2017 ■ Vol. 13 ■ no. 1

Interpretation of the point-biserial correlation coefficient in the context of a school examination

Vincent LeBlanc[a], ✉ and Michael A. A. Cox[b]

[a]University of Ottawa
[b]Newcastle University

Abstract ■ When creating an examination for educational purposes, one must make sure that the entire curriculum is covered. However, it can be difficult to do so without requiring a large number of questions. Also, it is desirable for those questions to be capable of accurately discriminating students who understand from those who do not, without being too difficult or easy. A practical way of identifying questions that fit these criteria is to study the point-biserial correlation (r_{pb}) between the success on a question and the number of questions correctly answered. This tutorial explains what the r_{pb} is and how to use it through the interpretation of effect sizes and significance testing applied to real data. It also presents the corrected point-biserial correlation (r^*_{pb}), which is more suited for significance testing when one of the variables is partially determined by the other.

Keywords ■ Tutorial, Examination, Assessment, Point-Biserial Correlation, Corrected Point-Biserial Correlation.

✉ vlebl017@uottawa.ca

VLB: 0000-0003-0492-5564; *MAAC*: 0000-0001-7344-2393

10.20982/tqmp.13.1.p046

Acting Editor ■ Denis Cousineau (Université d'Ottawa)

Reviewers
■ One anonymous reviewer

Introduction

When evaluating a large group of students, both the teachers and the students like to keep the number of questions to a minimum. Teachers wishing to shorten their examinations should select questions that are reasonably difficult and that are able to discriminate students who understand from those who do not. Very hard questions might discriminate the top students from the rest, but will not provide information on what the average or weaker students understand. It will also result in a majority of low scores, with very few average and high scores. The inverse is not more desirable; very easy questions will only discriminate the weaker students from the rest and will produce a majority of high scores with very few average or low scores. Fortunately, there are methods to identify which questions of an examination are reasonably difficult and are accurate in their discrimination at all levels of understanding.

One of these methods consists in studying the Pearson correlation (r) between correctly answering a question and the number of correctly answered questions (the "score") on the examination. If a question is relevant and well formulated, correctly answering it will have a strong positive correlation with the score; those who answer it correctly should, on average, have a higher score than those who did not.

However, r can only be calculated for two continuous variables, which is not the case here because the success or failure on a question is dichotomous. To study the correlation between a dichotomous and a continuous variable, we must turn to a special instance of the Pearson correlation, called the point-biserial correlation, r_{pb} (not to be confused with the biserial correlation, which is used when one of the variables is artificially dichotomized). The r_{pb} is mathematically equivalent to r, but has a more intuitive formula which provides insights on what constitutes a "good" question.

In this tutorial, we define, compute and interpret r_{pb} in the context of improving an examination. We then discuss methods to interpret it as an effect size and to test its significance. We follow with the introduction of the corrected r_{pb}, which is more appropriate for our context and can be adequately tested for significance. We conclude on a comparison of the two methods of significance testing

presented in this paper with a naïve t-test on r_{pb}.

Defining r_{pb} as a special instance of r

The Pearson correlation, r, is a measure of the linear dependence between two variables. It is one of the most commonly used statistical tools, especially in the social sciences (see Meyer et al., 2001, for a review in psychology). It is defined as

$$r = \frac{cov\,(x,\ y)}{s\,(x) \times s\,(y)}, \tag{1}$$

where $cov\,(x,\ y)$ is the covariance between scores x and scores y, and $s\,(x)$ and $s\,(y)$ are the standard deviations of the scores in x and y. This definition expands to

$$r = \frac{n\sum(x_i y_i) - \sum x_i \times \sum y_i}{\sqrt{n\sum x_i^2 - \left(\sum x_i\right)^2} \times \sqrt{n\sum y_i^2 - \left(\sum y_i\right)^2}}, \tag{2}$$

in which all the summations are over the n elements, x_i and y_i, $i = 1, 2, \ldots, n$, are two continuous variables between which the correlation is being computed, and n is the number of $(x,\ y)$ pairs.

If we wish to study the correlation between a continuous and a dichotomous variable, we must instead use r_{pb}. Let us start by defining the continuous variable, x, as the score of a student on the test, and the dichotomous variable, y, as an indication of whether the student answered the question incorrectly (scored 0) or correctly (scored 1). Because we chose 0 and 1 as values of y, the average of y is the proportion of students who answered the question correctly, denoted p. Conversely, the proportion of students who answered incorrectly, denoted q, is $1 - p$. This choice, along with a few algebraic manipulations, allows to rewrite Equation 2 into

$$r_{pb} = \sqrt{pq} \times \frac{\mu_1 - \mu_0}{\sigma}, \tag{3}$$

where μ_1 and μ_0 are the average scores of the students who answered the question correctly and incorrectly, and σ is the population's standard deviation of the scores (computed by dividing the sum of squares by n, not by $n-1$). An intuitive explanation for using the population's standard deviation instead of the sample's is that we study all the students who took the examination, not a sample. Those interested in a proof that Equations 1, 2 and 3 are equivalent can easily find one online.

The first part of Equation 3, $\sqrt{pq}$, shows why it is desirable to have reasonably difficult questions on an examination. Because $p + q = 1$, the maximal value of $\sqrt{pq} = 0.5$, when $p = q = 0.5$. However, examinations composed solely of questions with a success rate of 50% would have a group average of 50%. In order to maintain both a reasonable class average and a high value of r_{pb}, we recommend to select questions where $0.2 \leq p \leq 0.8$. These values result in $\sqrt{pq} \geq 0.4$, which isn't too far from the maximal value of 0.5; questions with p beyond these values can still be used, but will likely have a small r_{pb} (when $p = 0$ or $p = 1$, the result of Equation 3 is 0). Also, one should ponder on the validity or pertinence of questions that are correctly answered or failed by more than 80% of the students before adding them to their examination. Still, these values are only guidelines and very hard or easy questions could be relevant in some contexts. To summarize, you can use questions of any level of difficulty to craft an examination with the desired group average, but only questions with an acceptable value of $\sqrt{pq}$ can potentially be identified as strongly correlated with the score.

The second part of Equation 3, $\frac{\mu_1 - \mu_0}{\sigma}$, is the normalized distance between the average scores of the students who answered the question correctly and incorrectly. Whether the r_{pb} is positive or negative depends on this term; it will be positive if the average of those who answered correctly is greater, and negative if it is smaller. This term also shows how the magnitude of the correlation increases with the difference between the two averages.

It would have been hard to draw any conclusion on what properties of a question lead to a higher correlation by looking at Equations 1 and 2. Equation 3, however, is much more transparent on that matter. We will now apply it to real data.

Computing r_{pb} on sample data

Attached to this manuscript is a file that contains the results of 165 students (n) on a 50 questions (k) examination carried out by one of the authors. In this file, each line corresponds to a student, and each column to a question. The sum of a line is x, the student's score on 50. Calculating the scores of all students allows us to find the population's standard deviation of the scores, which is constant and will be used to compute the r_{pb} of all 50 questions ($\sigma = \sqrt{\sum (x - \mu)^2 / n} \cong 6.245$). The average of a column is the p for that question (and $1 - p = q$). To compute μ_1, we average the scores of the students who answered that question correctly (and similarly for μ_0).

As an example: there are 57 students out of 165 who answered Question 4 correctly, thus $p \cong .345$, and the average score of those 57 students is $\mu_1 \cong 31.702$. Similarly, we find $q \cong .655$ and $\mu_0 \cong 26.954$. Applying Equation 3, we obtain

$$r_{pb}\,(Q4) \cong \sqrt{.345 \times .655} \times \frac{31.702 - 26.954}{6.245} \cong .362.$$

Using software to compute r_{pb}

We will now see how to use software to compute r_{pb}. Because $r_{pb} \equiv r$, we will simply use the correlation functions

included in those software. In the additional content of this manuscript, you will find an Excel spreadsheet and an SPSS file that use the methods described in this section.

Computing r_{pb} with Excel Prior to computing the r_{pb}, you must find the score of each student. If you are using our data and want the scores in column AY, the cell AY2 would contain

```
=SUM(A2:AX2).
```

Once you have the scores of all students, correlate them with their result on Question 4 in column D

```
=CORREL(D2:D166, AY2:AY166)
```

Computing r_{pb} with SPSS. Again, the first step will be to calculate the scores. Supposing you use our data and that the question variables are named from Q1 to Q50, this will be done by

```
COMPUTE Total = SUM(Q1 to Q50).
```

You can then compute the correlation between Question 4 and the scores with

```
CORRELATIONS VARIABLES = Q4 Total.
```

Interpreting r_{pb}

The goal of computing the r_{pb} of a question is to determine if it should be kept in the examination, based on the magnitude of its correlation and on whether this correlation is positive or negative. A question with a large positive correlation is a good predictor of the score, and is therefore informative. A question with a large negative correlation, on the other hand, indicates that those who answered it correctly usually have a lower score. If every question had a large negative correlation, then correctly answering all the questions would result in a negative score (which is impossible). Hence, questions with large negative correlations should be investigated for errors in their formulation or in the way they were corrected, and removed from the examination. Also, note that correlation does not mean causation; suppose that a student correctly answers only one question, and that this question has a strong positive correlation with the total score. Even if he did answer it correctly, his score will still be very low.

Before we proceed with the interpretation of r_{pb}, we offer a word of warning on removing questions with a small, non-significant correlation (positive or negative). The r_{pb} is not an intrinsic property of a question, but a contextual one determined by the other questions in the examination. This means that the r_{pb} of a question will change when you remove or add questions; altering the examination too much might turn a strong positive correlation into a weak one. In other words, questions with weak correlations set the context for questions with strong correlations to exist and should not be removed hastily. If you believe that a question with a weak r_{pb} should be removed from your examination, you should use more sophisticated methods to confirm your intuition. The combined use of correspondence analysis and procrustes analysis, for example, is a method that describes how questions relate to each other and can detect if this relationship changes through different samples. A weakly correlated question that is inconsistent between samples could be eliminated safely. For an in-depth coverage of multidimensional scaling, which touches on procrustes and correspondence analyses, see T. F. Cox and Cox (2000).

We will now look at different ways of interpreting a correlation as an effect size and of determining if it is significant.

Effect size of r_{pb}

To determine the impact of one variable on another (the "effect size"), we can look at the proportion of variance shared between both variables, which happens to be the definition of r (Equation 1). In other words, a large positive correlation means that the question is correctly answered more frequently by students with a high score than by those with a low score. Cohen's (1988) general rule of thumb is to consider $r = \pm .10$ as small, $r = \pm .30$ as medium and $r = \pm .50$ as large effect sizes, but these are only guidelines; a correlation of $.10$ between undergoing a brain surgery and permanent brain damage, for example, could be considered very large.

We now present how to determine the precision of a correlation by computing its confidence interval. We also offer two other measures that provide different interpretations of the value of a correlation.

Confidence interval. A confidence interval (CI) is a range that includes a proportion of the values we might expect to find if we repeated the same experiment many times. For example, if we choose $\alpha = .05$, we can find the bounds of a 95% CI, which would contain the average of 95 out of 100 repetitions of the same experiment. To determine these bounds, we would normally take the critical values from the Z score distribution for $\alpha/2 = .025$, which are ± 1.96 (or ± 2.576 if we wanted a 99% CI), and multiply them by the standard error (SE) of our data. However, because we are using values from the Z score distribution, we must first make sure that our data loosely follows this distribution. Specifically, in the case of correlations, we must adjust their variance to compensate for the fact that they only exist within the interval $[-1;\ 1]$, whereas the Z distribution goes from $[-\infty;\ +\infty]$. We do so by applying the

Fisher r to z' transformation:

$$\begin{aligned} z' &= \operatorname{arctanh}(r_{pb}) \\ &= \frac{\ln(1+r_{pb}) - \ln(1-r_{pb})}{2} \\ &= \ln\left(\sqrt{\frac{1+r_{pb}}{1-r_{pb}}}\right), \end{aligned} \tag{4}$$

where $\ln(a)$ is the natural logarithm of a. Once the correlations are in z' scores, we determine the standard error of z' with:

$$SE_{z'} = \frac{1}{\sqrt{n-3}}. \tag{5}$$

Next, we apply the regular method to calculate confidence intervals:

$$\begin{aligned} 95\%\ CI_{z'} &= [z' - 1.96 \times SE_{z'}\ ;\ z' + 1.96 \times SE_{z'}] \\ &= \left[z'_{-}\ ;\ z'_{+}\right]. \end{aligned} \tag{6}$$

However, we are interested in a confidence interval of correlations, not of transformed scores. Hence, we must transform our z' back into r_{pb} with the inverse of Equation 4:

$$r_{pb} = \tanh(z') = \frac{e^{2z'} - 1}{e^{2z'} + 1}, \tag{7}$$

where e is the base of the natural logarithm. Finally, we apply Equation 7 to Equation 6 to obtain our confidence interval:

$$95\%\ CI_{r_{pb}} = \left[\tanh\left(z'_{-}\right),\ \ \tanh\left(z'_{+}\right)\right]. \tag{8}$$

Let us continue our previous example by finding the 95% CI of Question 4's r_{pb}:

$$\begin{aligned} z'(Q4) &= \operatorname{arctanh}(.362) \cong .378 \\ SE_{z'}(Q4) &= \frac{1}{\sqrt{165-3}} \cong .079 \\ 95\%\ CI_{z'}(Q4) &\cong [.378 - 1.96 \times .079\ ;\ .378 + 1.96 \times .079] \\ &\cong [.225\ ;\ .533] \\ 95\%\ CI_{r_{pb}}(Q4) &\cong [\tanh(.225)\ ;\ \tanh(.533)] \\ &\cong [.221\ ;\ .487] \end{aligned}$$

In other words, we can expect that if that examination was passed 100 times, there would be 95 times where the r_{pb} of Question 4 would be between .221 and .487, inclusively.

Coefficient of determination. The coefficient of determination, r^2, is the proportion of variance that is shared by the variables being correlated. For example, Question 4 shares, or can account for, $.362^2 \cong 13.07\%$ of the variance of the scores (and vice-versa). It is worth noting that some oppose the use of squared correlations (D'Andrade & Dart, 1990; Ozer, 1985) or think that r should be preferred to r^2 as an effect size measure (Cohen, 1988; Hunter & Schmidt, 1989; Kvalseth, 1985; Rosenthal, 1991).

Forecasting efficiency. Another way of understanding r is in term of the forecasting efficiency (FE; Vorhees, 1926), which indicates to what degree relying on x to predict y is better than a blind guess. It is computed using the following equation:

$$FE = 1 - \sqrt{1 - r^2}. \tag{9}$$

Looking at Question 4, the FE suggests that relying on this question to predict a student's score is $1 - \sqrt{1 - .362^2} \cong 6.76\%$ better than a blind guess. Figure 1 puts this value in perspective.

Significance testing of r_{pb}

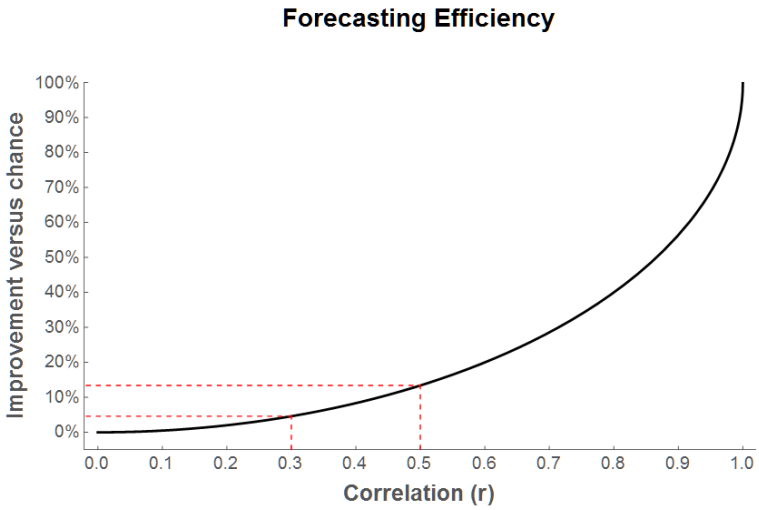
Although measuring an effect size is informative, it is also useful to know when an effect is large enough to be statistically different from no effect at all. Null hypothesis significance tests are the most common methods to detect this; with correlations, we use Student's t-test with the null hypothesis (H_0) that the two variables are not correlated ($\rho = 0$). However, this assumption that the two variables are uncorrelated does not hold if one of the variables is partially determined by the other, which is the case here; the score is determined by the number of questions that were correctly answered.

In the context of assessing an examination, if the questions are graded in an "all-or-nothing" fashion or are all worth the same number of points, we can predict the magnitude of the "built-in association" between the questions and the score. Here, every one of the k questions of an examination should have a coefficient of determination $r^2 = 1/k$ (see Appendix A for a demonstration). In other words, if a question that tests language competencies was inserted in an examination on statistical knowledge, we would expect that question to have $r = \sqrt{1/k}$ with the total score instead of 0. Hence, $H_0 : \rho = 0$ is invalid and $H_0 : \rho = \sqrt{1/k}$ should be tested instead. Sadly, this is not possible using regular null hypothesis significance tests.

This expected value $\rho = \sqrt{1/k}$ is not valid in most examinations, where grades are not "all-or-nothing" and questions are worth different numbers of points. However, the built-in association remains for all sorts of grading schemes, as the problem is that the score is partially determined by the question it is being correlated with.

Hence, we present a first alternative to the t-test on r_{pb}: confidence intervals (CI) around ρ. By using Equations 4 to 8, we determine bounds that will be used as our critical values of r_{pb}. Any value more extreme than these critical values is then considered as significantly negative or positive. Because the CI is centered on a positive value instead of 0, it should have less Type I and Type II errors on positive and negative correlations, respectively, compared to blindly using a t-test on r_{pb}.

Figure 1 ■ Forecasting efficiency as a function of r. The forecasting efficiency is a measure of how relying on the independent variable to predict the dependent variable is better than a blind guess. It increases slowly, and reaches high values only with extremely high correlations. It can be hard to understand the meaning of $r = .5$, but "13.40% better than chance" is clearer.



Let us apply this method to our data, with $n = 165$, $k = 50$ and $\alpha = .05$, to create a 95% CI. Using Equations 4 to 8 on $r_{pb} = \sqrt{1/50}$, we find the following bounds

$$95\%\ CI_{r_{pb}}\left(\sqrt{1/50}\right) \cong [-.012;\ .288]\,. \tag{10}$$

Although this approach to testing the significance of r_{pb} is not the most adequate, it is "good enough" for non-critical situations. Its first drawback stems from the fact that the r to z' transformation aims to make the distribution closer to a Normal Distribution, which is why we choose our bounds to be (precisely) ± 1.95996 SE from the mean. However, a t-test assumes a Student Distribution with $df = n - 2 = 163$, with critical t-values ± 1.97462. Hence, we do not obtain the same critical r_{pb} for a t-test and a confidence interval. For example, in a context where $H_0 : \rho = 0$ is valid, the 95% CI suggests that $r_{pb(crit)} = \pm .152786$, but the t-test would use $r_{pb(crit)} = \pm .152847$.

Its second drawback is that it requires the knowledge of the expected value of ρ, which can be tedious to find depending on the grading scheme of the examination. A workaround is to convert the examination into one where $\rho = \sqrt{1/k}$ is true. This is achieved by giving a score of either 0 or 1 to each question depending on if the answer was incorrect or correct, as we have done since the beginning of this tutorial.

Using the corrected r_{pb}

We now present an alternative to r_{pb} that is more rigorous in our context, the corrected point-biserial correlation (r^*_{pb}), which studies the correlation between correctly answering a question and the score computed without including this question's result (Crocker & Algina, 1986). The r^*_{pb} for a question is calculated with the same equation as r_{pb}, but the score on that question is not included in the students' total scores. Hence, the averages are calculated on $k - 1$ questions and the standard deviation is different for each question. Using the asterisk to denote this exclusion, we calculate r^*_{pb} with

$$r^*_{pb} = \sqrt{pq} \times \frac{\mu^*_1 - \mu^*_0}{\sigma^*}. \tag{11}$$

Consequently, the built-in association between the question and total score is entirely removed, and $H_0 : \rho = 0$ becomes a valid hypothesis. Those interested in understanding the impact of the built-in association and the benefits of r^*_{pb} versus r_{pb} can find more information in Appendix A.

Computing r^*_{pb} is slightly more tedious compared to r_{pb}, as σ changes for each question. Still, by using Equation 3, we can calculate r^*_{pb} for Question 4:

$$r^*_{pb}\,(Q4) = \sqrt{.345 \times .655} \times \frac{30.702 - 26.954}{6.090} \cong .293, \tag{12}$$

Using software to compute r^*_{pb}

As it was the case for r_{pb}, you will find an Excel spreadsheet and SPSS syntax in this article's additional content that compute the value of r^*_{pb} for our sample data.

Computing r^*_{pb} esing Excel. To find the r^*_{pb} for Question 4, we first compute the scores in column AZ, excluding the score of Question 4 (in column D). Hence, the cell AZ2 should contain

```
=SUM(A2:C2) + SUM(E2:AX2)
```

Alternatively, if the total scores are already in column AY, then you can use

```
=AY2 - D2
```

We then correlate this score with the question using the same formula as before

```
=CORREL(D2:D166, AZ2:AZ166)
```

Computing r^*_{pb} using SPSS. In SPSS, with the question variables still named Q1 to Q50, run the following command to compute the scores without including Question 4

```
COMPUTE TotalQ4 = SUM(Q1 to Q3) + SUM(Q5
    to Q50).
```

Alternatively, if you still have the variable Total, you can use this command instead

```
COMPUTE TotalQ4 = Total - Q4.
```

Then, find r^*_{pb} by using the same formula as before,

```
CORRELATIONS VARIABLES= Q4 TotalQ4.
```

Effect size of r^*_{pb}

Just like we did for r (and r_{pb}), we can find a confidence interval, coefficient of determination and forecasting efficiency for r^*_{pb}. However, the interpretation of r^*_{pb} is slightly different from the one for r_{pb}. A strong positive r^*_{pb} means that the question is a good predictor of performance on an examination composed of the questions on which the score was computed, whereas a strong positive r_{pb} means that the question is a good predictor of the score on an examination including that question. This new measure can be used to pick questions that evaluate how prepared the students are for this specific test.

Using the equations described earlier, we find that the 95% CI of Question 4's r^*_{pb} is $[0.146;\ 0.426]$, that its coefficient of determination is 8.57% and that its FE is 4.38%.

Significance testing of r^*_{pb}

With r^*_{pb}, the null hypothesis $H_0 : \rho = 0$ is appropriate, which means we can use a t-test to determine its significance. Compared to a t-test on r_{pb}, it will have less Type II errors for negative correlations and less Type I errors for positive correlations. The t-value of a correlation is calculated with:

$$t = r \times \sqrt{\frac{n-2}{1-r^2}}. \tag{13}$$

Going back to our example, we can determine the critical values of r^*_{pb} for our examination. We start by finding the critical t-values, with $\alpha = 0.05$ and $df = 163$, which gives us $t_{crit} \cong \pm 1.97$. We then solve for our critical r^*_{pb} using Equation 10, which gives us $r^*_{pb(crit)} \cong \pm .153$. We conclude that Question 4 is a good predictor of success on the other questions, as $.293 > .153$.

Comparing the significance testing methods

We will now apply a t-test and calculate a confidence interval on both the r_{pb} and r^*_{pb} to analyze and interpret Questions 38, 39, 18 and 20. We have previously computed the critical correlation values for confidence intervals, which are $r_{pb(crit)} \cong \{-.011616,\ .287985\}$ and $r^*_{pb(crit)} \cong \pm .152786$, as well as those for t-tests, which are $r_{pb(crit)} \equiv r^*_{pb(crit)} \cong \pm .152847$. Table 1 shows the conclusions drawn from each test. Please note that we do not support using a t-test on r_{pb} or a confidence interval on r^*_{pb} to test their significance. We strongly recommend to use the t-test on r^*_{pb} or, at the very least, to use the confidence interval on r_{pb}.

We clearly see that Question 38 is an excellent predictor of both the score (r_{pb}) and of how well students performed on the 49 other questions (r^*_{pb}). Question 20, however, is seriously problematic; such a strong negative correlation means that a correct answer predicts a lower score and a bad performance on the rest of the questions. It should be removed from future examinations.

Next, we turn to Question 39. Here, all alternative methods agree that this question is not positively correlated with the score or to successfully answering the other questions. In this situation, doing a naïve t-test on r_{pb} would lead to a Type I error. Finally, Question 18 is more complex to interpret. It is negatively correlated with the score, as suggested by the CI on r_{pb}, but not significantly correlated with the performance on the 49 other questions, according to r^*_{pb}. In this situation, depending on the interpretation of interest to the evaluator, a naïve t-test could lead to a Type II error.

Conclusion

In this article, we presented the point-biserial correlation, which is to be used instead of the Pearson correlation when one variable is naturally dichotomous. We studied how to compute its confidence interval using Fisher's r to z' transformation and different ways of interpreting a correlation as an effect size. We also explained the problem of using a

Table 1 ■ Regular (r_{pb}) and corrected (r^*_{pb}) point-biserial correlations of questions 38, 39, 18 and 20. The correlations and interpretation drawn from a t-test and a 95% confidence interval on both values follow. A "+" indicates that the method would declare the correlation significantly positive; "-" and "0" are used for significantly negative and non-significant correlations, respectively. Although the significance of questions 38 and 20 is the same for all tests, we find that Question 39 would result in a false positive if tested naïvely. As for question 18, the decision to consider it as weakly or as significantly negative depends on the tester's goal.

	r_{pb}			r^*_{pb}		
Q#	r_{pb}	t-test	95% CI	r^*_{pb}	t-test	95% CI
38	0.513	+	+	0.468	+	+
39	0.173	+	0	0.102	0	0
18	-0.038	0	-	-0.113	0	0
20	-0.673	-	-	-0.713	-	-

t-test on data that is intrinsically correlated, such as the r_{pb} between the success on a question and a score composed of that question. Finally, we presented two alternative methods to test the significance of correlations in the academic context.

Beyond testing a questionnaire post-hoc with these methods, there exist a priori techniques to create strong tests. One of these, the Cognitive Diagnosis Models, consists in identifying the common content underlying certain items (DiBello, Roussos, & Stout, 2006; George & Robitzsch, 2015). The strength of this approach is that items can include increasingly complex concepts when the person taking the test shows he mastered the pre-requisite, simpler concepts. Another paradigm that aims to improve tests is the Item Response Theory (van der Linden & Hambleton, 1996). This technique specializes in psychometric scales and multiple-choice questionnaires that measure abstract, latent concepts. Those who are willing to significantly improve their examinations and tests should look into these two methods.

As a final note, we think this paper highlights the importance of understanding the nature of the variables we manipulate. Identifying one of the variables as binary justified the choice of using r_{pb} instead of r, and reflecting on the way the score is computed made us look for alternative approaches to studying significance, such as r^*_{pb}. Paying close attention to the characteristic of our data can lead to great discoveries and should be a skill we hone at any given occasion.

References

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences (2nd ed.)* Hillsdale: Routledge.

Cox, T. F. & Cox, M. A. A. (2000). *Multidimensional scaling (2nd ed.)* Boca Raton: Chapman and Hall/CRC. doi:10.1201/9781420036121

Crocker, L. M. & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart and Winston.

D'Andrade, R. & Dart, J. (1990). The interpretation of r versus r^2 or why percent of variance accounted for is a poor measure of size of effect. *Journal of Quantitative Anthropology*, *2*, 47–59.

DiBello, L. V., Roussos, L. A., & Stout, W. (2006). Review of cognitively diagnostic assessment and a summary of psychometric models. *Handbook of Statistics*, *26*(6), 979–1030. doi:10.1016/S0169-7161(06)26031-0

George, A. C. & Robitzsch, A. (2015). Cognitive diagnosis models in R: A didactic. *The Quantitative Methods for Psychology*, *11*(3), 189–205.

Hunter, J. E. & Schmidt, F. L. (1989). *Methods of meta-analysis: Correcting error and bias in research findings (1st ed.)* Newbury Park: Sage Publications.

Kvalseth, T. O. (1985). Cautionary note about r^2. *The American Statistician*, *39*(4), 279–285. doi:10.1080/00031305.1985.10479448

Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., Dies, R. R., ... Reed, G. M. (2001). Psychological testing and psychological assessment: A review of evidence and issues. *American Psychologist*, *56*(2), 128–165. doi:10.1037/0003-066X.56.2.128

Ozer, D. J. (1985). Correlation and the coefficient of determination. *Psychological Bulletin*, *97*(2), 307–315. doi:10.1037/0033-2909.97.2.307

Rosenthal, R. (1991). *Meta-analytic procedures for social research (applied social research methods) (revised ed)*. Newbury Park: SAGE Publications.

Tate, R. F. (1954). Correlation between a discrete and a continuous variable. point-biserial correlation. *The Annals of Mathematical Statistics*, *25*(3), 603–607. doi:10.1214/aoms/1177728730

van der Linden, W. J. & Hambleton, R. K. (1996). *Handbook of modern item response theory*. New York: Springer.

Vorhees, J. F. (1926). A graphic and tabular aid to interpreting correlation coefficients. *Monthly Weather Review*, *54*(10), 423–423. doi:10.1175/1520-0493(1926)54<423:AGATAT>2.0.CO;2

Table 2 ■ Comparison of predicted and observed parameters of correlations between the score on a question and the total score on an examination, as well as the Anderson-Darling test on observed data fitted to a Normal Distribution with the expected parameters. The observed values r_{pb} and σ^2 were computed on 100,000 correlations with a sample size of 1,000 students, when there are 10, 25, 50, 100 and 250 questions with a success rate $p = .5$. These results confirm that there is indeed a built-in association of $\sqrt{1/k}$ when studying the r_{pb} between the score on a question and the total score on the examination, making t-tests inadequate.

k	$\sqrt{1/k}$	r_{pb}	$\frac{(k-1)^2\times(k-0.5)}{1000k^3}$	σ^2	A-D p-value
10	.3162	.3163	7.695×10^{-4}	7.694×10^{-4}	.1085
25	.2000	.1999	9.032×10^{-4}	9.032×10^{-4}	.1750
50	.1414	.1414	9.508×10^{-4}	9.508×10^{-4}	.9263
100	.1000	.0999	9.752×10^{-4}	9.752×10^{-4}	.7821
250	.0632	.0632	9.900×10^{-4}	9.900×10^{-4}	.2464

Note. A-D: Based on the Anderson-Darling test

Appendix A: Inadequacy of t-test on r_{pb} for data with a built-in association

The computation of t-values on Pearson correlations is based on the assumptions that the correlations are distributed approximately normally and that the two variables are independent, or that the null hypothesis $H_0 : \rho = 0$ is valid. We know that r_{pb} is distributed normally when the samples are large enough (Tate, 1954), but the two variables are not independent when correlating a score with one of the questions that make up this score. In this situation, even when a question should not have any predictive power on the score, its correlation will not be zero because of the way the score is computed. We call this a "built-in association".

In what follows, we ran two simulations in which we studied the correlation between the success on a question and the score on an examination. Our goal was to show that, even in unrealistically optimal situations, applying a t-test on r_{pb} is not appropriate. Hence, each simulation was composed of 100,000 examinations with $n = 1000$ students, $k \in \{10, 25, 50, 100, 250\}$ questions and a fixed success rate of $p = .5$ for each question. We had first randomly assigned the success rates of every question such that $p \in [.2, .8]$, but found no impact on our main conclusions. For each student, we sampled a question score from a Binomial distribution $B[1, .5]$ and a remaining score from $B[k-1, .5]$.

In the first simulation, the total score was the sum of the question and remaining score, and so the correlation between the question and total score was the r_{pb}. In the second simulation, the total score was equal to the remaining score, and the correlation was the r^*_{pb}. Our results show that t-tests are not adequate for r_{pb} in this context, but that they are for r^*_{pb}.

Simulation 1

The goal of the first simulation was to confirm that there was indeed a built-in association in r_{pb}. Because each question is worth the same amount of points and is graded dichotomously, they should, on average, have $r^2_{pb} = 1/k$, such that $\sum_{i=1}^{k} r^2_{pb_i} = 1$, and thus $r_{pb} = \sqrt{1/k}$. Also, Tate (1954) showed that the variance of r_{pb} is

$$\sigma^2 = \frac{\left(1-r^2\right)^2}{n} \times \left(1 + r^2 \times \frac{1-6pq}{4pq}\right). \tag{14}$$

Because $p = .5$, $n = 1000$, and $r^2 = 1/k$, we can simplify Equation 14 to

$$\sigma^2 = \frac{(k-1)^2 \times (k-0.5)}{1000k^3}. \tag{15}$$

Hence, we expect that r_{pb} will follow the Normal distribution, $N\left[\sqrt{1/k}, \frac{(k-1)^2\times(k-0.5)}{1000k^3}\right]$.

As we can see in Table 2, the expected averages, variances and distributions of r_{pb} are observed, as shown by the Anderson-Darling tests, for all values of k. Figure 2 shows the distribution of r_{pb} when $k = 50$, which follows $N[0.1414, .0009508]$. The red lines indicate the critical r_{pb}values of a t-test with $\alpha = .05$ and $df = 998$ for $H_0 : \rho = 0$, such that $r_{pb(crit)} \cong \pm.062$. This figure shows clearly that a t-test is not appropriate for data with a built-in association;

Figure 2 ■ The r_{pb} follows a Normal distribution with $\mu = \sqrt{1/k}$. This figure illustrates why this is a problem; if we use a t-test to detect significance, we will make many Type I and Type II errors. The red dotted lines indicate the critical t-values for a two-sided t-test with $\alpha = .05$, which are definitely not appropriate for this data ($.9947 \gg .05$).

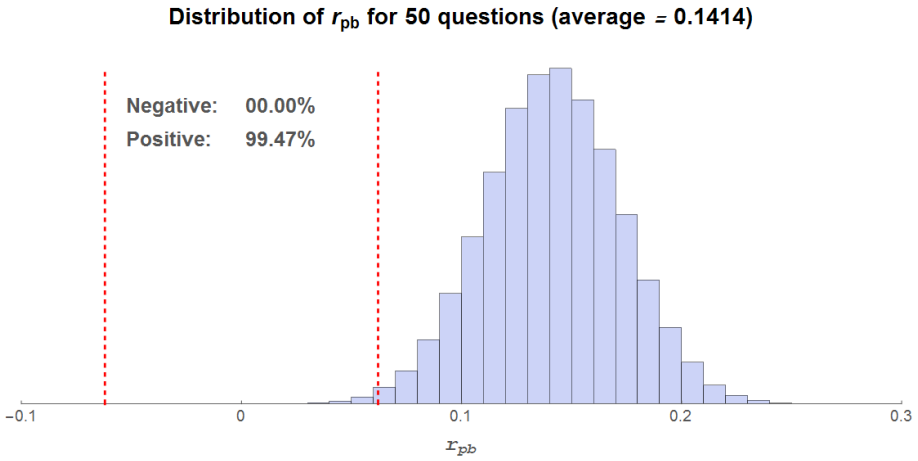


Table 3 ■ Observed parameters of correlations between the score on a question and the score of the remaining questions of an examination, as well as the Anderson-Darling comparing the observed data to $N\,[0,\ \ 0.001]$. The observed values r^*_{pb} and σ^2 were computed on 100,000 correlations with a sample size of 1,000 students, when there are 10, 25, 50, 100 and 250 questions with a success rate $p = .5$. These results confirm that t-tests are adequate for r^*_{pb}, as its distribution is not different from $N\,[0,\ \ 0.001]$.

k	r^*_{pb}	σ^2	A-D p-value
10	- .0001	0.001	.7915
25	- .0000	0.001	.6827
50	- .0000	0.001	.6934
100	- .0001	0.001	.4300
250	- .0001	0.001	.2763

Note. A-D: Based on the Anderson-Darling test

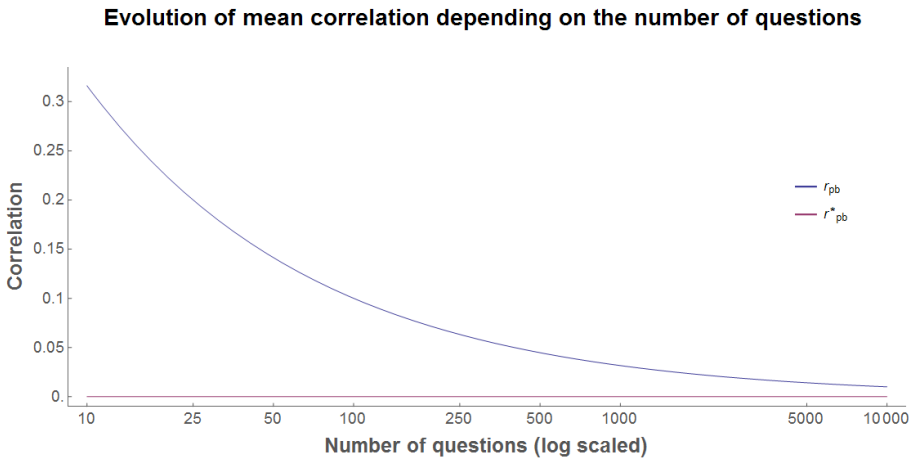
an examination with 50 questions would typically have 50 "significantly strong positive correlations" according to t-tests on r_{pb}. Finally, as shown in Figure 3, even with a large number of questions, $H_0 : \rho = 0$ is not appropriate for data with a built-in association.

Simulation 2

In Simulation 2, we correlated the question with the remaining scores instead of the total scores. Because the question and remaining scores are independent, we predict that $H_0 : \rho = 0$ will be adequate. Also, note that when $r_{pb} = 0$, Equation 14 becomes $\sigma^2 = 1/n$. Hence, for any value of k, r^*_{pb} should follow the same distribution, $N\,[0,\ \ .001]$. We ran an Anderson-Darling test on each condition to test that hypothesis, which was confirmed for all conditions (as can be seen in Table 3).

Figure 4 shows the distribution for r^*_{pb} when $k = 50$. This time, the distribution is centered at $r^*_{pb} \cong 0$ and there are approximately 5% of the observed values (2.46% to the left and 2.47% to the right) that fall beyond the critical values, $r^*_{pb(crit)} = \pm.062$. These results indicate that we can apply a t-test to r^*_{pb} and that the built-in association is absent from this correlation.

Figure 3 ■ Even though increasing the number of questions (k) does decrease the mean correlation for r_{pb}, it will never reach 0. In the case of r^*_{pb}, the correlation is independent of k and hence the mean correlation is always 0.



Conclusion

We have shown that applying a t-test to data with a built-in association, such as r_{pb}, is not appropriate. If one is interested in the correlation between correctly answering a question and the number of other questions correctly answered in an examination, replacing r_{pb} with r^*_{pb} is a safer alternative. If the correlation of interest is the one between the success on the question and the total number of correctly answered questions, it would be wiser to use the confidence interval method presented in this article.

Open practices

The *Open Material* badge was earned because supplementary material(s) are available on the journal's web site.

Citation

LeBlanc, V. & Cox, M. A. A. (2017). Interpretation of the point-biserial correlation coefficient in the context of a school examination. *The Quantitative Methods for Psychology*, *13*(1), 46–56. doi:10.20982/tqmp.13.1.p046

Received: 02/05/2014 ∼ Accepted: 29/09/2016

Figure 4 follows on next page.

Figure 4 ■ The r^*_{pb} ¬ follows a Normal distribution with $\mu = 0$. The red dotted lines indicate the critical t-values for a two-sided t-test with $\alpha = .05$, which are definitely appropriate for this data ($.0246 + .0247 \cong .05$).