the content or adjust the teaching method. Conducted this research to estimate and evaluate multiple-choice items on questions used in the air navigation engineering department consisting of evaluating items and testing quality exploring the correlation of distractor effectiveness, hardship, and discrimination index (DE).

2. Methodology

Forty-nine students take part doing this test. The test consists of 50 multiple-choice questions (MCQs) of transmission line and antenna subjects. The generation of multiple-choice questions (MCQs) from a text is a popular area of research. MCQs are widely accepted for large-scale assessment in a variety of fields and applications, (Balaha & Saafan, 2021) the application of multiple-choice questions (MCQs) as a summative assessment method has generated opinions. Polarity related to higher academic proficiency tests. For large classes, automatically graded MCQs provide instant feedback that is very appealing to students and teachers alike.

All respondents were students of air navigation engineering study program class X, composed of 2 classes, namely class A and B, at Aviation Polytechnic of Surabaya. This test

Technium Social Sciences Journal
Vol. 43, 205-212, May, 2023
ISSN: 2668-7798
www.techniumscience.com

uses a standard protocol determined by Surabaya Aviation Polytechnic. The answers from each student will be analyzed using Microsoft Excel. Each multiple-choice question is analyzed for the difficulty level, size of the discrimination as evaluated by the discriminant index and difficulty index (p-value), and psychoanalysis for all incorrect choices. Analysis data also includes the value of validity using the Pearson correlation with a significance level of 5% and reliability using Cronbach's Alpha for both of the items using SPSS 23. The findings demonstrate each test item's validity and dependability.

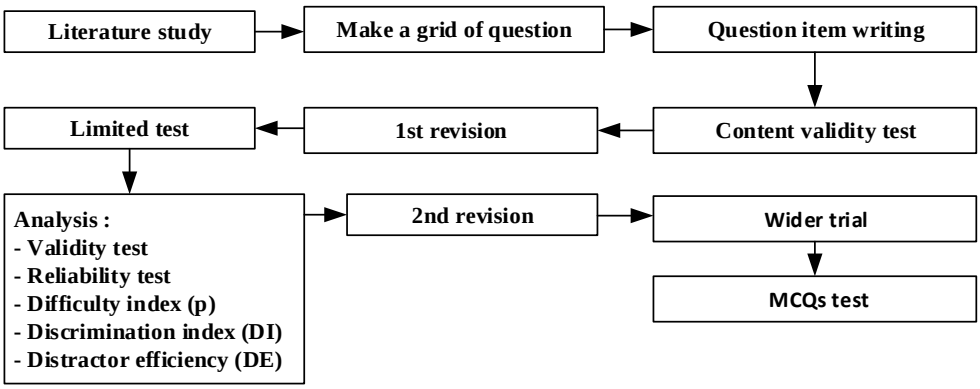


Figure 1. Test Development Research

In this research in figure 1 have a step we must do it first we need an assessment process to analyze the research we are doing. Literature study is the first step then we must make a grid of questions, and then make item question writing, then we must validate the question we have to do, If there are questions that are not appropriate, we will carry out the first revision process, then we do a limited test to measure the level of validity, reliability, difficulty, discrimination, distractor efficiency of the questions we make, and we do the second revision, the final stage of the wider trial, finally we can implement MCQs.

The time to do the test is 90 minutes, and each question has four answer options, One of them is the correct answer to the question, and three others become distracters. The correct answer is given a mark of 1, and no negative value is provided on the wrong answer. In general, questions always use four answer options to reduce the guessing effect performed by test takers. Most national and international tests, such as the TOEFL, have four answer options per question. So the maximum score from this test is 50, and the smallest one is 0.

2.1 Analytical Procedure

The difficulty index and discrimination index of each multiple-choice item were examined using the test participant achievement scores from the composite test. The formula 1 is used to determine the difficulty index as a proportion of the total number of correct answers:

$$p=R/T \quad (1)$$

Where p is the item difficulty index, R is the number of correct answers, and T is the number of questions answered, including incorrect and accurate answers. According to, the p-value (proportion) is between 0 and 1. (Gajjar et al., 2014). The percentage of students who correctly identified the element is calculated by converting the p-value to percent. The higher the value, the more likely the problem is to occur. So the higher the difficulty index value, the easier the items to understand. P values between 20% to 90% are reasonable and can be obtained

Technium Social Sciences Journal
Vol. 43, 205-212, May, 2023
ISSN: 2668-7798
www.techniumscience.com

because the range of p values between 40% to 60% is considered very well. Items of the A p-value (difficulty statistic) of less than 20% are considered very difficult, and if the p-value has a value of 90%, it is considered too easy to be accepted and should be modified first. The p-value is a measure of the test taker's behavior.

The point correlation consists of the correct answer value and all other total scores DI items. So total number of students above 27% gets the correct answer and 27% lower than the student who gets the correct answer. We can calculate DI using the patter 2:

$$DI=(UG - LG)/n \quad (2)$$

Where UG is the number of students in the above group who obtained the correct answer, and LG is the number of students at the bottom group who received the right answer, and n is the most significant number of students in the two groups (the highest group or lowest class). The ability to distinguish between kids with higher and lower exam scores improves with increasing DI.

Based on research (Tarrant et al., 2012) Regarding the guidelines for classic anchovy analysis test items categorized in the discriminatory index, the rules are: (1) If DI DI 0.40, then the item functions satisfactorily; (2) If $0.30 \leq DI \leq 0.39$, little or no revision is needed; (3) If $0.20 \leq IN \leq 0.29$, then the item is marginal and needs to be revised; (4) If $DI \leq 0.19$, the item must be removed entirely or revised.

DI cause of reflection is the level of an item and a comprehensive test to measure student reliability; Tests with a variety of content areas typically have lower coefficient values than tests with a more uniform distribution. In calculating DI, each student test is given a value, and the test value is computed using Microsoft Excel software. Furthermore, 27% of the students above and 27% below were separated for analysis. DI is calculated as the difference between the number of students who properly answered in the upper and lower groups, divided by the larger number of people in the two groups. It must be noticed that the element is better the higher the DI. The removal of these less discriminatory factors can seriously affect the validity of the test. Always consider the type of test being evaluated when interpreting the discriminant item index. A low discriminating index item should be scrutinized because confusing words are common. The reason a negative value was achieved for items with a negative index should be investigated.

3. Result and Discussion

There are 50 items on the test. The 49 pupils' grade ranges from 6 to 47. (out of 50). In other words, the test's standard deviation is 9.23, while the average result is 29.34. The average score is 30, so. The median score exceeds the mean score by a small margin. The distractor efficiency (DE) is 0.25, which indicates that there are more than 5% of suitable distractor values, the difficulty index (p), which is 0.59, indicates that the items are great.

A non-function distracter is one that has a positive discrimination power or a distractor efficiency below 5%. Additionally, there was a significant correlation between the overall score and inactive distractions. [13] It is advised that the DE range from 0 to 100% and be calculated for each element as a function of NFD quantity. The DE will be 0, 33.3, 66.6, and 100%, respectively, depending on whether a component has three, two, one, or no non-functional distractor (NFD). An overview of the experimental statistics is shown in Table 1.

Table 1. Test Parameter Summary

Parameter	Mean	Standart Deviation (SD)

Technium Social Sciences Journal
Vol. 43, 205-212, May, 2023
ISSN: 2668-7798
www.techniumscience.com

Difficulty Index (*p*)	0,59	0,1
Discriminator Index (DI)	0,3	0,1
Distractor Efficiency (DE)	0,25	0,22

The item difficulty index categories are displayed in Table 2. All items 50 (100%) are between 20 and 90 percent in p-value, which is an appropriate level of difficulty. Items 49 (98%) among them have an exceptional p-value (40-60 percent).

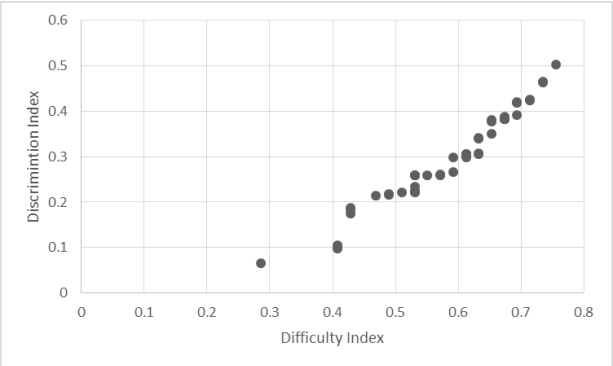
Table 2. Distribution of difficulty Index

Indicators of difficulty	Periodicity	Percent
<10	0	0
10.5-19.5	0	0
19.5-29.5	1	2
29.5-39.5	0	0
39.5-49.5	8	16
49.5-59.5	14	28
59.5-69.5	22	44
69.5-79.5	5	10
Final	50	100

The frequency distribution of the DI items is shown in Table 3. Item 36 has the highest DI (50%) while item 16 has the lowest DI (0.65 percent). Nineteen items (38%) and seven (14%) are good on the discrimination index (DI 0.3), respectively. Table 3 shows that 18 (36%) of the items need to be altered due to their insufficient discrimination power, per the research.

Table 3. Distribution of discrimination indices

Indicators of difficulty	Periodicity	Percent
<0.1	2	4
0.1-0.19	4	8
0.2-0.29	18	36
0.3-0.39	19	38
0.4-0.49	6	12
0.5-0.59	1	2
Final	50	100



Technium Social Sciences Journal
Vol. 43, 205-212, May, 2023
ISSN: 2668-7798
www.techniumscience.com

Figure 2. Relationship between the difficulty index and the discrimination index shown in a scatter plot.

The link between the difficulty index and the discrimination index is depicted graphically in Figure 2. The difficulty index ranges from a value of 0.28 to a value between 0.7 and 0.8, while the discrimination index ranges from a value of almost 0.1 to a value of 0.5. The graph above demonstrates a correlation between the difficulty index and discrimination index.

Table 4. Distribution of the frequency of a working distraction

	Periodicity	Percent
Number of items	50	
Number of distractors	150	
Distractor with frequency <5%	7	4.67
Distractor with discrimination ≥ 5%	143	95.33

3.1. Difficulty Level Analysis Results

In all 150 distractors rated, 7 distractors (4.67%) have a choice frequency of <5%. On the other hand, 143 distractors (95.33%) have a choice of frequencies ≥5. The essence of scattering analysis is the identification of non-functioning items and functioning options. The intruder function is an independent indicator for the function of the item. Distractors One or more test takers are called function diversion, and no one chooses them called a non-functioning bully. Designing options with good common sense is a difficult task, especially at the semester exams. The confounding function it lacks in writing items and the optimal number of options is interrelated and affects the quality of the item, the performance of the item, and the test results

4. Conclusion

It must be synchronized with test items in order to obtain a validated instrument. This necessitates the creation of appropriate test items as well as their analysis. Teachers can utilize tools like the difficulty and discriminating index to assess how well-formulated the multiple-choice questions are. The distractor efficiency is yet another instrument utilized for additional study (DE). The average p-value in this investigation was 59 percent, which falls within the excellent difficulty range (p = 40–60 percent). The study's average DI, which was 0.3, was deemed to be quite good.

Distracting factors are analyzed to determine their relative usefulness in each response item. If students consistently fail to choose specific multiple-choice options, those options cannot make sense and, therefore, are not used as much foil in multiple-choice items. Thus, the practical design of a distractor and the reduction of NFD are essential aspects for framing multiple choice quality problems. Distractions are an essential part of an item because they illustrate the correlation between overall exam scores and the distractions that students choose.

This study shows that not only difficulty and discrimination can make a good items, an NFD is needed to deep the better items, the data show that an NFD tend to fill in the high difficulty index and discrimination index.

Acknowledgement

Our gratitude goes to the lecturers at the State University of Surabaya for the knowledge that has been given and all Vocational Education doctoral friends who support each other in all

respects and friends from aviation polytechnics Surabaya. They also continue to provide support in completing this research

References

[1] Abed, E. R., Al-Absi, M. M., & Abu shindi, Y. A. (2015). Developing a Numerical Ability Test for Students of Education in Jordan: An Application of Item Response Theory. *International Education Studies*, *9*(1), 161. https://doi.org/10.5539/ies.v9n1p161

[2] Balaha, H. M., & Saafan, M. M. (2021). Automatic exam correction framework (AECF) for the MCQS, essays, and equations matching. *IEEE Access*, *9*(1), 32368–32389. https://doi.org/10.1109/ACCESS.2021.3060940

[3] Gajjar, S., Sharma, R., Kumar, P., & Rana, M. (2014). Item and test analysis to identify quality multiple choice questions (MCQS) from an assessment of medical students of Ahmedabad, Gujarat. *Indian Journal of Community Medicine*, *39*(1), 17–20. https://doi.org/10.4103/0970-0218.126347

[4] Harris, T. S. (1987). Discussion of Signaling and Monitoring in Public-Sector Accounting. *Journal of Accounting Research*, *25*(1987), 159. https://doi.org/10.2307/2491084

[5] Johnson, M. S., & Kress, R. (1971). Task Analysis for Criterion-Referenced Tests. *Reading Teacher*.

[6] Kim, E., Rothrock, L., & Freivalds, A. (2016). The effects of Gamification on engineering lab activities. *Proceedings - Frontiers in Education Conference, FIE, 2016-November*. https://doi.org/10.1109/FIE.2016.7757442

[7] Lange, A., Lehmann, I. J., & Mehrens, W. A. (1967). *USING ITEM ANALYSIS TO IMPROVE TESTS*. *4*(2), 1965–1968.

[8] Mehta, G., & Mokhasi, V. (2014). Item Analysis of Multiple Choice Questions- An Assessment of the Assessment Tool. *Historical Aspects of Leech Therapy*, *4*(7), 1–3.

[9] Quaigrain, K., & Arhin, A. K. (2017). Using reliability and item analysis to evaluate a teacher-developed test in educational measurement and evaluation. *Cogent Education*, *4*(1). https://doi.org/10.1080/2331186X.2017.1301013

[10] Sheng, Y., & Lou, Y. (2018). A New Method to Determine Validity of Student Teaching Evaluation Scores. *Open Journal of Social Sciences*, *06*(04), 337–345. https://doi.org/10.4236/jss.2018.64026

[11] Slepkov, A. D., Van Bussel, M. L., Fitze, K. M., & Burr, W. S. (2021). A Baseline for Multiple-Choice Testing in the University Classroom. *SAGE Open*, *11*(2). https://doi.org/10.1177/21582440211016838

[12] Tarrant, M., Calitri, R., & Weston, D. (2012). Social Identification Structures the Effects of Perspective Taking. *Psychological Science*, *23*(9), 973–978. https://doi.org/10.1177/0956797612441221