applied sciences

MDPI

Article

Exploring the Efficacy of Binary Surveys versus Likert Scales in Assessing Student Perspectives Using Bayesian Analysis

Andrés Suárez-García [1], María Álvarez-Hernández [1], Elena Arce [2],* and José Roberto Ribas [3]

[1] Defense University Center, Spanish Naval Academy, University of Vigo, 36920 Marín, Spain; andsuarez@uvigo.es (A.S.-G.); maria.alvarez@cud.uvigo.es (M.Á.-H.)
[2] Polytechnic School of Engineering of Ferrol, University of A Coruña, 15403 Ferrol, Spain
[3] Polytechnic School, Federal University of Rio de Janeiro, Rio de Janeiro 21941-853, Brazil; ribas@poli.ufrj.br
* Correspondence: elena.arce@udc.es

Abstract: Likert-scale surveys are the undeniable protagonists of online evaluations. They ask the respondent to express their degree of agreement with a series of statements related to the development of a subject. In contrast, in social networks, dichotomous surveys are mostly used. They force respondents to polarize their opinions by selecting "like" or "dislike". This study compares the efficacy of binary and Likert surveys in gathering student opinions on mechanical engineering program subjects. Using Bayesian analysis, it analyzes the similarity of responses obtained from both formats. For each question and scale, the ratio of "I like" among the total responses collected was calculated. The Bayesian factor method was used to compare the ratios obtained. The null hypothesis was equality between the ratios obtained by the different scales for the same question. This hypothesis was rejected in only 7 of the 49 questions evaluated—less than 15%. Finally, the students were surveyed on the preference for use of both scales. More than 80% stated a preference for the use of the dichotomous format. In view of the results obtained, we recommend more frequent use of the dichotomous scale to gather students' opinions.

Keywords: survey; student; Likert scale; binary scale; Bayes factor

Citation: Suárez-García, A.; Álvarez-Hernández, M.; Arce, E.; Ribas, J.R. Exploring the Efficacy of Binary Surveys versus Likert Scales in Assessing Student Perspectives Using Bayesian Analysis. *Appl. Sci.* **2024**, *14*, 4189. https://doi.org/10.3390/app14104189

Academic Editor: Luis Javier Garcia Villalba

Received: 1 April 2024
Revised: 10 May 2024
Accepted: 12 May 2024
Published: 15 May 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Rating systems are a widely used tool for gathering crucial information for decision making. They offer a way to summarize opinions in an organized way. They are used in marketing, politics, and teaching, among many other fields. For example, investors make decisions based on ratings of financial products, consumers compare products by examining ratings of previous buyers, students consult university rankings to decide where to apply, and universities use rating systems to monitor the performance of professors. A ranking system is defined by a metric and an aggregation rule that combines individual rankings into a single overall score. Several rating systems exist. Two of the most popular are the binary and Likert systems. For example, seller reputation systems on buy–sell platforms may request users to answer a questionnaire on a {1, 2, 3, 4, 5} scale, where 1 is considered the worst and 5 the best. Another everyday example is found on streaming platforms, which collect feedback on content via a "like" or "dislike" score that is usually coded as "1" or "0". Using the information about the set of opinions collected, regardless of the scale used, a decision maker can draw a conclusion about a product or service. Usually, this conclusion has a binary outcome such as buying a product or rewarding a teacher for his or her good work.

Likert-scale surveys dominate opinion gathering in marketing and academia [1]. However, there is a good reason to consider using other formats: survey size is one of the most influential factors in the successful collection of voluntary opinions [2]. If a questionnaire takes a long time to complete, many respondents may only fill it out partially. This negatively impacts the quality of the data collected [3,4]. Sometimes surveys are necessarily long. Such is the case for brand image measurement or student surveys on the quality of teaching. At the University of Vigo, each questionnaire consists of 20 questions to evaluate each teacher. A subject may be taught by multiple teachers, making it necessary to carry out the same survey several times. This survey has questions related to the teacher and to the general development of the subject. The student is given one hour to evaluate all the professors teaching the same subject. Thus, measures are needed to make the questionnaires easier and faster.

There is no good explanation for the predominance of five-point and seven-point Likert scales. A survey is a scientific measure of opinion. Its format has to be thought out, justified, and tested to be valid. There are studies showing that binary responses perform as well as multiple-choice responses [5–8]. In addition, it has been shown that seven-point Likert items suffer from a cultural bias in response style [9–11] that does not depend on the item being assessed. For example, there are cultures where intermediate positions are preferred to avoid direct confrontations. This means that extreme responses will be infrequent for reasons unconnected to the survey. As mentioned above, shorter questionnaires are likely to increase data quality due to reduced respondent fatigue. In addition, survey interviewer staffing costs can be reduced by inviting respondents to participate in online surveys. The interviewer is paid for work time. If the interviewee were to complete the survey in less time, it would mean a lower business cost. Therefore, a reduction in the length of the questionnaire could result in significant savings in data collection costs. If binary surveys produce measurements that are just as valid as Likert surveys in a faster and simpler way, it makes no sense not to promote their use.

The motivation of the present study is to analyze the feasibility of using binary questions in surveys involving student opinions of teachers. Currently, the Likert scale is used at the University of Vigo, whereby respondents are asked to evaluate their agreement with a statement on a scale from 1 (do not agree at all) to 5 (totally agree). This provides good richness, since a distribution of scores is obtained. However, when it comes to establishing reports and objectives, results are totally diluted, with the distribution being simplified since the values are averaged according to their relative frequency or expected value. In light of this, why not directly collect the information in a binary "like/dislike" format? This would make it easier for students to fill out the questionnaire. In addition, students are typically familiar with the binary format since it is used in a large number of social networks where a binary opinion is asked ("thumb up/thumb down"). Will a binary survey provide the same information as the usual Likert-scale survey? There is a possibility that the results derived from binary response formats may differ from those obtained with multi-category ordinal formats, be less reliable, or be perceived as more difficult to obtain because respondents are less familiar with multi-category scales.

Hypothesis testing is the statistical method commonly adopted in psychological research. Over the last century, several proposals have emerged to perform hypothesis testing in a statistically valid way. One of the best-known approaches is the null hypothesis statistical significance test (NHST) of [12], an approach situated within the frequentist school of thought [13]. In it, the evidence against a null hypothesis (H_0: no effect or difference) is quantified using a probability. When a predetermined threshold α is exceeded, it is rejected. Usually, a 5% threshold is used. This is expressed as a level of 0.05 and the null hypothesis is rejected when the p-value is less than that value. The p-value quantifies the probability of obtaining data equal to or more extreme than the results observed, under the assumption that the null hypothesis is true. With the rejection of H_0, the test is said to be significant. For a given sample size, the lower the p-value, the greater the evidence against H_0. Possible causes for a non-significant contrast include (1) sampling variability or poor luck in sampling that prevents rejecting H_0, even though the intervention contemplated in the experiment produces a change; and (2) the intervention actually produces no effect. As the sample size increases, the problem of sampling variability can be solved. However, even if significant p-values indicate evidence against H_0, non-significant

p-values do not allow one to conclude that the data support the null hypothesis. This is elegantly summarized as "absence of evidence is not evidence of absence" [14].

Bayesian thinking is rapidly gaining popularity among psychologists and neuroscientists [15] for reasons such as flexibility, higher accuracy of data with noise and small samples, lower tendency for type I errors or false positives, the possibility of introducing prior knowledge into the analysis, and the intuitive and simple interpretation of results [16–19]. In recent years, as discussed, frequentist thinking has become associated with the p-value approach and null hypothesis significance testing (NHST). The misinterpretation and misuse of p-values, so-called "p-hacking" [20], has contributed to the reproducibility crisis in psychological science [21].

Regarding Bayesian hypothesis testing as a replacement or alternative to NHST and p-values, there are a variety of a posteriori indices that have been proposed in the statistical literature [19]. Conceptually, these indices are all based on the posterior distribution in some form, and are employed to test a null hypothesis such as H_0 against an alternative H_1. The mathematical theory behind each of the proposed posterior indices differs substantially, and examples include the Bayes factor [22,23], the region of practical equivalence (ROPE) [24], the direction probability (DP), and the Full Bayesian Significance Test (FBST) [25]. Some works have compared the results of these indices in models such as linear regression or two-sample parametric tests [26]. For this work, we decided to use the Bayes factor mainly because it is able to support or reject the null hypothesis, in addition to being one of the most traditional metrics.

2. Methodology

The study was divided into three stages: conducting surveys, data pre-processing, and comparison of results by the Bayes factor (Figure 1). First, surveys were carried out in Likert and binary formats. The Likert survey was performed by an interviewer from the University of Vigo, while the binary-format survey was conducted online. A lapse of one week was left between the completion of both, in an attempt to minimize the use of short-term memory. It should be noted that the binary version focused exclusively on questions related to the subject. Both surveys included the same questions. Second, the results were transformed from the Likert format to a dichotomous scale. It was essential that the samples being compared were of a similar nature. On one hand, there is the qualitative Likert scale with scores ranging from 1 to 5, each denoting a level of qualitative agreement. On the other hand, there is the binary scale with two extremes. It would be impossible to compare both of them without a common framework. Finally, once the results of both surveys were on the same scale, they were compared to find out if the scale influenced the respondent and the results.

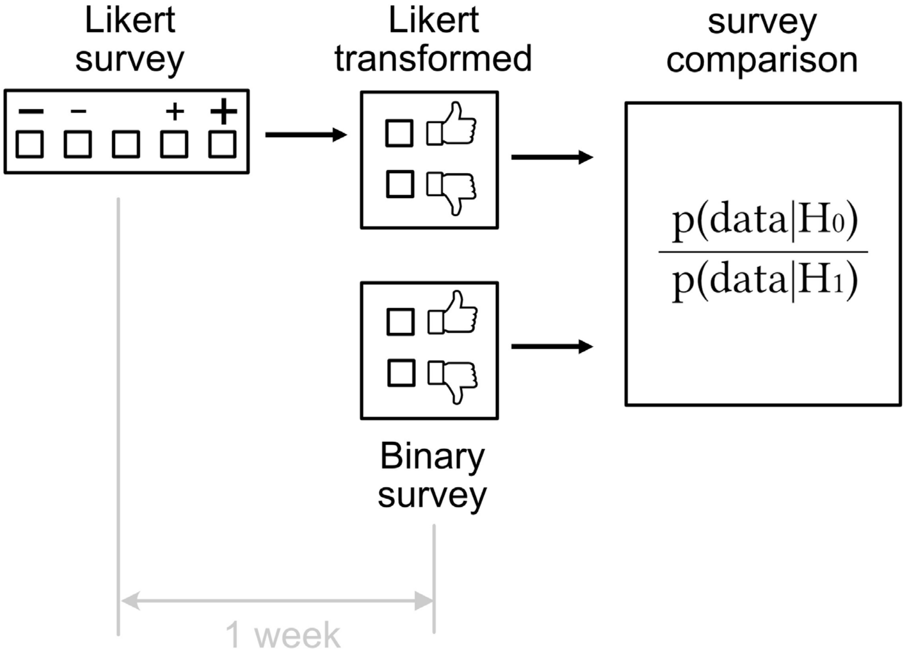


Figure 1. Steps taken in the present work.

2.1. Survey

Table 1 shows the questions used. They were focused on aspects related to the subject or the conditions for the development of the subject. In the questionnaire, there were 16 more questions related to each teacher of the subject, although these were left out. In this way, the online questionnaire kept a short format. In the case of keeping the questions about each teacher, it would be necessary to fill in 16 questions for each teacher. The center where the study was conducted has two to three teachers per subject. This would mean that the length of the questionnaire would need to be extended beyond 50 questions. Since the dichotomous survey was administered online, without a surveyor or class time devoted to filling it out, an attempt was made to make it as non-repulsive as possible. A 7-question questionnaire should take a student less than 2 min to complete. If there had been a much larger number of questions, it might have deterred completion. As mentioned, an interval of 1 week was left between the two surveys.

Table 1. Questionnaire for each subject.

Label	Question
Q1.1	This subject is important for my learning.
Q1.2	The credits assigned to the subject are commensurate with the amount of work required to pass.
Q1.3	The teaching guide (or program) of the subject is available and easily accessible.
Q1.4	The teaching guide (or program) of the subject includes the objectives, contents, methodology, bibliography and evaluation system in an understandable and detailed way
Q1.5	The coordination between teachers of the subject is adequate.
Q2.1	The conditions (space, equipment, material, etc.) in which the teaching takes place are satisfactory as far as theoretical classes are concerned.
Q2.2	The conditions (space, equipment, material, etc.) in which teaching takes place are satisfactory in terms of practical classes (laboratory, workshops, field classes…).

In addition to the 7 questions about the subjects and their development, the students were also asked about their preferences regarding the scales of the questionnaires (Table

2). For this purpose, there were 3 additional questions where they were asked to compare the ease, speed, and pleasantness of both formats. The results thus collected served as further evidence to be taken into account. If the respondent feels strong favor for the binary format and it collects the same information as the Likert format, this would be another compelling reason to encourage the use of the binary format. These questions were only asked in the binary format.

Table 2. Questions on the scales of the questionnaires.

Label	Question
Q3.1	I found the scale ("I like"/"I don't like") easier than the one usually used (1 to 5).
Q3.2	I found the scale ("I like"/"I don't like") easier than the one usually used (1 to 5).
Q3.3	I found the scale ("I like"/"I don't like") easier than the one usually used (1 to 5).

2.2. Data Pre-Processing

In the case of the Likert surveys carried out at the University of Vigo, a different one was formulated for each teacher. In each survey, all the questions were asked, with questions focused on the development of the subject and others on the teacher. Therefore, the questions focused on the subject were asked repeatedly, as many times as there were teachers. For example, if a subject was taught by 3 teachers and was taken by 50 students, 50 questionnaires would be obtained for each teacher with a total of 150 questionnaires for the subject. This fact must be taken into account when comparing the binary responses. This was carried out only once and focused on the subject.

It is curious to note that the respondents were inconsistent in repeating the information for the subject. This block of questions evaluated aspects related to the subject (importance of the training, workload, teaching guide...). They were independent of the individual work of the teachers teaching it. We expected the answers of the same students to be the same in the various repetitions of the same blocks of questions. This would imply that all frequencies recorded for each of the possible scores would be multiples of the number of teachers. For example, if the survey had been carried out for a subject with 3 teachers, the frequencies recorded for each score would have to be multiples of 3. Therefore, it can be seen that the teacher's perception influences questions that are not his or her own. To solve the problem of triplicate information in the Likert-scale surveys, we decided to divide the frequencies obtained in the survey results returned by the University of Vigo by the number of teachers. In addition, when a decimal place occurred, the score was rounded up or down to the nearest whole number. The information thus produced represented the results of each subject according to the Likert scale.

Once the size adjustment was made to the Likert survey, a clear difference in population size was observed between this and the dichotomous-scale surveys (Figure 2). The Likert population was always larger, reaching twice the size for most of the subjects analyzed. There were only two subjects, IT1 (Thermal Engineering I) and MFL (Fluid Mechanics), with a similar sample size. The explanation lies in the mode of gathering responses. While the Likert surveys were conducted face to face and collected on-site during a class, the binary surveys were performed remotely on a delayed basis (by providing the link to the students through the MOOVI teaching platform). The second case required a greater effort by the students, since they had to use part of their personal time to fill in the questionnaire. For this reason, a smaller sample size was obtained via the dichotomous-scale surveys.

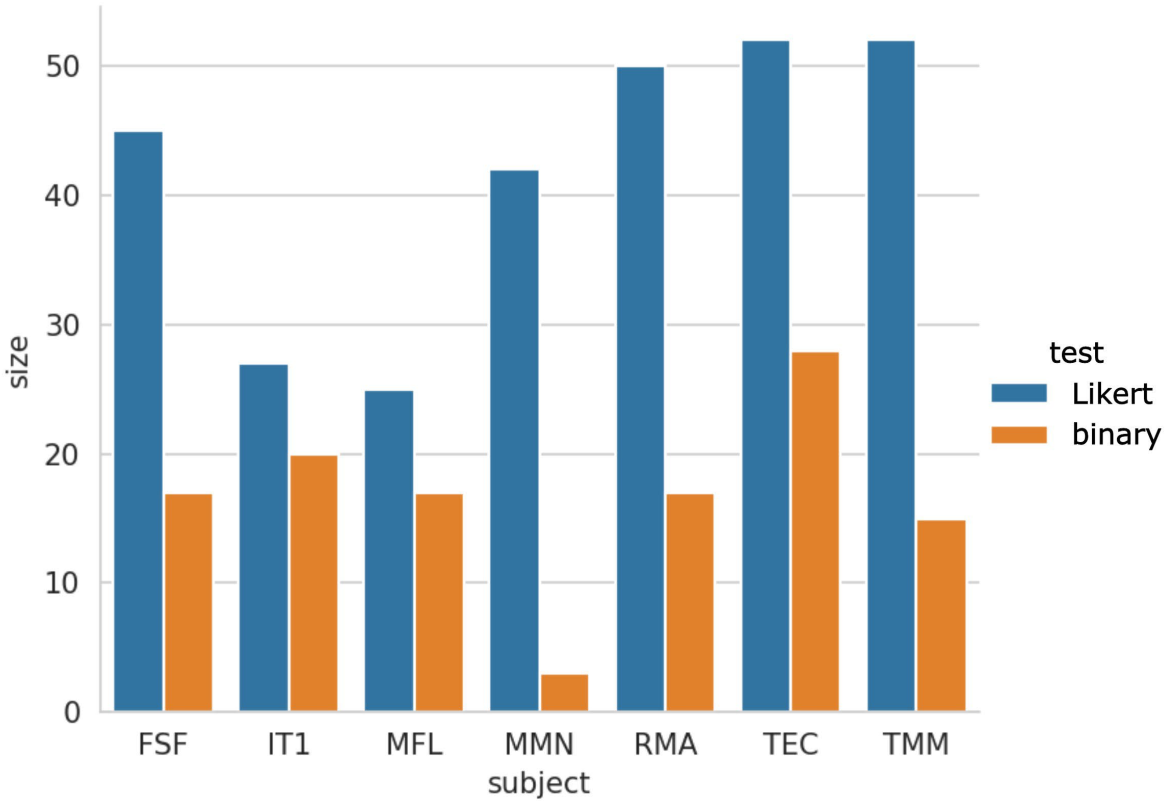


Figure 2. Number of survey responses by subject and format.

2.3. Likert–Dichotomous Conversion

It was necessary to put the results of both surveys on the same scale in order to compare the results. The Likert scale had five different values, while the binary scale had only two. It is easier to collapse this scale than to create five values from the dichotomous one, although the latter is not impossible. For the transformation, it was necessary to select a Likert score from which a positive response was considered. For example, by choosing the score "2" as the threshold, "2, 3, 4 and 5" would add up as positive responses. What threshold should be chosen to minimize the differences between the responses collected from the two surveys? The Likert scale involves a degree of subjective judgment influenced by cultural background. Emotions play a role in whether individuals respond positively or negatively to a question [27,28]. Different cultures perceive the midpoint differently; it represents a positive emotion for some, while it is more neutral for others. It was not known a priori the profile of the students surveyed. Choosing the wrong cutoff could introduce systematic bias into our analysis. This was possible for each threshold analyzed.

Since our aim was to look for similarities, the transformation was performed for all the responses collected, that is, from 1 to 5. After the transformation, a ratio of the affirmative responses among the total responses was obtained for each subject and question. This transformed quantity was compared with that obtained directly from the dichotomous survey. The histograms of the differences between the two magnitudes revealed that the threshold "3" produced the result with the least bias (Figure 3). Therefore, threshold "3" was chosen to transform the Likert surveys into binary or dichotomous surveys.

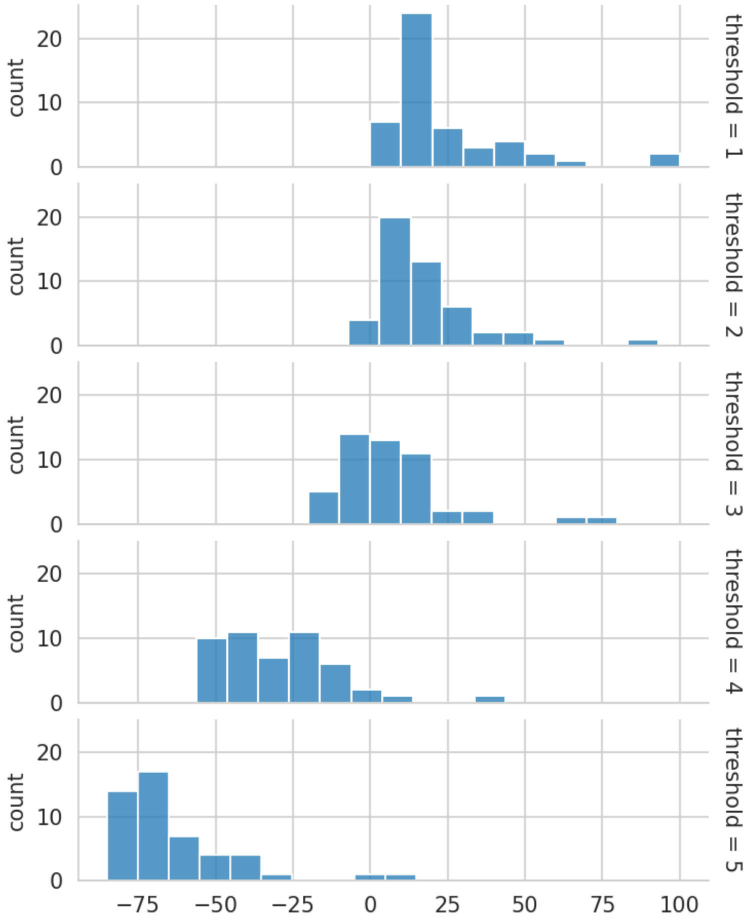


Figure 3. Threshold selection for transformation from Likert to dichotomous scale.

2.4. Bayes Factors

People update their knowledge as they have new experiences or obtain new data. In childhood, it takes a few attempts to understand that, unlike cartoons, cats cannot talk. A child does not undertake such an investigation without context: the child carries a degree of prior information based on previous experiences. This prior information helps to interpret the new data so that, weighing the new and old, the child can develop his/her updated information (posterior information). The child continues to refine this information while gathering new evidence. Allowing the posterior to balance prior knowledge and data is critical to the knowledge-building process. As new data are collected, the prior information loses weight. So even if two people had different prior knowledge, they would come to the same conclusion after collecting sufficient data.

In the context of hypothesis testing, at the beginning, the observer has a set of two rival hypotheses. In this case, the two hypotheses are the null hypothesis H_0 and the alternative hypothesis H_1. The null hypothesis states that there is no difference between two sets of data or no effect between before and after treatment. These two datasets are usually associated with the subjects of a treatment and a control group. In the case of this paper, the datasets are the binary survey success ratio θ_{bin} and the Likert-transformed survey θ_{likert}. Therefore, the null hypothesis is stated as $H_0\colon \theta_{bin} = \theta_{likert}$, and the alternative as $H_1\colon \theta_{bin} \neq \theta_{likert}$. These can be rewritten as $H_0\colon \theta_{bin} - \theta_{likert} = 0$ and $H_1\colon \theta_{bin} - \theta_{likert} \neq 0$. Both options can be assumed to be equally plausible, with $p(H_0) = p(H_1) = 0.5$. By using information from the data, the probabilities are updated: the hypothesis that best describes the data increases in credibility and the one that worst reflects the data suffers a decrease [17]. This is a process similar to reinforced learning. This updating is performed using Bayes' rule (1). Below, we describe this rule for one of the hypotheses. In it, $p(H_0)$

is the a priori belief in H_0, $p(H_0|data)$, the a posteriori belief given the data, and finally, $p(data|H_0)/p(data)$ is the predictive update factor. The a posteriori belief reflects the probability that the observed data conform to that hypothesis.

$$p(H_0|data) = p(H_0)\,\frac{p(data|H_0)}{p(data)} \tag{1}$$

To compare two hypotheses, Bayes' rule can be written as an odds ratio [22,29]. This equation shows that the change from the prior hypothesis odds $p(H_1)/p(H_0)$ to the posterior hypothesis odds $p(H_1|data)/p(H_0|data)$ occurs due to the ratio of the predictive update factors $p(data|H_1)/p(data|H_0)$, commonly known as the Bayes factor. The Bayes factor is abbreviated as BF_{10}, where the subscripts indicate the hypotheses placed in the numerator and denominator positions, respectively.

$$BF_{10} = \frac{p(data|H_1)}{p(data|H_0)} = \frac{p(H_0)\,p(H_1|data)}{p(H_1)\,p(H_0|data)} \tag{2}$$

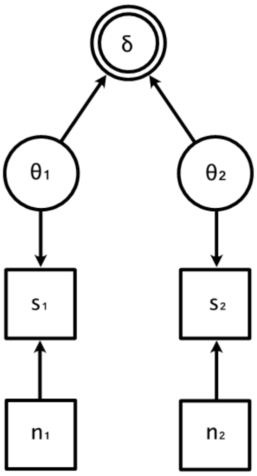
For example, suppose that the rival hypotheses are equally plausible beforehand; that is, $p(H_0) = p(H_1) = 0.5$. Then, the odds ratio of the prior hypotheses is equal to one. If the observed data are 10 times more likely under H_0 than under H_1, the assumption would be that the update factor is 10 times larger for H_0. This would cause the posterior probabilities of H_0 also to be 10 times greater. Since the hypotheses are mutually exclusive, with $p(H_0) + p(H_1) = 1$, this means that the data have increased the probability of H_0 from 0.5 (prior probability of H_0) to $10/11 \approx 0.91$ (posterior probability of H_0). The Bayes factor quantifies the degree to which the data justify a change in beliefs and thus represents the strength of the evidence that the data provide. Note that this measure of strength is symmetric: the evidence can support H_0 as well as H_1. A priori, none of the rival hypotheses enjoys any special status.

For a scientist who wants to know whether or not a treatment had an effect, the share of posterior probabilities might seem the most obvious metric, as this reflects the plausibility of one hypothesis over another after considering the data. However, posterior probabilities depend on both the evidence provided by the data (i.e., the Bayes factor) and the prior probabilities. Prior probabilities capture beliefs before the experiment and introduce an often undesirable element of subjectivity: conclusions drawn from posterior beliefs could be biased. Scientists can disagree strongly about these a priori probabilities, even if they agree about the evidence, i.e., about the extent to which the data should change their beliefs. Since beliefs are considered less valuable for scientific information than evidence, the data-fueled Bayes factor is the least controversial, and therefore the preferred metric for making decisions.

The Bayes factor has three qualitatively different logical states: (1) $BF_{10} > x$, according to which there is convincing evidence of the effect of interest; (2) $1/x < BF_{10} < x$, indicating the data do not carry enough weight to be able to make a diagnosis; and (3) $BF_{10} < x$, where the data are sufficient proof of the absence of an effect. x represents the objective level of evidence defined by the researcher. Harold Jeffreys proposed a logarithmic scale for interpreting the strength of evidence [30]. In it, the logarithmic values are equidistant such that $10^{0.5} \approx 3$, $10^1 = 10$, $10^{1.5} \approx 30$, etc. He then compared these values with the critical values of the t-tests and χ^2 values, noting the equivalence between the p-value 0.05 and $BF_{10} = 3$, in addition to the p-value 0.01 and $BF_{10} = 10$. These reference values are still used: $BF > 3$ is considered moderate evidence for the numerator hypothesis, and $BF > 10$ is considered strong evidence. Because $BF_{10} = 1/BF_{01}$, this also defines the bounds for the hypothesis in the denominator such that $BF_{10} < 1/3$ is moderate evidence in favor of H_0 and $BF_{10} < 1/10$ is strong evidence. Values of BF between 1/3 and 3 indicate that there is insufficient evidence to draw a conclusion for or against any hypothesis. For new findings, Jeffreys suggested that $x = 0$ is more appropriate than $x = 3$. These values are frequently used to categorize Bayes factors with slight differences in their interpretation [31]. However, each scientist in each field will have to decide whether to prefer test sensitivity for small samples or effects by using smaller x values, such as 3, or to avoid false conclusions by using higher x values, such as 10. In either case, readers can judge the strength of evidence directly from the numerical value of the BF, since a BF twice as high provides evidence twice as strong. Crucially, the three-state system of the Bayes factor allows differentiating between evidence of absence and absence of evidence. This represents a fundamental conceptual advance in the way data are interpreted: instead of one knowledge-generating outcome ($p < \alpha$) of the frequentist approach, there are now two ($BF > x$ and $BF < x$).

Now that the Bayes factors have been explained, it is necessary to define the different elements of the present study (Figure 4). To begin with, the hypotheses to be compared (H_0 and H_1) are defined. It should be recalled that the null hypothesis defines the absence of effect or difference between samples. In other words, $H_0\colon \theta_1 = \theta_2$ and $H_1\colon \theta_1 \neq \theta_2$, where θ_1 represents the affirmative responses to a Likert-scale question converted to binary and θ_2 is the ratio of affirmative responses in binary-scale surveys. This could be rewritten as $H_0\colon \delta = 0$ and $H_1\colon \delta \neq 0$ where $\delta = \theta_2 - \theta_1$. The a priori belief in both ratios is modeled by a vague prior distribution, such that $\theta_1 \sim Beta(1{,}1)$ and $\theta_2 \sim Beta(1{,}1)$. It is called vague in the sense that it reflects a belief that is very weak and easily molded by exposure to new information. In other words, the updating factor will have more strength than the prior distribution and will leave a greater role for the calculation of the a posteriori belief or distribution. Speaking of the updating or likelihood factor, the transformed Likert survey is modeled as $s_1 \sim Binomial(\theta_1, n_1)$ and the binary format survey as $s_2 \sim Binomial(\theta_2, n_2)$, where n represents the total number of responses and s the number of positive responses.

Typically, the use of Bayesian inference and the calculation of Bayes factors are associated with numerical computation methods. This is caused by the appearance of integrals when probability distributions are used to represent beliefs and likelihoods. However, comparisons similar to the present study have been performed by developing a simple equation for calculating the Bayes factor for each question asked [25,32], where n_1 denotes the total number of responses collected in the transformed Likert survey, n_2 denotes the total number of responses collected in the binary survey, s_1 denotes the number of affirmative responses in the transformed Likert survey, and s_2 denotes the total number of affirmative responses in the binary survey. This will be the expression used in our calculation due to its simplicity.



$$\begin{gathered} s_1 \sim Binomial(\theta_1, n_1) \\ \theta_1 \sim Beta(1{,}1) \\ s_2 \sim Binomial(\theta_2, n_2) \\ \theta_2 \sim Beta(1{,}1) \\ \delta \sim \theta_1 - \theta_2 \end{gathered}$$

Figure 4. Bayesian graphical model of the survey data.

$$BF_{01} = \frac{\binom{n_1}{s_1}\binom{n_2}{s_2}}{\binom{n_1+n_2}{s_1+s_2}} \frac{(n_1+1)\,(n_2+1)}{n_1+n_2+1} \tag{3}$$

In a general case, there is an easy way to estimate the Bayes factor. We will consider the hypotheses in a general manner as $H_0 : \phi = \phi_0$ and $H_1 : \phi \neq \phi_0$. The Bayes factor can be obtained by only considering H_1 (4) and dividing the height of the posterior for δ by the height of the prior for δ at the point of interest. This surprising method was first published by [33] and is called the Savage–Dickey density ratio method. The magnitude in this method has to be estimated using numerical methods. The revision of (Wagenmakers et al., 2010) is recommended for a simple explanation with written examples of WinBUGS. Nevertheless, in the present work, thanks to the model used, Equation (3) was employed without the need for numerical methods.

$$BF_{01} = \frac{p\left(\phi = \phi_0 | D, H_1\right)}{p\left(\phi = \phi_0 | H_1\right)} \tag{4}$$

After ascertaining the Bayes factor for each question, the null hypothesis posterior probability was also calculated. It helps to measure the confidence in what the evidence says about the null hypothesis. In other words, how strongly we believe in $H_0 : \delta = 0$. In order to calculate this, we use Equation (5), where $\Pr(H_0)$ represents the prior belief in H_0 being true. In the equation, $\Pr$ represents probability and p represents the probability distribution function. Since there are only two options and no prior leaning towards either the null or alternative hypothesis, the value of $\Pr(H_0) = 0.5$ is chosen. Because there are only two hypotheses, $\Pr(H_1) + \Pr(H_0) = 1$, and therefore, $\Pr(H_1) = 1 - \Pr(H_0)$. This equation can also be solved using probability distribution functions as the prior. In fact, Equation (3) is derived assuming a Beta(1,1) prior. However, this approach would necessitate the use of computational Bayesian inference and a posterior probability density function would be obtained. In the present work, the simplicity and information of the closed equations were chosen over computational methods.

$$\Pr(H_0|data) = \frac{\Pr(H_0) \cdot p(data|H_0)}{\Pr(H_0) \cdot\ p(data|H_0) + \Pr(H_1) \cdot p(data|H_1)} = \frac{\Pr(H_0) \cdot BF_{01}}{\Pr\,(H_0) \cdot BF_{01} + (1 - \Pr\,(H_0))} \tag{5}$$

3. Results

This section reports the results obtained. It focuses on three sections: the data preprocessing, the statistical comparison, and the opinion survey. The pre-processing focuses on the scale transformation necessary to compare both surveys, which are in binary format. The statistical comparison is performed using Bayesian factors. Finally, the opinion survey will show the students' preferences for the use of the scales.

3.1. Likert–Dichotomous Conversion

The Likert surveys converted to binary (left) and dichotomous surveys (right) for each question and for each subject are shown below. The color is the probability of agreement with the statement elicited by the question, this being the ratio between options marked YES and TOTAL. This is a ratio similar to the likes-to-view or view-to-likes ratio used to catalogue the popularity of a YouTube video, although in the case of YouTube, the rating is voluntary. Figure 5 shows the breakdown of responses categorized as "likes" or "hits" for Likert surveys converted to dichotomous format (s_1 and $n_1 - s_1$) and like/dislike surveys (s_2 and $n_2 - s_2$). In both cases, the variable s represents the "likes" and $n - s$ the "dislikes". As stated previously, a difference in the number of samples collected is observed. However, the color pattern assigned independently in each matrix clearly indicates similarity. This coloring is performed by means of a scale between the maximum and minimum values recorded in each matrix. The matching color intensity between each question and subject implies a clear agreement between the two surveys. For example, question Q1.1 has a similar color for variables $n_1 - s_1$ and $n_2 - s_2$, in addition to variables s_1 and s_2. In areas where there is color disparity, as will be seen below, this is indicative of a discrepancy in results between the two surveys.

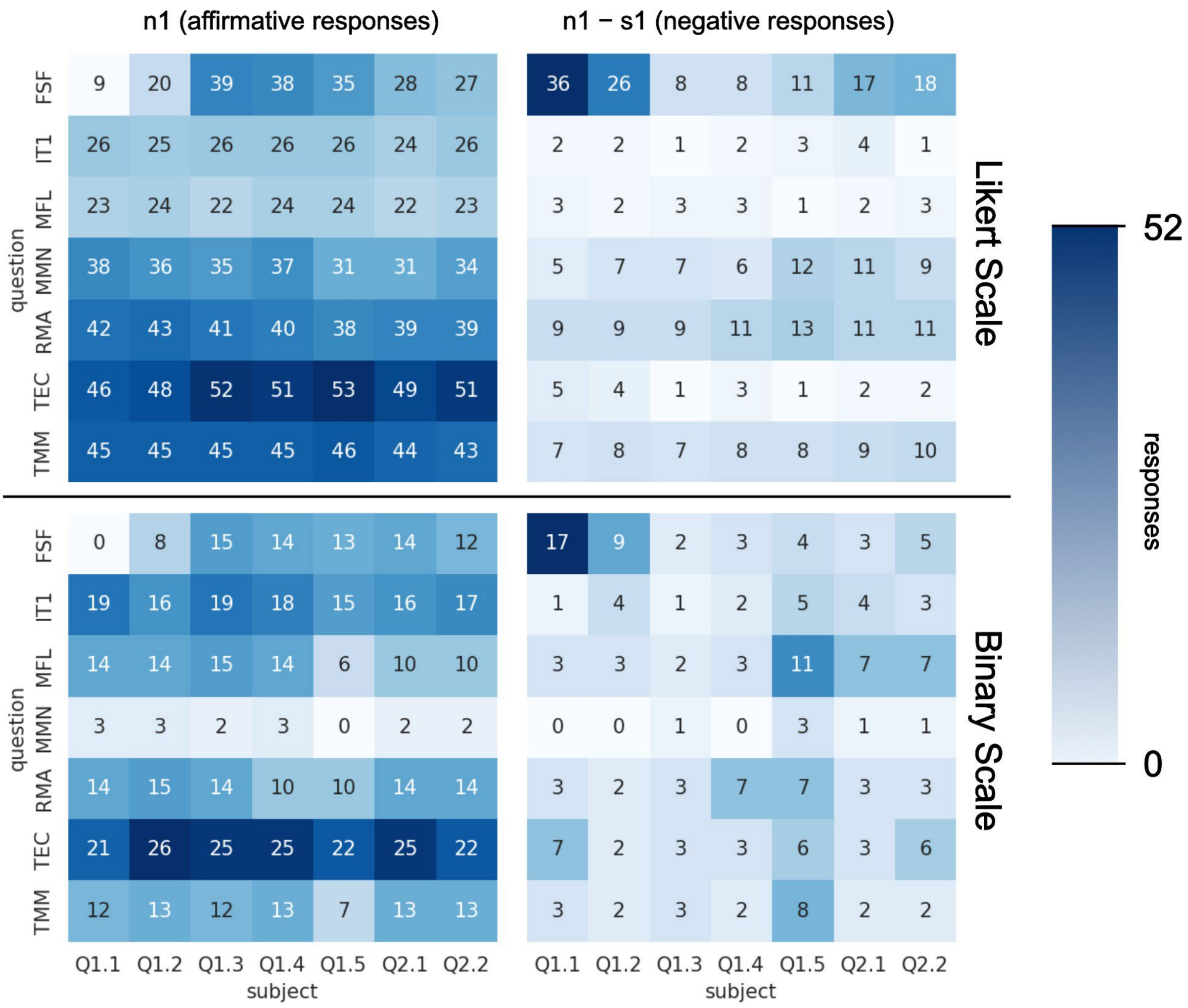
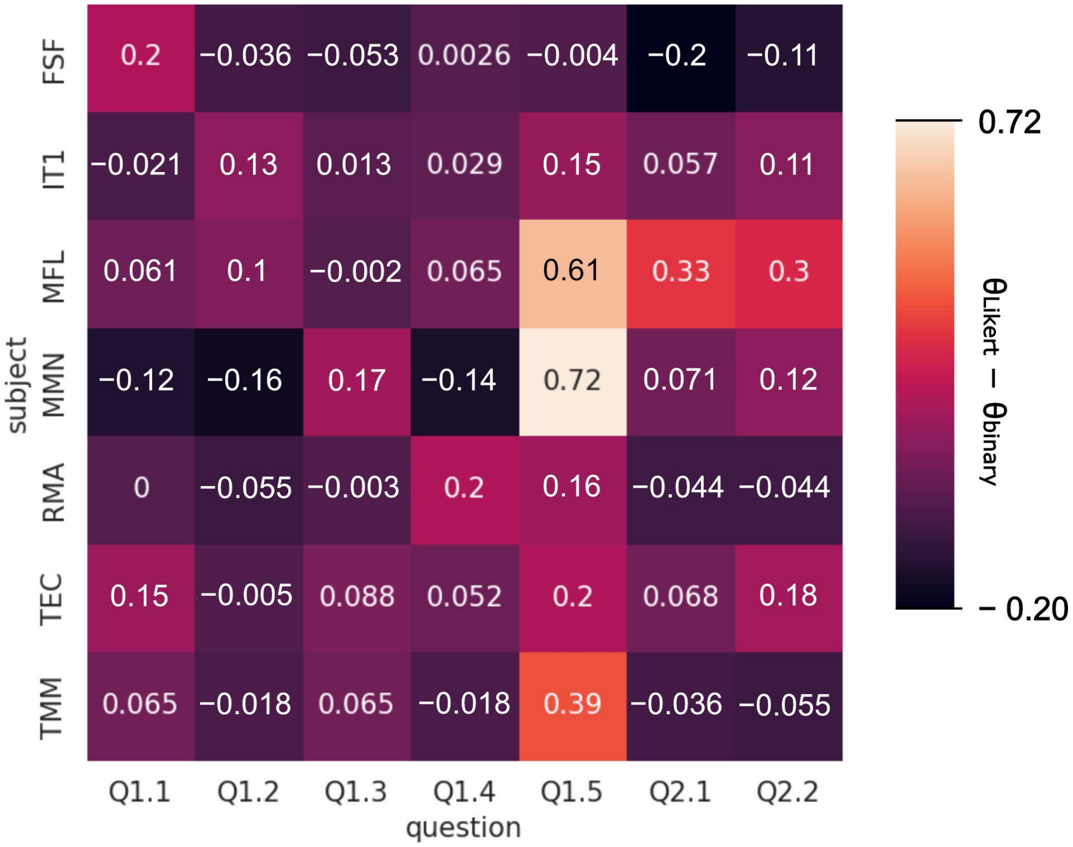


Figure 5. Distribution of affirmative and negative answers.

The differences between the results obtained can be summarized by comparing the frequency of success of both survey formats (Figure 6). The frequency of success is defined for each subject and question as the total number of affirmative responses divided by the total number of responses collected n. The results by subject in the rows show that the first subject represents a very similar visual result. The rest of the subjects, in general, also present a very similar appearance (e.g., FSF (Fundamentals of Manufacturing Systems and Technologies); RMA (Resistance of Materials); TEC (Theory of Structures and Industrial Constructions); and IT1 (Thermal Engineering I). However, there are areas where there is a clear difference, such as Q1.5 for the subjects MFL (Fluid Mechanics), MMN (Naval Engines and Machinery), and TMM (Theory of Machines and Mechanisms). Beyond colors, it is necessary to use some kind of metric to help discern if the differences are significant. This can be improved by looking at the differences between the two surveys.



subject	Q1.1	Q1.2	Q1.3	Q1.4	Q1.5	Q2.1	Q2.2
FSF	0.2	−0.036	−0.053	0.0026	−0.004	−0.2	−0.11
IT1	−0.021	0.13	0.013	0.029	0.15	0.057	0.11
MFL	0.061	0.1	−0.002	0.065	0.61	0.33	0.3
MMN	−0.12	−0.16	0.17	−0.14	0.72	0.071	0.12
RMA	0	−0.055	−0.003	0.2	0.16	−0.044	−0.044
TEC	0.15	−0.005	0.088	0.052	0.2	0.068	0.18
TMM	0.065	−0.018	0.065	−0.018	0.39	−0.036	−0.055

Figure 6. Difference between frequency of success for each question and subject.

3.2. Bayes Factor

Using Bayes factors, the equivalence between both survey formats can be deduced by comparing the success ratios. Figure 7 reports the values of BF_{01}. Of the 49 questions on which results were collected in both survey formats, the Bayes factor allowed the null hypothesis to be accepted for 24 questions (49%), according to the established criteria. In all of them, the value of the Bayes factor was greater than 3. In 18 others (36%), the Bayes factor did not allow the null hypothesis to be rejected or accepted. The null hypothesis was rejected in only seven (14%) questions where the Bayes factor was less than 1/3. The rejection quantity is the most limiting factor since it reflects the worst-case scenario. No conclusion can be drawn about the 36% that neither rejected nor accepted the null hypothesis with the data collected. In this case, the evidence in favor of the null hypothesis is also quantified.

We reiterate that when we talk about "equivalence", we are not talking about the equality of information collected. It is clear that the Likert scale contains more information because five values are collected. However, when it comes to quantifying the success ratio, set at 3 on the Likert scale, the results are actually equivalent. For example, in the MMN subject, question Q1.5 has a $BF_{01} = 0.11$ associated with it. This means that the data are about 1/0.11 ≈ 9 times more plausible under the alternative hypothesis than under the null hypothesis. It is curious to note that four of the seven questions where the null hypothesis was rejected were associated with question Q1.5, which asks students about teacher coordination. It may be that in this question opinions are naturally polarized due to the binary format. It is worth remembering that the Likert format allows the selection of a midpoint that can be interpreted as "fair" or a "fair pass". It could be that the polarization of the binary format allows for a more accurate collection of student opinions.

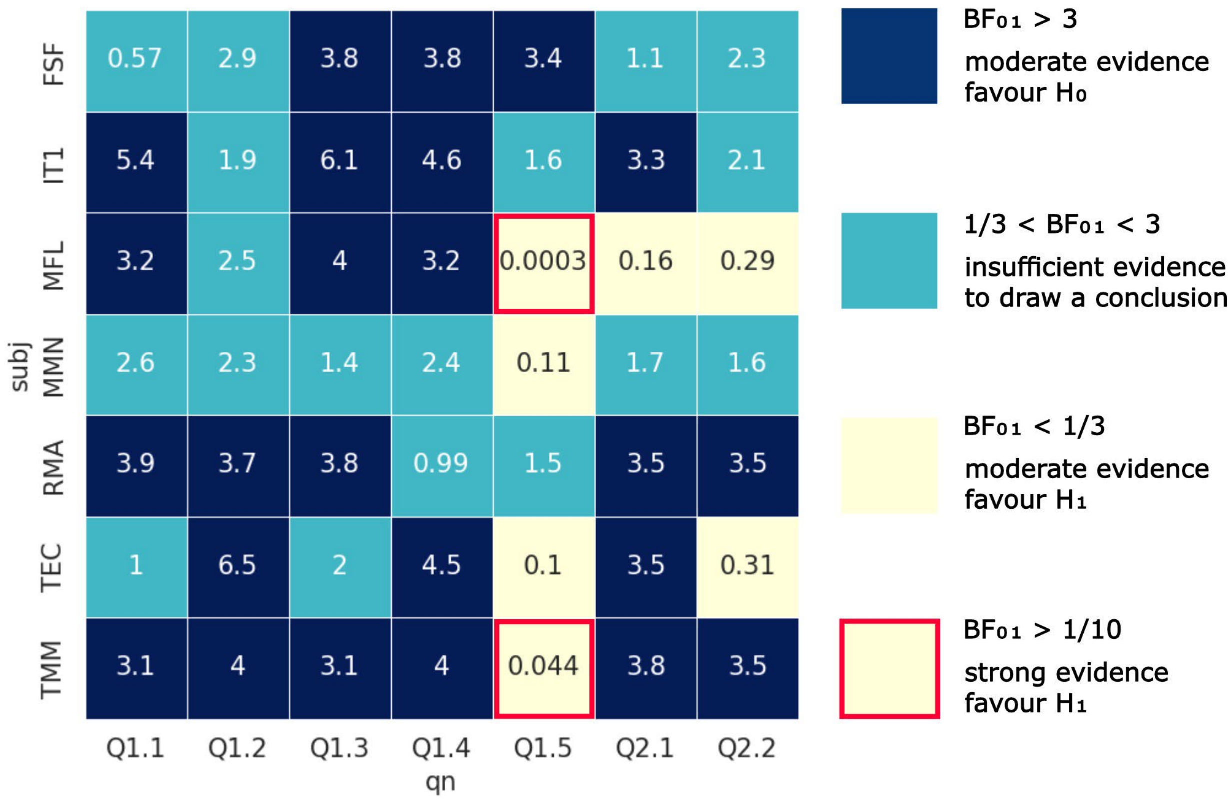


Figure 7. Bayes factor BF_{01} by subject and question.

Figure 8 represents the null hypothesis posterior probabilities. The pattern is similar to the Bayes factor figure. It is easy to see the questions where the alternative hypothesis was accepted (MFL: Q1.5, Q2.1, Q2.2; MMN: Q1.5; TEC: Q1.5, Q2.2; TMM: Q1.5). In all of them, the probability is below 25 %. Also, the questions where the null hypothesis was accepted can be distinguished. They have probabilities above 75%, reaching the maximum value of 86%.

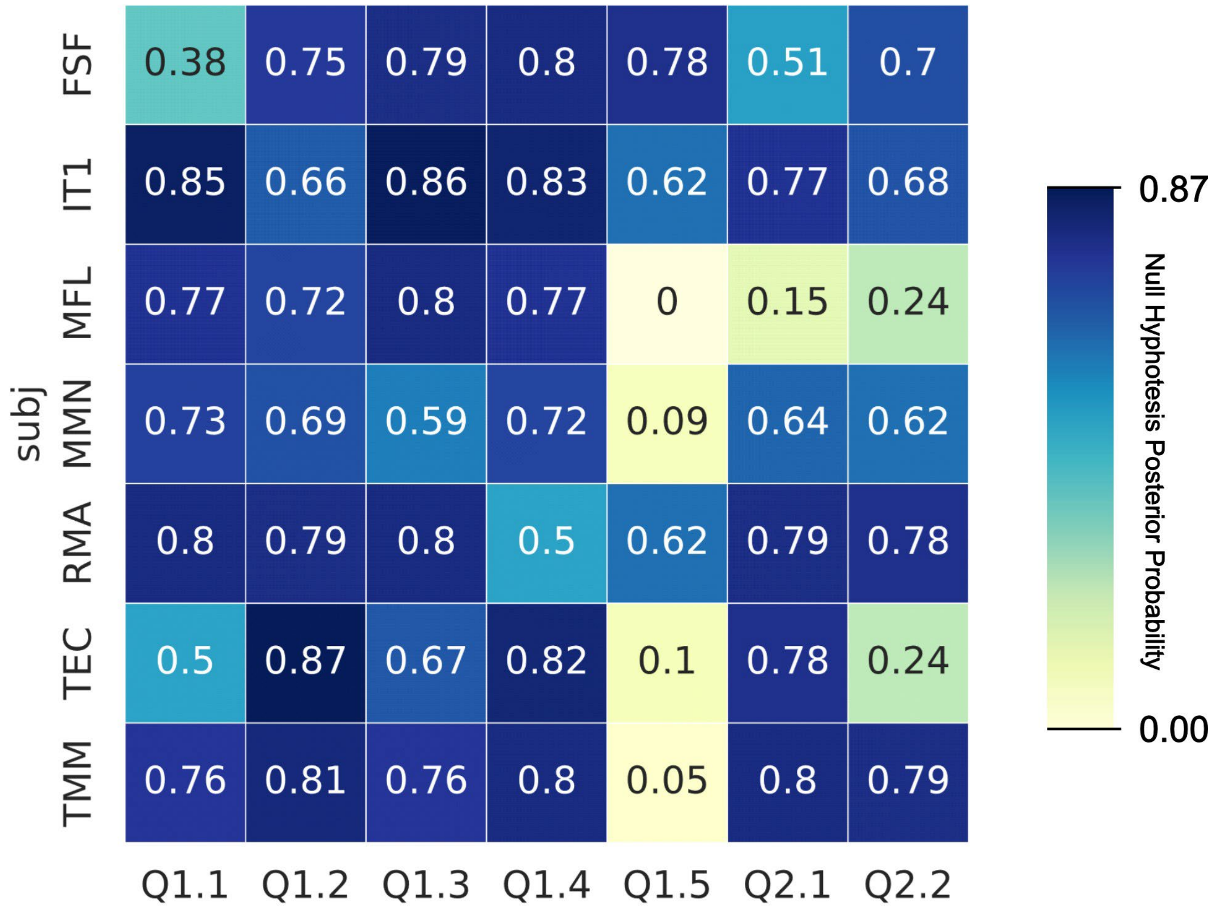


Figure 8. Null hypothesis posterior probabilities.

3.3. Student Opinions

The opinion surveys show a clear preference for the binary scale over the Likert scale (Figure 9). Recall that the survey focused on comparing the ease (Q3.1), speed (Q3.2), and pleasantness (Q3.3) of the two formats. Responses were elicited using "like/dislike". The figure represents the percentage of likes received out of the total number of responses received. The sample size corresponds to the size of the binary surveys, since this opinion questionnaire was an add-on. For all subjects, we observed that the minimum percentage of likes corresponded to 80% for all questions. Regarding speed, the percentage reached 100% in almost all subjects. However, these figures should be taken with caution. Despite the fact that the questions asked students to rate each "scale", the binary-scale questionnaire (10 questions) was considerably shorter than the Likert-scale version (16 questions per teacher). This may have influenced some of the opinions. However, the widespread use of the binary scale in social networks and hence the familiarity of using this format should not be forgotten. Students are more accustomed to choosing between two options. Choosing between five takes a greater effort for them because they are asked to be more precise in their verdict. For all these reasons, a clear preference is given to the binary format.

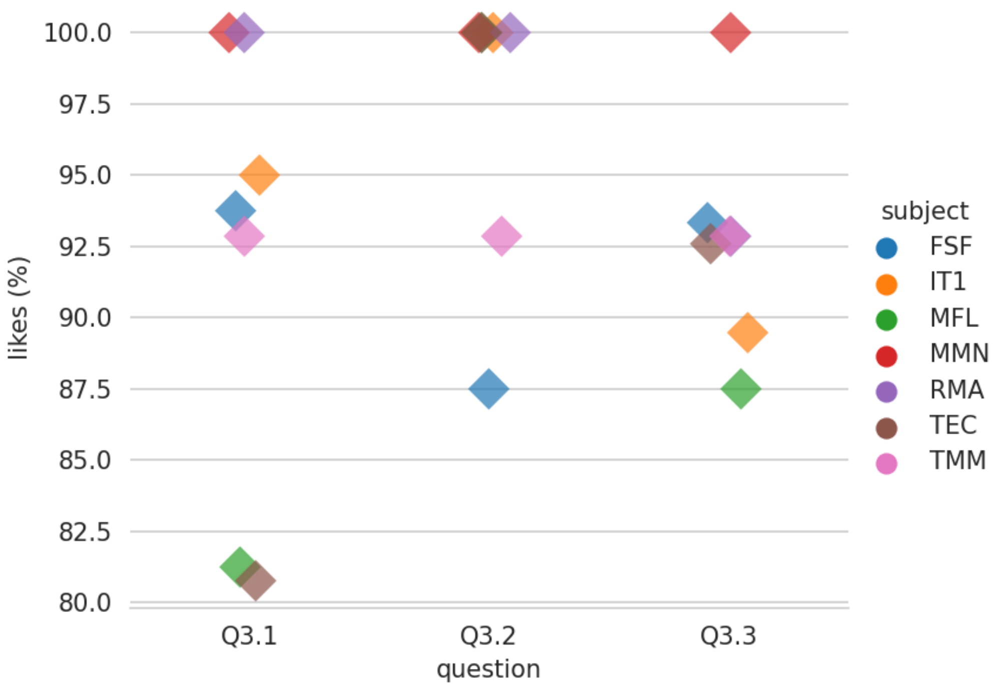


Figure 9. Student feedback on survey format.

4. Conclusions

In this work, two types of scales for collecting student opinions have been compared. In the comparison of the expected value collected by both types, the hypothesis that they were equal was rejected in only 14% of the questions collected. Equality means that the percentage of positive responses to a question is equal in both scenarios. This represents a very low percentage. It should also be noted that these results were achieved with a low volume of data. The sample size of the binary surveys was clearly lower. In addition, the opinions collected showed a clear preference for the binary format. This greater preference may lead to greater voluntary participation. It could even produce scenarios where the collection of these opinions is purely online, reducing the costs associated with hiring interviewers. In light of these results, we recommend preference for the binary scale.

The recommendation in favor of using the binary scale is made with an awareness of the loss of information obtained. The Likert scale collects five values, whereas the binary scale collects two. The responses elicited by the Likert scale can give rise to a probability mass function, and various characteristics that define its shape (skewness, kurtosis, etc.) can be studied. Although the binary scale can also be analyzed, the variety of values of the Likert scale brings more richness. However, this is not the usual process to analyze this information. It is common to use a figure that represents a success rate to study the evolution of opinions over the years. This is usually the expected value of responses to a question. The common use of the Likert scale adds another point in favor of the binary scale. The metric usually obtained from the Likert scale can also be obtained using the binary scale.

For future research, three paths can be established. The first is to test the method in more scenarios. This would imply the use of computational methods. Although multiple subjects from various courses have been evaluated, there is always an option for the collection of new data and the re-evaluation of results. The second is the use of other approaches to hypothesis testing, both frequentist and Bayesian. Finally, approaching the problem from different points of view can reinforce the idea of using the binary scale. In addition, this could lead to a comparison work that would serve to confront and analyze the differences and similarities between the scales.

Author Contributions: Methodology, A.S.-G., M.Á.-H., E.A. and J.R.R.; Validation, J.R.R.; Formal analysis, M.Á.-H. and E.A.; Data curation, A.S.-G. and M.Á.-H.; Writing—original draft, A.S.-G., E.A. and J.R.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Acknowledgments: The authors thank the Defense University Center at the Spanish Naval Academy (CUD-ENM) for the support provided for this research. The authors thank the Transformative Education: Science, Communication and Society UNESCO chair support.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Van Der Eijk, C. Measuring Agreement in Ordered Rating Scales. *Qual. Quant.* **2001**, *35*, 325–341. https://doi.org/10.1023/A:1010374114305.
2. Dolnicar, S. Asking Good Survey Questions. *J. Travel Res.* **2013**, *52*, 551–574. https://doi.org/10.1177/0047287513479842.
3. Johnson, M.D.; Lehmann, D.R.; Horne, D.R. The Effects of Fatigue on Judgments of Interproduct Similarity. *Int. J. Res. Mark.* **1990**, *7*, 35–43. https://doi.org/10.1016/0167-8116(90)90030-Q.
4. Drolet, A.L.; Morrison, D.G. Do We Really Need Multiple-Item Measures in Service Research? *J. Serv. Res.* **2001**, *3*, 196–204. https://doi.org/10.1177/109467050133001.
5. Komorita, S.S.; Graham, W.K. Number of Scale Points and the Reliability of Scales. *Educ. Psychol. Meas.* **1965**, *25*, 987–995. https://doi.org/10.1177/001316446502500404.
6. Martin, W.S.; Fruchter, B.; Mathis, W.J. An Investigation of the Effect of the Number of Scale Intervals on Principal Components Factor Analysis. *Educ. Psychol. Meas.* **1974**, *34*, 537–545. https://doi.org/10.1177/001316447403400307.
7. Grassi, M.; Nucera, A.; Zanolin, E.; Omenaas, E.; Josep; Anto, M.; Leynaert, B. Performance Comparison of Likert and Binary Formats of SF-36 Version 1.6 across ECRHS II Adults Populations. *Value Health* **2007**, *10*, 478–488. https://doi.org/10.1111/j.1524-4733.2007.00203.x.
8. Dolnicar, S.; Grün, B. How Constrained a Response: A Comparison of Binary, Ordinal and Metric Answer Formats. *J. Retail. Consum. Serv.* **2007**, *14*, 108–122. https://doi.org/10.1016/j.jretconser.2006.09.006.
9. Zax, M.; Takahashi, S. Cultural Influences on Response Style: Comparisons of Japanese and American College Students. *J. Soc. Psychol.* **1967**, *71*, 3–10. https://doi.org/10.1080/00224545.1967.9919760.
10. Paulhus, D.L. Measurement and Control of Response Bias. In *Measures of Personality and Social Psychological Attitudes*; Elsevier: Amsterdam, The Netherlands, 1991; pp. 17–59. https://doi.org/10.1016/B978-0-12-590241-0.50006-X.
11. Welkenhuysen-Gybels, J.; Billiet, J.; Cambré, B. Adjustment for Acquiescence in the Assessment of the Construct Equivalence of Likert-Type Score Items. *J. Cross-Cult. Psychol.* **2003**, *34*, 702–722. https://doi.org/10.1177/0022022103257070.
12. Fisher, R.A. Statistical Methods for Research Workers. In *Breakthroughs in Statistics*; Kotz, S., Johnson, N.L., Eds.; Springer Series in Statistics; Springer: New York, NY, USA, 1935; pp. 66–70. https://doi.org/10.1007/978-1-4612-4380-9_6.
13. Nuzzo, R. Scientific Method: Statistical Errors. *Nature* **2014**, *506*, 150–152. https://doi.org/10.1038/506150a.
14. Altman, D.G.; Bland, J.M. Statistics Notes: Absence of Evidence Is Not Evidence of Absence. *BMJ* **1995**, *311*, 485. https://doi.org/10.1136/bmj.311.7003.485.
15. Mark, A.; Baguley, T. Prior Approval: The Growth of Bayesian Methods in Psychology. *Br. J. Math. Stat. Psychol.* **2013**, *66*, 1–7. https://doi.org/10.1111/bmsp.12004.
16. Kruschke, J.K.; Aguinis, H.; Joo, H. The Time Has Come. *Organ. Res. Methods* **2012**, *15*, 722–752. https://doi.org/10.1177/1094428112457829.
17. Wagenmakers, E.-J.; Morey, R.D.; Lee, M.D. Bayesian Benefits for the Pragmatic Researcher. *Curr. Dir. Psychol. Sci.* **2016**, *25*, 169–176. https://doi.org/10.1177/0963721416643289.
18. Dienes, Z.; Mclatchie, N. Four Reasons to Prefer Bayesian Analyses over Significance Testing. *Psychon. Bull. Rev.* **2018**, *25*, 207–218. https://doi.org/10.3758/s13423-017-1266-z.
19. Makowski, D.; Ben-Shachar, M.S.; Chen, S.H.A.; Lüdecke, D. Indices of Effect Existence and Significance in the Bayesian Framework. *Front. Psychol.* **2019**, *10*, 2767. https://doi.org/10.3389/fpsyg.2019.02767.
20. Simmons, J.P.; Nelson, L.D.; Simonsohn, U. False-Positive Psychology. *Psychol. Sci.* **2011**, *22*, 1359–1366. https://doi.org/10.1177/0956797611417632.
21. Szucs, D.; Ioannidis, J.P.A. Empirical Assessment of Published Effect Sizes and Power in the Recent Cognitive Neuroscience and Psychology Literature. *PLoS Biol.* **2017**, *15*, e2000797. https://doi.org/10.1371/journal.pbio.2000797.
22. Kass, R.E.; Raftery, A.E. Bayes Factors. *J. Am. Stat. Assoc.* **1995**, *90*, 773–795. https://doi.org/10.1080/01621459.1995.10476572.

23. Morey, R.D.; Romeijn, J.-W.; Rouder, J.N. The Philosophy of Bayes Factors and the Quantification of Statistical Evidence. *J. Math. Psychol.* **2016**, *72*, 6–18. https://doi.org/10.1016/j.jmp.2015.11.001.
24. Kruschke, J.K. Rejecting or Accepting Parameter Values in Bayesian Estimation. *Adv. Methods Pract. Psychol. Sci.* **2018**, *1*, 270–280. https://doi.org/10.1177/2515245918771304.
25. De Bragança Pereira, C.; Stern, J. Evidence and Credibility: Full Bayesian Significance Test for Precise Hypotheses. *Entropy* **1999**, *1*, 99–110. https://doi.org/10.3390/e1040099.
26. Kelter, R. Analysis of Bayesian Posterior Significance and Effect Size Indices for the Two-Sample t-Test to Support Reproducible Medical Research. *BMC Med. Res. Methodol.* **2020**, *20*, 88. https://doi.org/10.1186/s12874-020-00968-2.
27. Capik, C.; Gozum, S. Psychometric features of an assessment instrument with likert and dichotomous response formats. *Public Health Nurs.* **2015**, *32*, 81–86. https://doi.org/10.1111/phn.12156.
28. Lee, J.W.; Jones, P.S.; Mineyama, Y.; Zhang, X.E. Cultural differences in responses to a Likert scale. *Res. Nurs. Health* **2002**, *25*, 295–306. https://doi.org/10.1002/nur.10041.
29. Wrinch, D.; Jeffreys, H. On Certain Fundamental Principles of Scientific Inquiry, (Second Paper). *Lond. Edinb. Dublin Philos. Mag. J. Sci.* **1923**, *45*, 368–374. https://doi.org/10.1080/14786442308634125.
30. Jeffreys, H. *Theory of Probability (Oxford Classic Texts in the Physical Sciences)*; Paperback; Oxford University Press: Oxford, UK, 1998.
31. Kelter, R. How to choose between different Bayesian posterior indices for hypothesis testing in practice. *Multivar. Behav. Res.* **2023**, *58*, 160–188. https://doi.org/10.1080/00273171.2021.1967716.
32. Wagenmakers, E.-J.; Lodewyckx, T.; Kuriyal, H.; Grasman, R. Bayesian Hypothesis Testing for Psychologists: A Tutorial on the Savage–Dickey Method. *Cogn. Psychol.* **2010**, *60*, 158–189. https://doi.org/10.1016/j.cogpsych.2009.12.001.
33. Dickey, J.M.; Lientz, B.P. The weighted likelihood ratio, sharp hypotheses about chances, the order of a Markov chain. *Ann. Math. Stat.* **1970**, *41*, 214–226.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.