Health Professions Education

Volume 4 | Issue 4 Article 10

2018-10-18

A Pragmatic Approach to Statistical Testing and Estimation (PASTE)

Jimmie Leppink
School of Health Professions Education, Maastricht University, PO Box 616, 6200 MD Maastricht, The Netherlands, jimmie.leppink@maastrichtuniversity.nl

Recommended Citation
Leppink, Jimmie (2018) "A Pragmatic Approach to Statistical Testing and Estimation (PASTE)," *Health Professions Education*: Vol. 4: Iss. 4, Article 10.
DOI: 10.1016/j.hpe.2017.12.009
Available at: https://hpe.researchcommons.org/journal/vol4/iss4/10

Available online at www.sciencedirect.com

ScienceDirect

Health Professions Education 4 (2018) 329–339

www.elsevier.com/locate/hpe

A Pragmatic Approach to Statistical Testing and Estimation (PASTE)

Jimmie Leppink

School of Health Professions Education, Maastricht University, PO Box 616, 6200 MD Maastricht, The Netherlands

Received 12 December 2017; accepted 29 December 2017
Available online 11 January 2018

Abstract

The *p*-value has dominated research in education and related fields and a statistically non-significant *p*-value is quite commonly interpreted as ‘confirming’ the null hypothesis (H_0) of ‘equivalence’. This is unfortunate, because *p*-values are not fit for that purpose. This paper discusses three alternatives to the traditional *p*-value that unfortunately have remained underused but can provide evidence in favor of ‘equivalence’ relative to ‘non-equivalence’: two one-sided tests (TOST) equivalence testing, Bayesian hypothesis testing, and information criteria. TOST equivalence testing and *p*-values both rely on concepts of statistical significance testing and can both be done with confidence intervals, but treat H_0 and the alternative hypothesis (H_1) differently. Bayesian hypothesis testing and the Bayesian credible interval aka posterior interval provide Bayesian alternatives to traditional *p*-values, TOST equivalence testing, and confidence intervals. However, under conditions outlined in this paper, confidence intervals and posterior intervals may yield very similar interval estimates. Moreover, Bayesian hypothesis testing and information criteria provide fairly easy to use alternatives to statistical significance testing when multiple competing models can be compared. Based on these considerations, this paper outlines a pragmatic approach to statistical testing and estimation (PASTE) for research in education and related fields. In a nutshell, PASTE states that all of the alternatives to *p*-values discussed in this paper are better than *p*-values, that confidence intervals and posterior intervals may both provide useful interval estimates, and that Bayesian hypothesis testing and information criteria should be used when the comparison of multiple models is concerned.

Keywords: Equivalence testing; Information criteria; Bayesian hypothesis testing; Confidence intervals; Posterior intervals

1. Introduction

The *p*-value has dominated research in education and related fields and a statistically non-significant *p*-value is quite commonly interpreted as ‘confirming’ the null hypothesis (H_0) of ‘equivalence’. Given the critiques on the latter practice and on statistical significance testing in a broader sense,[1–14] the persisting dominance of *p*-values and misuse such as in confirming a H_0 is quite remarkable. Where does the *p*-value actually come from and what does it actually mean?

1.1. Three distributions

In statistical significance testing, three distributions need to be considered. Firstly, there is the *population distribution* of the variable of interest, for instance mental effort needed to learn a particular topic, with an average or mean effort and standard deviation around

E-mail address: jimmie.leppink@maastrichtuniversity.nl

Peer review under responsibility of AMEEMR: the Association for Medical Education in the Eastern Mediterranean Region

https://doi.org/10.1016/j.hpe.2017.12.009

that mean. These population values are also called *parameters* and are – at least at a given time – assumed to have *fixed* values which are usually unknown. Secondly, once we draw a sample from that population, we obtain a *distribution of the sample* with a mean and standard deviation that are probably somewhat different from the population mean and standard deviation. The latter are also referred to as *statistics*. Thirdly, under the assumption that samples are drawn *randomly*, the *sampling distribution* is the probability distribution of a statistic of interest – in practice often a sample mean, a mean difference or a correlation – across all possible samples of the same size drawn from the same population. For statistical tests on means, mean differences and similar statistics this distribution is commonly assumed to be Normal (i.e., bell-shaped) or at least approximately Normal. Whenever the variables of interest follow a Normal distribution in the population, the sampling distribution also follows a Normal distribution. However, when these variables do not follow a Normal distribution in the population, the sampling distribution will gradually approach a Normal distribution with increasing sample size. Although samples of sizes like $N = 40$ will in practice often be sufficient to guarantee an approximately Normal sampling distribution, the assumption of an approximately Normal sampling distribution may be problematic when samples are of much smaller sizes such as $N = 10$. Unfortunately, samples of this size are not uncommon in some areas of educational and psychological research.

Even if an approximately Normal sampling distribution can be assumed, smaller samples are problematic for other reasons including the following. Firstly, given that many differences of interest in fields like education and psychology are of a somewhat smaller magnitude or of ‘medium’ size (e.g., half a standard deviation or Cohen's $d = 0.5$)[15] at best, researchers who perform statistical significance tests on the difference in means between two samples of sizes of 10–20 are generally poorly equipped to detect such differences of interest.[16] Secondly, although that limited statistical power to detect differences of interest not rarely stimulates an expectation that a statistically significant p-value must reflect a real effect, it is easily demonstrated that in a pile of statistically significant outcomes more false alarms can be expected in smaller than in larger samples.[17] Thirdly, the smaller the sample sizes, the more findings can be expected to vary from one random sample to the next.[18] Many researchers appear to underappreciate that variability and speak of ‘failed’ replications when a statistically significant outcome from an initial study cannot be replicated in a direct replication of that study. Replicating a *study* does not imply that the *findings* of the initial study will be replicated, and when dealing with small samples findings may vary wildly from one study to the next even if in the population we sample from the phenomenon of interest (e.g., mean, mean difference or correlation) remains the same over time.

Under the aforementioned assumptions of *fixed* population parameters, *random* samples and usually at least approximately Normal sampling distributions, the *standard error* is the standard deviation of the sampling distribution that can be used to calculate p-values and confidence intervals (CIs). The p-value is the probability of finding the difference, relation or effect observed in the sample or further away from H_0, assumed that H_0 is true. If the p-value is very small, it is reasoned that a result (i.e., difference, relation or effect) at least as extreme as observed in the sample is very unlikely under H_0 but probably more likely under an alternative hypothesis H_1 that states that there is a difference, relation or effect in the population of interest. Both p-values and CIs can indicate whether at statistical significance level α we can reject H_0. For example, a 95% CI not including the population value or difference specified under H_0 corresponds with p being smaller than $\alpha = 0.05$ in a default two-sided test (i.e., extreme deviations from H_0 in both negative and positive direction can result in H_0 being rejected). Moreover, CIs are widely used to provide interval estimations around means and other statistics. However, it is important to note that in this approach to statistical testing and estimation, all inference is about what would happen across all possible samples of the same size drawn from the same population. Given *fixed* population parameters and *random* samples, a 95% CI will include the population parameter of interest in 95% of all possible samples of the same size drawn from that same population.

1.2. Equivalence testing

Given the very definition of the p-value, it cannot provide evidence in favor of a H_0.[11,16] Hence, if H_0 is defined in terms of ‘equivalence’, a statistically non-significant p-value cannot be used to confirm equivalence. However, in an alternative to this traditional H_0 significance testing approach, two one-sided tests (TOST) equivalence testing,[19–21] this problem is circumvented by treating H_0 and H_1 differently. For instance, we may agree that differences between a treatment and control condition smaller than 0.3 standard deviations (i.e., $d < 0.3$) are too small to really matter for educational or psychological practice in a given context. In the traditional approach, the p-value is an outcome with regard to H_0: $d = 0$, and a statistically significant p-value results in that H_0 being rejected thinking that the difference observed or more extreme is more likely under H_1: $d≠0$. In TOST equivalence testing, two H_0s are tested: $H_{0.1}$: $d < -0.3$ (i.e., more negative than $d = -0.3$) and $H_{0.2}$: $d > 0.3$ (i.e., more positive than $d = 0.3$). In other words, contrary to the traditional approach, the H_0s in TOST equivalence testing are about *differences* in a certain range, and *if and only if both* $H_{0.1}$ *and* $H_{0.2}$ can be rejected we have evidence for (relative) equivalence. That is, the rejection of both these H_0s does not imply that H_0: $d = 0$ is true, but it does provide evidence for the hypothesis that differences are in a 'zero-to-small' range representing little if any practical meaning. In terms of CIs, if both $H_{0.1}$ (lower bound: $d < -0.3$) and $H_{0.2}$ (upper bound: $d > 0.3$) are rejected at $\alpha = 0.05$, a 90% CI around the statistic of interest (here: d observed in the sample) will *not* include any of the values under $H_{0.1}$ and $H_{0.2}$; whenever it includes a value (under $H_{0.1}$ and/or $H_{0.2}$) the statistical significance test associated with that value (i.e., $H_{0.1}$ and/or $H_{0.2}$) will not be statistically significant and hence we have no evidence for (relative) equivalence.

In other words, while we can still use the CIs that we should report when using traditional p-values, by testing two H_0s which represent meaningful differences, TOST equivalence testing enables us to avoid a fallacy associated with traditional p-values: to interpret particular testing outcomes as evidence in favor of (relative) equivalence. However, TOST equivalence testing is not the only alternative to the traditional p-value.

1.3. From before to after seeing the data

In another alternative, the Bayesian approach, uncertainty with regard to population parameters of interest is expressed in a *probability* distribution that is updated when data comes in. For example, before seeing any data, we may not know anything with regard to a correlation of interest, and that will translate in a probability distribution *before* seeing the data (i.e., *prior distribution*) that indicates that any correlation may be equally likely (i.e., a Uniform prior distribution).[13] Observing a substantial correlation (e.g., $r = 0.4$) in a sample of $N = 30$ participants will result in a peaked probability distribution *after* seeing the data (i.e., *posterior distribution*) with a 95% posterior interval (PI) aka *credible* interval around $r = 0.4$, and the value '0' (i.e., the traditional H_0 in statistical significance testing) will be less likely than before seeing the data. However, observing a smaller correlation (e.g., $r = 0.1$) in that sample would result in a peaked posterior distribution with a 95% PI around $r = 0.1$, and $r = 0$ would be more likely than it was before seeing the data. Finally, keeping other factors constant, in small samples (e.g., $N = 10$), the posterior distribution will be less peaked than in a large sample (e.g., $N = 100$); a small sample has less weight (i.e., provides less information) than a large sample.

The shift in likelihood of a H_0 vs. a H_1 from before to after seeing the data can be expressed in a so-called *Bayes factor* (*BF*).[13,16] When formulated as H_0 vs. H_1 (BF_{01}), values > 1 indicate that after seeing the data H_0 has become more likely than before seeing the data, whereas when formulated as H_1 vs. H_0 (BF_{10}), values > 1 indicate that after seeing the data H_0 has become less likely than before seeing the data (i.e., $BF_{10} = 1/BF_{01}$). The more a BF is away from 1, the more evidence for one hypothesis vs. the other, and a value of 1 indicates that seeing the data has not resulted in a shift in likelihood of one hypothesis vs. the other. In the aforementioned case of observing $r = 0.4$, BF_{10} will be greater than 1, whereas in the case of observing $r = 0.1$, BF_{01} will be greater than 1.

Although in former times a common critique to Bayesian hypothesis testing and PIs was that specifying a prior distribution (i.e., before seeing the data) may be highly subjective, the influence of the prior on the PI decreases as more data comes in (i.e., with increasing sample size). Moreover, for an increasing variety of statistical models so-called 'default Priors' (i.e., that take into account the mathematical properties of different types of variables and their probability distributions but effectively assume little or no prior knowledge about the population parameter of interest) have been proposed and used[22–25] and are available in free-of-charge (i.e., non-commercial) software packages.[26–28] With regard to the interpretation in terms of evidence, BFs can then be interpreted as follows: 1–3.2: anecdotal evidence, not worth more than a bare mention; 3.2–10: substantial evidence; 10–32: strong evidence; 32–100: very strong evidence; 100+: 'decisive'.[29] Hence, $BF_{10} = 5.29$ indicates substantial evidence in favor of H_1 vs. H_0, whereas $BF_{10} = 0.05$ corresponds with $BF_{01} = 1/0.05 = 20$ and indicates strong evidence in favor H_0 vs. H_1.

1.4. Several competing models

In many practical situations, not two but several hypotheses or combinations of hypotheses may compete. For instance, consider a randomized controlled experiment in which students are randomly allocated to four cells of a 2 (factor A) × 2 (factor B) factorial design. They study during a certain period of time following the format outlined in their respective conditions and subsequently complete a posttest for the researchers to compare the conditions in terms of posttest performance. Given a 2 × 2 factorial design, we can distinguish three effects: a main effect of factor A, a main effect of factor B, and an A-by-B interaction effect (i.e., a combined effect of A and B). This results in five possible models: (1) a model with main effect A only, (2) a model with main effect B only, (3) a model with *both* main effects; (4) a model with both main effects *and* the interaction effect; and (5) a model without any of the aforementioned terms (i.e., none of the effects mattering). The same holds for a study where the interest lies in group differences (i.e., the one and only factor under study) in terms of a response variable of interest but there is a third variable, a covariate: if in the previous we replace 'main effect B' by 'effect of covariate', we have our five models that can be compared. Bayesian hypothesis testing provides a very straightforward and transparent approach to comparing these models. A priori, all models may be equally likely. However, after seeing the data, some models will be more likely than others. Bayes factors can be calculated and compared across models to facilitate decision-making which model we may prefer now that we have seen the data.

A somewhat similar approach to comparing multiple models is found in so-called *information criteria.*[30–36] Although a full discussion on the wide variety of information criteria available, which ones should be used for which kind of analyses and how they relate to the other approaches already discussed would require at least one full article in itself, succinctly put the logic underlying information criteria is to help researchers determine which of the models under comparison best captures the information in the data. The different criteria differ in their trade-off between model fit on the one hand and model complexity and sample size on the other hand. Three criteria that are widely available and will be used in this paper are Akaike's information criterion (AIC),[32] Schwarz' Bayesian information criterion (BIC)[35] and the sample-size adjusted BIC.[37–39] Although a full analysis of these criteria and their possible uses is beyond the scope of this paper, in short smaller values indicate better models.

1.5. The current study: a comparison of traditional p-values and alternatives

In a nutshell, from Sections 1.1–1.4. it becomes clear that several alternatives to traditional p-values exist and enable researchers to do things that cannot be done with traditional p-values. TOST equivalence testing, Bayesian hypothesis testing, and information criteria provide three alternatives that in their own way enable researchers to obtain evidence in favor of a 'no (meaningful) difference' H_0 relative to a '(meaningful) difference' H_1. In the case of a simple two-group comparison (e.g., a two-samples t-test or one-way analysis of variance),[40] TOST equivalence testing and Bayesian hypothesis testing provide easy alternatives to traditional p-values to obtain evidence in favor of (relative) equivalence. In the case of more complex models, Bayesian hypothesis testing and information criteria – and to some also TOST equivalence testing – provide researchers with tools to determine whether or not to include a particular term in a model without having to rely on traditional p-values. For example, the assumption of non-interaction underlying analysis of covariance (ANCOVA)[41] has commonly been tested using traditional p-values. A statistically non-significant p-value then results in dropping the interaction term from the model, as if it does not matter. Instead, TOST could be applied by inspecting the 90% CI for an effect size such as standardized β and check for overlap with the lower or upper bound (e.g., $\beta < -0.1$ and $\beta > 0.1$ and or perhaps $\beta < -0.15$ and $\beta > 0.15$). In the case of no overlap, we would drop the interaction term from the model. Also, Bayes factors or information criteria can be calculated to see if a model with interaction term or one of the models without interaction term should be preferred.

The remainder of this paper focuses on a simulated practical example in the form of a study in which two groups of learners are asked about the previous experience with a topic to be studied, then each study that topic in a different way, and finally complete a posttest on the topic studied. The main interest lies in group differences in average posttest performance followed by group differences in average previous experience followed by the relation between posttest performance and previous experience and finally the combined use of the grouping variable and previous experience as predictors of posttest performance. Traditional p-values and all alternatives discussed in this paper are compared. Based on the analysis thus far and the comparison in the following sections, this paper outlines a *pragmatic approach to statistical testing and estimation* (PASTE) for research in education and related fields. Concisely put, PASTE states that all of the alternatives to p-values discussed in this paper are better than p-values, that CIs and PIs may both provide useful interval estimates, and that Bayesian hypothesis testing and information criteria should be used when the comparison of multiple models is concerned. PASTE is a *pragmatic* approach because it does not reject one approach or another for reasons of philosophy, worldview or definition of probability, but considers different approaches as having the potential to help researchers study questions of interest from somewhat different perspectives. PASTE is an *approach to testing and estimation*, because *both* testing *and* estimation are considered useful in helping researchers to make decisions with regard to questions of interest.

2. Method

2.1. Participants and design

In this study, two groups of $n = 100$ learners each (i.e., total $N = 200$) are asked about the previous experience with a topic to be studied, then each study that topic in a different way, and finally complete a posttest on the topic studied. The reader may wonder why such large numbers if they are not common for much of educational and psychological research. With two groups of $n = 100$ each and testing at $\alpha = 0.05$, our traditional two-samples t-test has a statistical power of 0.8 or 80% for $d = 0.4$.[42] To achieve that statistical power when testing at $\alpha = 0.05$ and assuming $d = 0.5$, we would need $n = 64$ per group. Using two groups of $n = 100$ each, testing at $\alpha = 0.05$ and assuming $d = 0.3$, we have a statistical power of around 0.56.

2.2. Materials and procedure

Both groups of learners are asked to self-report their previous experience with the topic that will be studied on a visual analog scale (VAS) ranging from 0 (minimum) to 100 (maximum). Next, they study the topic in their own way – as they are used to do (i.e., no random allocation to condition) – and complete a posttest on the topic studied that yields a score ranging from 0 (minimum) to 100 (maximum).

2.3. Statistical analysis

As the main research question is whether the two groups differ in terms of posttest performance, a two-samples t-test[40] along with a TOST equivalence testing[21] and Bayesian[26] alternative will be performed for posttest performance. As the second research question is whether the two groups differ in terms of previous experience (i.e., this is not a randomized controlled experiment), the same will be done for previous experience. Next, as the third research question is on the relation between posttest performance and previous experience, a traditional test on the Pearson's correlation between posttest performance and previous experience[40] along with a TOST equivalence testing[21] and Bayesian[26] alternative will be performed. Finally, to examine the combined use of the grouping variable and previous experience as predictors of posttest performance, we compare five models using Bayes factors, AIC, BIC, and (sample-size) adjusted BIC similar to the previously discussed two-way factorial example: (1) only the groups; (2) only previous experience; (3) groups and previous experience (i.e., ANCOVA);[41] (4), groups, previous experience, and their interaction; and (5) a model without any of the aforementioned terms (i.e., as if none matter).

For all analyses and models, descriptive statistics as well as traditional significance testing and Bayesian hypothesis testing were done with JASP 0.8.4,[26] information criteria were calculated in M*plus* 8,[38] TOST equivalence testing was performed with JAMOVI 0.8.1.5,[28] and the graphs (i.e., Figs. 1 and 2) come from SPSS 25.[43]

3. Results

Fig. 1 graphically depicts the univariate distributions (i.e., histograms and boxplots) of posttest score and self-reported previous experience for each of the two groups.

Fig. 2 demonstrates the bivariate distribution (i.e., scatterplot) of posttest score and self-reported previous experience with separate icons for members of the two groups (i.e., squares for learners from group A, triangles for learners from group B).

Table 1 presents means and standard deviations along with skewness and kurtosis for posttest score (0-100) and previous experience (0-100) for each group.

3.1. Research question 1: group difference in posttest performance

Table 2 presents the outcomes of traditional significance testing, TOST equivalence testing, and Bayesian hypothesis testing for posttest performance and previous experience.

In terms of standard deviations, the difference between groups A and B in average posttest performance is small: $d = 0.142$. The 95% CI extends from −0.135 to 0.420 and thus includes H_0: $d = 0$, hence the result is not statistically significant at $\alpha = 0.05$: $p = 0.315$. However, the 90% CI (from −0.091 to 0.375)

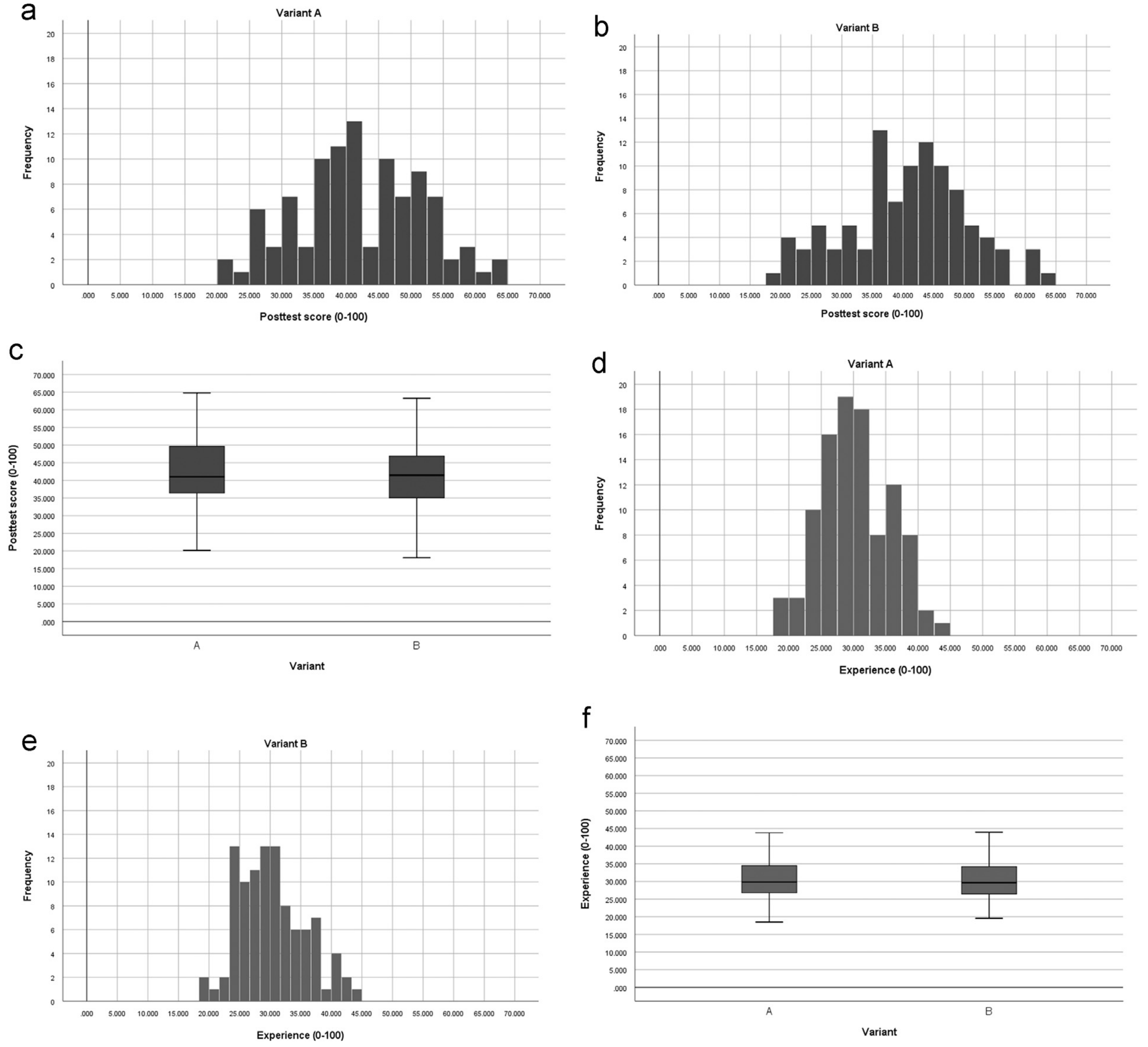

Fig. 1. Histograms and boxplots of posttest score (0–100) and experience (0–100) per group. *Software: SPSS 25.*

and logically p-values obtained from TOST equivalence testing indicate that we have insufficient evidence to assume (relative) equivalence either: although we can reject $H_{0.1}$: $d < -0.3$ ($p = 0.001$), we fail to reject $H_{0.2}$: $d > 0.3$ ($p = 0.133$).

Bayesian hypothesis testing yields $BF_{10} = 0.247$, which corresponds with $BF_{01} = 1/0.247 \approx 4.049$. In other words, from Bayesian hypothesis testing, we conclude 'substantial evidence' in favor of H_0 vs. H_1. The 95% PI extends from −0.132 to 0.398 and is thus quite similar to the 95% CI. This is quite normal; even when Cauchy not Uniform priors are used, 95% PIs and 95% CIs for mean differences tend to become more similar with increasing sample size.

3.2. Research question 2: group difference in previous experience

In terms of standard deviations, the difference between groups A and B in average previous experience is close to zero: $d = -0.017$. The 95% CI extends from −0.294 to 0.260 and thus includes H_0: $d = 0$, hence the result is not statistically significant at $\alpha = 0.05$: $p = 0.904$. In this case, the 90% CI (from −0.250 to 0.216) includes neither the lower nor upper bound value used in TOST equivalence testing, and hence both $H_{0.1}$: $d < -0.3$ ($p = 0.023$) and $H_{0.2}$: $d > 0.3$ ($p = 0.013$) can be rejected. In other words, for previous experience, we appear to have sufficient

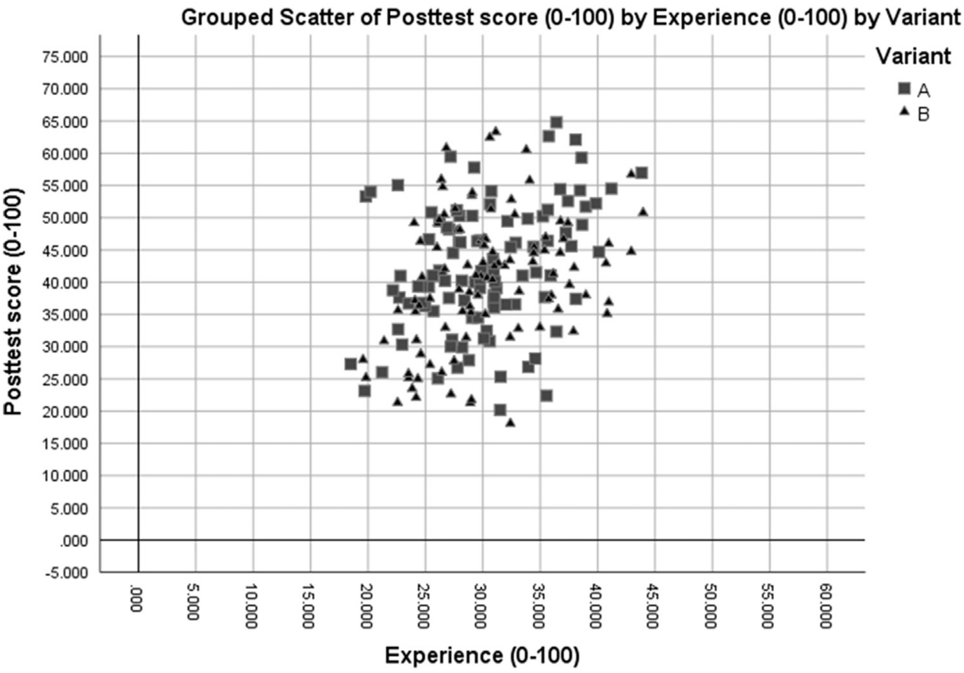

Fig. 2. Scatter plot of posttest score (0-100) by experience (0-100) per group. *Software: SPSS 25.*

Table 1
Descriptive statistics of posttest score and experience per group. *Software: JASP 0.8.4.*

	Posttest score (0-100)		Experience (0-100)	
	Variant A	Variant B	Variant A	Variant B
Sample size	100	100	100	100
Mean	42.013	40.597	30.259	30.351
Standard deviation	9.842	10.056	5.362	5.417
Skewness	0.019	−0.111	0.127	0.482
Kurtosis	−0.500	−0.351	−0.439	−0.270

evidence to assume (relative) equivalence (i.e., $-0.3 < d < 0.3$).

Bayesian hypothesis testing yields $BF_{10} = 0.155$, which corresponds with $BF_{01} = 1/0.155 \approx 6.452$. In other words, from Bayesian hypothesis testing, we conclude 'substantial evidence' in favor of H_0 vs. H_1. The 95% PI extends from −0.279 to 0.251 and is thus quite similar to the 95% CI.

3.3. Research question 3: the relation between previous experience and posttest performance

Table 3 presents testing and estimation outcomes with regard to the correlation between posttest score and experience.

The observed correlation is $r = 0.332$, and the 95% CI ranges from 0.202 to 0.450. Hence, a traditional significance test yields $p < 0.001$. The 90% CI ranges from 0.224 to 0.432, and therefore, in TOST equivalence testing, $H_{0.1}$: $r < -0.3$ can rejected ($p < 0.001$) but $H_{0.2}$: $r > 0.3$ cannot be rejected ($p = 0.690$). Bayesian hypothesis testing yields $BF_{10} = 8153.056$ or 'decisive' evidence in favor of H_1 vs. H_0, and the 95% PI (ranging from 0.201 to 0.447) almost coincides with the 95% CI.

3.4. Comparing models for the prediction of posttest performance

Table 4 presents a comparison of the five models mentioned previously: only the grouping variable (factor: 'X'), only previous experience (covariate: 'C'), both grouping variable and covariate ('X, C'), both grouping variable and covariate as well as their interaction ('X, C, XC'), and a model without any of the aforementioned ('None').

Both in terms of *BF*s (i.e., larger is better) and all three information criteria (i.e., smaller is better), the model with only previous experience (i.e., 'C') is to be preferred. Note that we have decided on whether or not to include terms other than previous experience (i.e., grouping variable and perhaps additionally group-by-experience interaction) without having to engage in statistical significance testing.

Table 2
Tests for group differences in posttest score and experience. Software: JASP 0.8.4 (classical, Bayesian), JAMOVI 0.8.1.5 (equivalence).

		Posttest	Experience
Cohen's d	Estimate[a]	0.142	−0.017
Confidence	95% Lower	−0.135	−0.294
	95% Upper	0.420	0.260
	90% Lower	−0.091	−0.250
	90% Upper	0.375	0.216
Classical	H_0: $d = 0$	0.315	0.904
Equivalence	$H_{0.1}$: $d < -0.3$	0.001	0.023
	$H_{0.2}$: $d > 0.3$	0.133	0.013
Bayesian[b]	H_1: $d \neq 0$ *vs.* H_0: $d = 0$	0.247	0.155
Posterior[c]	95% Lower	−0.132	−0.279
	95% Upper	0.398	0.251

[a] Positive difference is in favor of variant A.
[b] Bayes factor for H_1 vs. H_0.
[c] Prior: default (Cauchy with location 0 and scale 0.707).

Table 3
On the correlation between posttest score and experience. Software: JASP 0.8.4 (classical, Bayesian), JAMOVI 0.8.1.5 (equivalence).

Pearson's r	Estimate	0.332
Confidence	95% Lower	0.202
	95% Upper	0.450
	90% Lower	0.224
	90% Upper	0.432
Classical	H_0: $r = 0$	< 0.001
Equivalence	$H_{0.1}$: $r < -0.3$	< 0.001
	$H_{0.2}$: $r > 0.3$	0.690
Bayesian[a]	H_1: $r \neq 0$ *vs.* H_0: $r = 0$	8153.056
Posterior[b]	95% Lower	0.201
	95% Upper	0.447

[a] Bayes factor for H_1 vs. H_0.
[b] Prior: default (Uniform).

Table 4
Comparison of five models. Software: JASP 0.8.4 (R^2, Bayesian), Mplus 8 (information criteria).

In model	X	C	X, C	X, C, XC	None
R^2	0.005	0.110	0.116	0.116	0.000
adjusted R^2	< 0.001	0.106	0.107	0.102	0.000
Bayes factor[a]	0.247	8983.093	2832.186	629.420	1.000
AIC[b]	1490.587	1468.285	1469.044	1471.024	1489.608
BIC[b]	1500.482	1478.180	1482.237	1487.515	1496.205
Adjusted BIC[c]	1490.978	1468.676	1469.565	1471.675	1489.869

[a] Larger = more in favor (a priori: all models equally likely).
[b] Smaller = better.
[c] Sample-size adjusted; smaller = better.

4. Discussion

This paper started with the observations that p-values have dominated research in education and related fields, that statistically non-significant p-values are not rarely used to 'confirm' a H_0 that states some kind of equivalence, and that the use of p-values in small samples comes with additional problems. The alternatives to the traditional p-value discussed in this paper are not immune to issues related to small samples (e.g., $N = 10$) and all methods suffer from the same common problem: *the user*. Although a common quote attributed to Benjamin Disraeli is "*Lies, damned lies, and statistics*", a more appropriate way of summarizing statistics and the way they are used is probably "*Lies, damned lies, but not statistics*." As long as methods are used in a context where they make sense and for the purposes they are designed for, there is no problem at all. As long as researchers interpret the p-value as it is and refrain from using it for purposes it was not designed for, there is no problem. The problem occurs when we start to use methods inappropriately or in essence correctly but for the wrong questions.

Like any method, statistical methods only make sense in a particular context. Numbers, in isolation, say nothing, and outcomes of statistical testing, without presentation of descriptive statistics and/or graphical presentation thereof, tell nothing. Methods only make sense when they are used appropriately and for the purposes and kind of questions they have been designed for. Therefore, PASTE recommends three basic principles:

1. *Describe and plot*: with a clear descriptive and graphical presentation of key statistics we generally provide a much more meaningful picture to our audience than by just reporting outcomes of statistical testing.
2. *Equivalence testing and interval estimation*: when we engage in statistical significance testing, we should not rely solely on traditional p-values but report CIs and the outcomes of TOST equivalence testing as well.
3. *Model comparison*: when our interest lies in comparing two or more statistical models – which represent two or more (sets of) hypotheses – information criteria and Bayesian hypothesis testing provide useful alternatives to statistical significance testing.

Each of these three principles is discussed in the following.

4.1. PASTE (1): describe and plot

A descriptive and/or graphical check usually provides us with a much more meaningful picture than outcomes of statistical testing in absence of that descriptive and/or graphical check. Researchers commonly check assumptions like Normality and equality of variances by using a statistical significance test. Using a statistically non-significant p-value as evidence in favor of our assumption of Normality or equal variance – or any assumption for that matter – we are back to our fallacy of using a p-value to 'confirm' the null. Moreover, while such tests may fail to reach statistical significance despite a substantial deviation from our assumption of Normality or equal variance when applied to small samples, they may yield statistically significant outcomes for practically possibly meaningless deviations when used in large samples. Using graphs (e.g., Figs. 1 and 2) and descriptive statistics (e.g., Table 1) provides a much more practical approach to assessing assumptions than some statistical testing outcomes. The same holds for group differences on aspects of interest, such as average performance and previous experience. If we report only descriptive statistics (e.g., Table 1) a reader can compute the outcomes of a statistical test using these descriptive statistics; the other way around (e.g., trying to calculate the descriptive statistics and hence actual difference from a statement like '$p < 0.001$') does not really work.

When we deal with two equally large samples like the ones compared in this paper that have more or less the same standard deviations on the variables of interest (here: posttest performance and previous experience), assuming equal standard deviations seems fair and accounting for unequal standard deviations would result in almost identical testing and estimation outcomes. In studies where the groups differ substantially in their standard deviations (e.g., one group has a standard deviation about twice as large as the other group), accounting for unequal variances may yield somewhat different results compared to assuming equal variances and accounting for unequal variances is probably to be preferred. When in doubt, reporting both (i.e., with and without accounting for unequal standard deviations) – along with the actual means and standard deviations – is always an option.

When samples are very small, it becomes very hard to engage in a meaningful check of assumptions. For instance, even if two populations of interest have the same standard deviation (or in squared form, the same variance), drawing a small random sample from each of these populations may yield sample standard deviations that differ substantially from one another. The other way around is of course also possible: population standard deviations may differ substantially, yet a study in which a small random sample is drawn from each population results in approximately equal standard deviations. Keeping other factors constant, the deviation between sample statistic and population parameter tends to decrease with increasing sample size, and the same holds for the difference in a statistic from sample to sample.

Whether we need to engage in statistical testing when dealing with very large samples partly depends on the complexity of the statistical models to be compared; more complex models generally require larger samples. When interested in a fairly simple comparison in terms of average performance, such as in the example discussed in this paper, testing at $\alpha = 0.05$ yields a statistical power of 0.8 for $d = 0.3$ with two samples of $n = 176$ each (i.e., total $N = 352$), a statistical power of 0.8 for $d = 0.2$ with two samples of $n = 394$ each (i.e., total $N = 788$) and a statistical power of 0.8 for $d = 0.1$ with two samples of $n = 1571$ each (i.e., total $N = 3142$). Eventually, even very small differences in average performance observed in the sample may result in H_0: $d = 0$ being rejected, but descriptive statistics will indicate whether we are dealing with a practically meaningful difference.

4.2. PASTE (2): equivalence testing and interval estimation

When we engage in statistical significance testing, we should not rely only on traditional p-values that reject some H_0 of equivalence but report CIs and the outcomes of TOST equivalence testing as well. Regardless of the p-value, CIs provide us with intervals around our statistics of interest. Keeping other factors constant, the smaller the samples the wider the CIs and the more findings can be expected to vary from sample to sample. If a p-value obtained from a test of a H_0 of equivalence is statistically significant, CIs and the outcomes of TOST equivalence testing can provide useful information with regard to the practical significance of a difference, relation or effect of interest. In very large samples, or in a meta-analysis on a large series of studies,[15] the p-value associated with H_0: $d = 0$ may be statistically significant at $\alpha = 0.05$, but the 90% CI for d may range from 0.05 to 0.25 meaning that it overlaps with neither the lower bound ($H_{0.1}$: $d < -0.3, p > 0.05$) nor the upper bound ($H_{0.2}$: $d > 0.3, p > 0.05$) of TOST equivalence testing and hence the practical implications of the difference under study may be limited. At the same time, the current paper provides an example of how even in a study that compares two samples of $n = 100$ each a traditional significance test may yield a statistically non-significant p-value (i.e., for posttest performance: $p = 0.315$) but TOST equivalence testing fails to provide evidence in favor of equivalence as well (i.e., the upper bound of the 90% CI of 0.375 exceeds 0.3, and hence the p-value associated with $H_{0.2}$: $d > 0.3$ is larger than 0.05, here 0.133).

4.3. PASTE (3): model comparison

TOST equivalence testing, through p-values associated with $H_{0.1}$ and $H_{0.2}$ and/or through CIs, provides researchers with a means to reflect on questions with regard to whether we should consider two treatments or two groups as meaningfully different or more or less equivalent. Bayesian hypothesis testing also provides a useful approach for that kind of questions, and 95% PIs provide a useful alternative to 95% CIs, although with the kind of prior distributions used in this paper and/or sample sizes like the one in the example study or larger 95% PIs and 95% CIs for a mean difference or for a correlation may well be very similar.

Whenever multiple competing models can be compared, such as in the case of two-way factorial designs or the example study in this paper with one grouping variable and a covariate, Bayesian hypothesis testing and information criteria can help researchers to reflect on which model we may prefer without having to engage in statistical significance testing. This way, we may decide to drop an interaction term from the model (e.g., to perform ANCOVA, which assumes no interaction between grouping variable and covariate) without having to engage in the erroneous reasoning of using a statistically non-significant p-value as evidence in favor of H_0 of no interaction.

4.4. To conclude: PASTE as an alternative to traditional p-values

In a nutshell, PASTE states that all of the alternatives to p-values discussed in this paper are better than p-values, that CIs and PIs may both provide useful interval estimates, and that Bayesian hypothesis testing and information criteria should be used when the comparison of multiple models is concerned. Given that TOST equivalence testing and CIs, Bayesian hypothesis testing and PIs as well as information criteria such as the ones discussed in this paper are readily available in more than one statistical software package, the times that we need to rely on p-values associated with H_0s of no difference have passed. And given the advantages that these alternatives have to traditional p-values, we may as well stop using traditional p-values altogether.

Ethical approval

Not applicable.

Funding

None.

Other disclosures

No conflicts of interests.

References

1. Bakker M, Van Dijk A, Wicherts JM. The rules of the game called psychological science. *Perspect Psychol Sci* 2012;7:543–554.
2. Cohen J. Things I have learned (thus far). *Am Psychol* 1990;45:1304–1312.
3. Cohen J. The earth is round (p < .05). *Am Psychol* 1994;49:997–1003.
4. Cumming G. Replication and p intervals: p values predict the future only vaguely, but confidence intervals do much better. *Perspect Psychol Sci* 2008;3:286–300.
5. Fidler F, Cumming G. The new stats: attitudes for the 21st century. In: Osborne JW, editor. *Best Practices in Quantitative Methods*. London: Sage; 2010.
6. Gigerenzer G. Mindless statistics. *J Socio-Econom* 2004;33:587–606.
7. Kline RB. *Beyond Significance Testing: Reforming Data Analysis Methods in Behavioral Research*. Washington, DC: APA; 2004.
8. Kruschke JK. *Doing Bayesian data analysis: a tutorial with R and BUGS*. London: Elsevier; 2011.
9. Meehl PE. Theoretical risks and tabular asterisks: sir Karl, Sir Ronald, and the slow progress of soft psychology. *J Consult Clin Psychol* 1978;46:806–834.
10. Nickerson RS. Null hypothesis significance testing: a review of an old and continuing controversy. *Psychol Methods* 2000;5:241–301.
11. Rozeboom WW. The fallacy of the null hypothesis significance test. *Psychol Bull* 1960;57:416–428.
12. Wagenmakers EJ. A practical solution to the pervasive problems of p values. *Psychon Bull Rev* 2007;14:779–804.
13. Wagenmakers EJ, Marsman E, Jamil, T, et al. Bayesian inference for psychology. Part I: theoretical advantages and practical ramifications. *Psychon Bull Rev* 2017http://dx.doi.org/10.3758/s13423-017-1343-3.
14. Wetzels R, Matzke D, Lee MD, Rouder JN, Iverson GJ, Wagenmakers EJ. Why psychologists must change the way they analyze their data: the case of psi: comment on Bem (2011). *J Pers Soc Psychol* 2011;100:426–432.

15. Lipsey MW, Wilson DB. *Practical Meta-analysis*. London: Sage; 2001.
16. Leppink J, O'Sullivan P, Winston K. Evidence against vs. in favour of a null hypothesis. *Perspect Med Educ* 2017;6:115–118.
17. Leppink J, Winston K, O'Sullivan P. Statistical significance does not imply a real effect. *Perspect Med Educ* 2016;5:122–124.
18. Leppink J, O'Sullivan P, Winston K. On variation and uncertainty. *Perspect Med Educ* 2016;5:231–234.
19. Goertzen JR, Cribbie RA. Detecting a lack of association: an equivalence testing approach. *Br J Math Stat Psychol* 2010;63: 527–537.
20. Hauck DWW, Anderson S. A new statistical procedure for testing equivalence in two-group comparative bioavailability trials. *J Pharm Biopharm* 1984:83–91.
21. Lakens D. Equivalence tests: a practical primer for t tests, correlations, and meta-analyses. *Soc Psychol Pers Sci* 2017;8: 355–362.
22. Rouder JN, Speckman PL, Sun D, Morey RD. Bayesian t tests for accepting and rejecting the null hypothesis. *Psychon Bull Rev* 2009;16:225–237.
23. Wagenmakers EJ, Lodewyckx T, Kuriyal H, Grasman R. Bayesian hypothesis testing for psychologists: a tutorial on the Savage-Dickey method. *Cogn Psychol* 2010:158–189.
24. Rouder JN, Morey RD, Speckman PL, Province JM. Default Bayes factors for ANOVA designs. *J Math Psychol* 2012;56: 356–374.
25. Van der Zee T, Admiraal W, Paas F, Saab N, Giesbers B. Effects of subtitles, complexity, and language proficiency on learning from online education videos. *J Med Psychol* 2017;29:18–30.
26. Love J, Selker R, Marsman M, et al. JASP; 2017. Retrieved from ⟨https://jasp-stats.org/⟩ [Accessed 12 December 2017].
27. Morey RD, Rouder JN BayesFactor: computation of Bayes factors for common designs [Computer software manual], http://dx.doi.org/10.1016/j.jmp.2012.08.001.
28. Jamovi project. Jamovi (version 0.8.1.5). Retrieved from ⟨https://www.jamovi.org⟩ [Accessed 12 December 2017].
29. Jeffreys H. *Theory of Probability*. Oxford: University Press; 1961.
30. Anderson DR. *Model based inference in the life sciences: a primer on evidence*. New York, NY: springer; 2008.
31. Burnham KP, Anderson DR. *Model Selection and Multimodel Inference: a Practical Information-theoretic Approach*. New York, NY: Springer; 2002.
32. Akaike H Information theory and an extension of the maximum likelihood principle. In: Petrov BN, Csaki F (Eds.), Proceedings of the Second International Symposium on Information Theory. Budapest: Academiai Kiado: 267–81.
33. Bozdogan H. Model selection and Akaike's information criterion (AIC): the general theory and its analytical extensions. *Psychometrika* 1987;52:345–370.
34. Hurvich CM, Tsai CL. Regression and time series model selection in small samples. *Biometrika* 1989;76:297–307.
35. Schwarz G. Estimating the dimensions of a model. *Ann Stat* 1978;6:461–464.
36. Spiegelhalter DJ, Best NG, Carlin BP, Van der Linde A. Bayesian measures of model complexity and fit (with discussion). *J R Stat Soc* 2002;64:583–639.
37. Enders CK, Tofighi D. The impact of misspecifying class-specific residual variance in growth mixture models. *Struct Eq Mod: Multidisc J* 2008;15:75–95.
38. Muthén LK, Muthén B Mplus user's guide. Version 8; 2017. Retrieved from: ⟨https://www.statmodel.com/download/usersguide/MplusUserGuideVer_8.pdf⟩ [Accessed 12 December 2017].
39. Tofighi D, Enders CK. Identifying the correct number of classes in mixture models. In: Hancock GR, Samuelsen KM, editors. *Advances in Latent Variable Mixture Models*. Greenwich, CT: Information Age; 2007. p. 317–341.
40. Field A. *Discovering Statistics Using IBM SPSS Statistics*, 4th ed., London: Sage; 2013.
41. Huitema B. *The Analysis of Covariance and Alternatives: Statistical Methods for Experiments, Quasi-experiments, and Single-case Studies*. Chichester: Wiley; 2011.
42. Buchner A, Erdfelder E, Faul F, Lang AG G*Power: statistical power analyses for Windows and Mac, G*Power version 3.1.2. [software]; 2009. Retrieved from ⟨http://www.gpower.hhu.de/⟩ [Accessed 12 December 2017].
43. IBM Corporation. SPSS 25; Retrieved from: ⟨http://www−01.ibm.com/support/docview.wss?Uid=swg24043678⟩ [Accessed 12 December 2017].