Do We Always Need a Difference? Testing Equivalence in a Blended Learning Setting

Claude Mueller[a], Thoralf Mildenberger[b] and Maren Lübcke[c]

[a] Center for Innovative Teaching and Learning, Zurich University of Applied Sciences, Winterthur, Switzerland
[b] Zurich University of Applied Sciences, Winterthur, Switzerland
[c] HIS Hochschul-Informations-System, Hannover, Germany

Abstract:

Evidence-based research is becoming increasingly important in educational research. Calculation and test methods available in statistical software packages such as SPSS and STATA are widely used. To evaluate teaching innovations such as blended learning against classical classroom settings, for example, previous studies have mainly applied inference methods such as the t-test or variance analyses. The problem with these methods is that they test for the difference. A non-significant result does not automatically mean equivalence of the treatments examined, which is why we propose the use of equivalence testing. This paper introduces the equivalence test as complementary to the classical t-test and briefly discusses other approaches based on confidence intervals and Bayesian methods. As an example, the introduction of a blended learning format to a Bachelor's degree program is used to demonstrate the procedure and discuss the results of conducting an equivalence test. By combining tests for difference and equivalence successfully, it was possible to arrive at more informative statistical statements: Whereas a t-test alone only produced results for three out of 22 courses, a t-test and an equivalence test in combination yielded statistically confirmed statements for 12 out of 22 courses.

Keywords: blended learning; equivalence testing; flexible learning; learning effectiveness; methodology

This is an Accepted Manuscript of an article published by Taylor & Francis in the *International Journal of Research & Method in Education* on 24 Oct 2019, available online: https://doi.org/10.1080/1743727X.2019.1680621

Universities are facing major challenges such as technological innovations (MOOCs), increasing competition, and a highly mobile and globalized student body. One way of responding to this is the introduction of flexible learning, which allows students to learn independently of time and place, and to choose their learning paths. In the case of flexible learning, according to Chen (2003), flexibility must be present in at least one of the following learning dimensions: time, place, pace, learning style, content, assessment, or learning path. From an institutional point of view, this definition also means a change in the organization of teaching and learning. For example, the content must be made available in such a way that students can access it anytime and anywhere. This is the most basic form of flexible learning and in this sense flexible learning is often used synonymously with terms such as e-learning, open learning, distance learning, or blended learning (Tucker & Morris, 2012). Blended learning is commonly understood as a combination of face-to-face instruction and computer-mediated learning (Graham, 2006). Brown (2016) points out that although an increasing number of online tools are being used to enrich face-to-face learning, it is only possible to speak of a truly blended learning setting if the online elements and face-to-face elements are sensually and purposefully combined with each other; the simple upload of documents on an LMS is not enough.

Such a blended learning setting becomes a flexible learning design when not only the classroom teaching is enriched, but also a new composition of the module with higher degrees of freedom for learners takes place. This means they can study more independently of time and place than before or they can individually determine the content and pace of learning.

It is difficult to say whether a learning format such as this is more effective than a traditional one. Previous meta-analyses (e.g., Vo, Zhu, & Diep, 2017), which identify blended learning as more efficient than traditional forms of learning, usually do not indicate whether conventional teaching is supplemented by e-learning or whether it has been replaced. The interesting question is whether online elements are capable of replacing face-to-face courses in part and enabling more flexible learning. To answer this question, the determination of learning performance is of central importance. Only when flexible learning can offer the same or better results than face-to-face learning – that is if flexibility is not at the expense of quality – can universities offer and extend this new learning format successfully.

Looking at 40 blended learning studies compiled by Vo et al. (2017) in their current meta-analysis, 22 of them compare face-to-face lessons with blended learning while the other studies either examine e-learning-enriched, face-to-face courses or compare face-to-face courses with purely online sequences. Regarding learning effectiveness, the 22 blended learning studies present the following picture[1]:

- Eight studies conclude that blended learning produces better results than pure face-to-face teaching (Alonso, Manrique, Martínez, & Viñes 2011; Al-Qahtani & Higgins, 2013; Day & Foley, 2006; Lim, Kim, Chen, & Ryder, 2008; Melton, Bland, & Chopak-Foss, 2009; Pereira et al., 2007; Uzun, & Senturk, 2010; Vernadakis, Giannousi, Derri, Michalopoulos, & Kioumourtzoglou, 2012).
- Eight studies find no difference (Aly, Elen, & Willems, 2004; Demirer, & Sahin, 2009; Delialioglu, & Yildirim, 2007; Frederickson, Reed, & Clifford, 2005; Howerton, Enrique, Ludlow, & Tyndall, 2004; Larson & Sung, 2009; Reasons, Valadares, & Slavkin, 2005; Utts, Sommer, Acredolo, Maher, & Matthews, 2003).
- Two studies find better results for face-to-face learning compared with blended learning (Gundlach, Richards, Nelson, & Levesque-Bristol, 2015; Senn, 2008).
- The other studies cannot be classified owing to their study design or results (Dowling, Godfrey, & Gyles, 2003; Hui, Hu, Clark, Tam, & Milton, 2008; Maki & Maki, 2002; Schunn & Patchan, 2009).

This article takes this inconsistent picture as an opportunity to consider whether we might be asking the wrong methodological question when comparing blended learning with face-to-face settings. Currently, comparative pedagogical studies usually rely on classical statistical tests for significant differences. The simplest example is the (two-sided) two-group (unpaired) t-test, where the null hypothesis of equal means in two populations is tested against the alternative hypothesis of a non-zero difference. Variants and extensions include analysis of variance (ANOVA), where more than two groups can be compared or several factors can be investigated, and linear regression and analysis of covariance, which can be used to investigate the influence of numerical

[1] See detailed references in Vo et al. (2017).

independent variables on a response. Alternatively, nonparametric analogues of these tests which are valid under less stringent assumptions are also available, for example, the Wilcoxon rank-sum test (also known as Mann-Whitney U-test). Generally, all of these are used to test a null hypothesis of "no effect" (no differences between groups) against the alternative of a non-zero effect. A significant result then means the existence of a non-zero effect is statistically proven (with some probability of error and if all assumptions necessary for the test are fulfilled). Since classical statistical hypothesis tests (Neyman-Pearson theory) treat the null and alternative hypotheses asymmetrically, these methods cannot be used to prove the null hypothesis of a zero effect statistically (Schmidt, & Hunter, 1997). If the null hypothesis is not rejected, it may either be true or we might not be able to reject it because of low power of the test (e.g., due to small sample size or high variance). This is analogous to a court case where a defendant is convicted only if there is enough evidence of his or her guilt but is otherwise acquitted. Consequently, a defendant who is acquitted may be innocent or might be guilty (with insufficient evidence for a conviction).

In many cases, however, the aim of a study is not to establish a difference between groups but to show equality, or more precisely equivalence, i.e., near-equality up to practically irrelevant differences. While the – still very common – practice of taking a failure to reject the null hypothesis of no difference as proof for equality is fundamentally flawed, alternative approaches exist both within and without the classical significance testing paradigm but are usually not applied in pedagogical research. For example, in the 40 studies on blended learning used by Vo et al. (2017), 18 cases used a t-test and ten cases ANOVA or ANCOVA.

The t-test and ANOVA or ANCOVA, however, are tests for inequality. It is questionable whether this makes sense for studies related to blended learning in which face-to-face teaching is not supplemented by e-learning (in the sense of an enrichment strategy,) but, instead, a substantial part of face-to-face teaching is replaced. In certain contexts, however, statistical evidence of equivalence can be of great importance for decision-makers. According to Owston and York (2018):

> Faculty and institutions typically decide a priori to use a blended approach for reasons such as providing more convenience and flexibility to students or better utilization of classroom space, as long as they are assured that students will achieve at least as well as they would in face-to-face classes. (p. 22)

To provide this proof, another statistical method should be employed to test the logic in reverse. In this case, H_0 (mean values are not equal) is rejected in favour of H_1 (mean values are equal up to practically irrelevant differences). A testing method of this kind would be the two-sample test for equivalence (Wellek, 2010; see also Meyners, 2012). Equivalence tests were originally developed in epidemiology to prove that a newly developed, cheaper drug works just as well as an existing product (Schuirmann, 1987). However, as Dinno (2014, p. 2344) states "…evaluating evidence of equivalence is generally useful to the sciences because it allows the burden of evidence to be shared evenly between demonstrating the existence of a relationship and demonstrating the absence of a relationship."

This paper is structured as follows: First, we discuss alternative approaches to address this issue and introduce a test procedure similar to the t-test – the equivalence test – which is better tailored to the context of flexible or blended learning. We use the [university anonymous] flexible learning program as an example to illustrate objectives and considerations when implementing flexible learning in a blended learning format and demonstrate the application of the equivalence test using courses from the new FLEX programs examples. Finally, we raise the question of whether a combination of the t-test and the equivalence test could lead to more informative results in the evaluation of blended learning.

The Methodology of Testing Equivalence

To test statistically for equality instead of a difference as in the commonly used t-test, one possibility is to abandon the Neyman-Pearson significance testing framework altogether and adopt a Bayesian approach which avoids the asymmetry between null and alternative hypotheses present in classical significance testing. Bayesian statistics are based on philosophical foundations different from those of classical frequentist statistics, in particular with respect to the nature and interpretation of probabilities. Practically speaking, the main difference is that the uncertainty about parameters is formulated using probability distributions (so-called priors) while parameters are treated as unknown but fixed quantities in classical frequentist statistics. Bayesian approaches to hypothesis testing have also been suggested as a remedy for the inconclusiveness of non-significant results in classical hypothesis testing, see for example Dienes (2014) and Foster (2018). In Bayesian hypothesis testing, the null and alternative hypotheses are treated on an equal footing, but the researcher has to be able to specify prior

probabilities for both hypotheses before seeing the data. For Bayesian hypothesis testing, the influence of the prior distribution and the data (entering via the so-called Bayes factor) can largely be separated, and while different researchers may assign different prior probabilities to the hypotheses, if they see the same data, they will calculate their posterior probabilities using the same Bayes factor. In this sense, the Bayes factor is often considered an objective measure of evidence for the hypotheses (Kass, & Raftery, 1995). However, while the prior probabilities *for* the hypotheses do not enter in the calculation of the Bayes factor, when a hypothesis consists of more than one possible value for a parameter, a prior distribution on the parameter values *under* the hypothesis is still needed. Hence, for a "nonzero effect" hypothesis, one would still have to specify the probabilities for effects of different sizes given that some effect exists, making the approach again dependent on prior probabilities. Just replacing the classical t-test by its Bayesian counterpart still tests a strict "no-effect" hypothesis, with the difference that, unlike in the frequentist framework, evidence in favour of a zero effect can be assessed. Usually, it is not the goal of the researcher to prove that the effect is exactly zero but to show that if it does exist, it is small enough to be practically irrelevant. This leads to Bayesian testing of interval hypotheses, for which Bayes factors can be calculated (Morey & Rouder, 2011). This approach can be described as a Bayesian counterpart to the frequentist equivalence test we describe below. Another widely used Bayesian approach is based on the region of practical equivalence (ROPE) (Kruschke, 2013), where a range of parameters around zero is specified so that true parameter values within this region correspond to practically irrelevant effects (similar to the interval specified in the equivalence hypothesis). Then a highest posterior density interval (HDI) is calculated, which is a Bayesian counterpart to a frequentist confidence interval. Based on whether the HDI is completely in the ROPE, falls completely outside the ROPE, or is partly contained in the ROPE and partly falls outside the ROPE, one concludes equivalence (practically irrelevant effect) or non-equivalence (effect large enough to be pratically relevant), or one declares the results inconclusive. While these approaches are motivated differently (as testing or interval estimation procedures), Liao, Midya, & Berg (2019) have recently shown interesting formal relations between the two.

Confidence intervals are often seen as more informative than *p*-values from a hypothesis test, as they provide a range of plausible values for the parameter under scrutiny (for a fixed confidence level). Confidence intervals hence provide not only

information about the uncertainty of the existence of an effect but also indicate its size and direction. Reporting of confidence intervals instead of *p*-values has also been suggested as a remedy for some problems with hypothesis testing in the behavioural sciences (Cumming, 2012). Although different in interpretation, confidence intervals like hypothesis tests originate from classical (frequentist) statistical theory. Indeed, they are generally largely equivalent to Neyman-Pearson hypothesis tests in the sense that a level α test can be performed by checking whether the parameter value from a (point) null hypothesis is included in a $(1-\alpha)$ confidence interval and a confidence interval with coverage probability of $(1-\alpha)$ can be constructed by including all parameter values for which the corresponding (point) null hypothesis would not be rejected at significance level α. Hence, confidence intervals and hypothesis tests generally lead to similar conclusions and confidence intervals share some of the difficulties with hypothesis tests (e.g., coverage probabilities of confidence intervals are prone to similar misinterpretations as *p*-values or significance levels in testing).

Confidence intervals on either a difference of mean values or a standardized difference of mean values can be used to assess equivalence or non-equivalence of treatments by comparing the location of the confidence interval calculated from the sample to a pre-defined interval where treatments are considered equivalent (see also Lakens, Scheel, & Isager, 2018):

1. If the confidence interval is entirely included in the equivalence interval, one can conclude that the treatments are equivalent.
2. If the confidence interval is entirely outside the equivalence interval, one can conclude that the treatments are not equivalent, i.e., the effect is larger than the allowed tolerance.
3. If the confidence interval includes both values inside and outside the equivalence interval, the result is inconclusive, and neither equivalence nor non-equivalence can be ruled out.

Note that this is similar to the Bayesian approach using the Region of Practical Equivalence.

If equivalence of two treatments is to be proven, this is however also entirely possible within the classical Neyman-Pearson testing framework, provided the hypotheses are formulated correctly. Equivalence tests basically switch the more traditional “no effect” hypothesis and the alternative. The null hypothesis of the equivalence tests then states that there is an effect that is at least as large as a specified threshold, while the

alternative states that the effect is smaller than the threshold (but not necessarily zero). Rejecting the null hypothesis, therefore, statistically establishes equivalence. While there are also mathematical reasons why the alternative cannot be formulated as a zero effect, usually researchers want to prove it is so small that it is practically irrelevant.

The wide-spread flaws when statistically proving equivalence hence are not a problem of classical hypothesis testing per se but of a common misapplication of the t-test, and the problem can be correctly addressed within the framework that is (still) the most familiar to many researchers. There are of course reasons why researchers may prefer either confidence intervals or Bayesian approaches over classical significance testing on more general grounds, and as pointed out above, the problem can be adequately addressed in these settings as well. But as such, the equivalence problem is neither an argument for nor against classical hypothesis testing, and some of the challenges – like determination of reasonable equivalence limits – are intrinsic to the problem and have to be addressed whether using Bayesian or frequentist approaches. In the following, we restrict ourselves to the description of equivalence testing within a Neyman-Pearson framework.

Equivalence tests exist for many different situations; see Wellek (2010) and Meyners (2012). They are routinely used in pharmaceutical research but are not as well known in other fields.

If an equivalence analogue of the classical two-sample, unpaired t-test for a zero effect is desired, an important choice to be made is whether the effect of the treatment should be specified as the absolute difference in means or a standardized difference (difference in means divided by the standard deviation). The former hypothesis can be tested by a combination of two one-sided t-tests (TOST); see Schuirmann (1987). This is probably the most widely known method for tests of equivalence and is most appropriate in fields such as pharmaceutical research, where effects are measured on scales with well-defined units. There are also several articles advocating the use of TOST in fields like psychology or educational research; see for example Rogers, Howard, and Vessey (1993) or Lakens (2017). However, in educational research, a relative measure of the effect, such as the standardized mean difference (difference of population means divided by the standard deviation), is often desired since outcomes can be measured on very different scales. An optimal test (in the sense of maximizing power) exists for this situation (Wellek, 2010) and we will describe this approach now.

Assume that there are two independent samples of normally distributed observations and that the unknown variances are equal:

$$X_i \sim N(\mu_1; \sigma^2) \text{ for } i = 1, \ldots, m$$

$$Y_j \sim N(\mu_2; \sigma^2) \text{ for } j = 1, \ldots, n$$

Define the standardized difference in means by

$$\theta = \frac{\mu_1 - \mu_2}{\sigma}$$

The null hypothesis to be tested is

$$H_0: \quad \theta \leq -\varepsilon_1 \quad \text{or} \quad \theta \geq +\varepsilon_2$$

against the alternative

$$H_1: \quad -\varepsilon_1 < \theta < +\varepsilon_2$$

We will also refer to the (open) interval $(-\varepsilon_1, \varepsilon_2)$ as the equivalence interval and to values of θ outside this interval as relevant effects. The equivalence limits ε_1 and ε_2 have to be chosen based on subject-matter considerations so that a standardized mean difference within these limits can be considered too small to be practically relevant. The choice may be guided by several different interpretations of θ, ε_1 and ε_2:

- The parameter θ is the theoretical version of Cohen's d, a widely used measure of effect size. The equivalence limits may thus be interpreted on Cohen's scale (Cohen, 1992).
- Different values of θ correspond to different amounts of overlap between two normal distributions with the same variance, which allows for a choice based on a graphical representation; see Figure 1.
- In case of a zero effect, the probability p that a randomly chosen student receiving the treatment has a higher outcome than a randomly chosen student not receiving the treatment is exactly 0.5. If there is an effect, p will differ from 0.5. The limits ε_1 and ε_2 on θ can be translated to limits on the deviation of p from 0.5 and vice versa; see Wellek (2010, Chapter 1.7). Figure 1 also gives some examples.

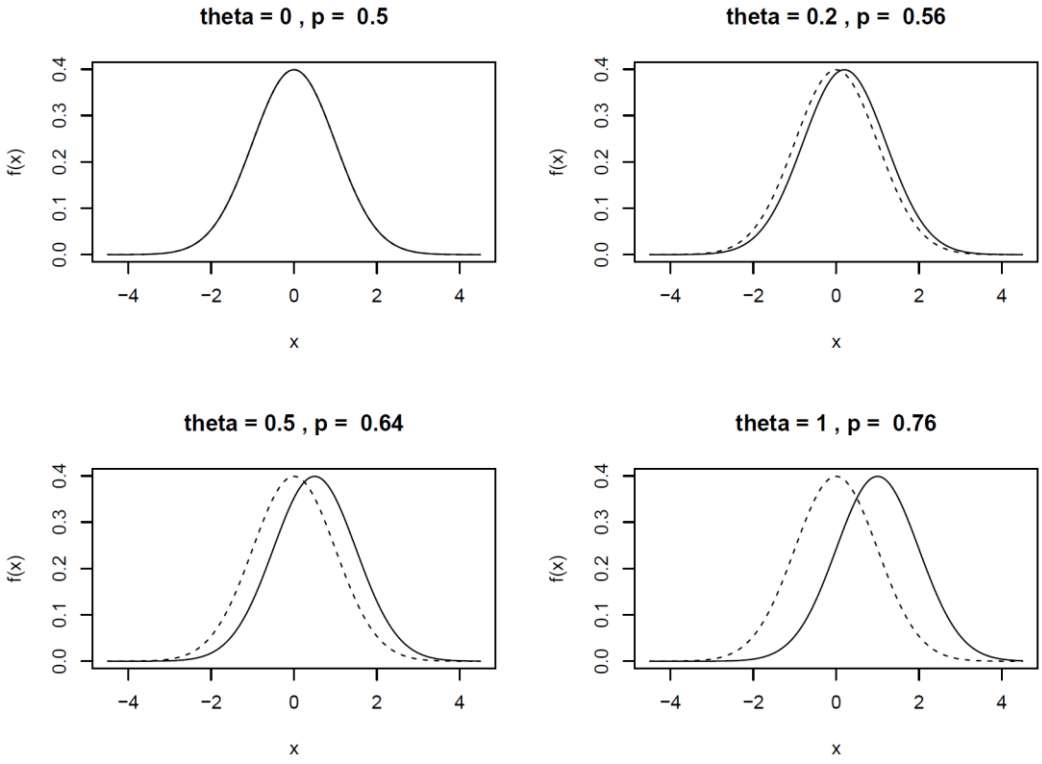

Figure 1. Schematic view of treatment (solid line) and reference (broken line) groups for different values of standardized mean reference theta (θ).

In the situation described above, a uniformly most powerful invariant (UMPI) level-α test can be derived (see Wellek 2010, Chapter 6.1). The UMPI test is optimal in a large class of reasonable tests in the sense that there is no test for which

- the rejection probability is at most α for any value of θ belonging to the null hypothesis,
- the p-value does not change when all observations $(x_1, \ldots, x_m, y_1, \ldots y_n)$ are replaced by $(ax_1 + b, \ldots, ax_m + b, ay_1 + b, \ldots, ay_n + b)$ for arbitrary constants $a > 0$ and b, i.e., a change of units does not change the conclusion, and
- the rejection probability (power) for at least one value of θ belonging to the alternative is strictly larger than the corresponding probability for the UMPI test.

The test statistic is the usual t-statistic (with pooled variance)

$$T = \sqrt{\frac{mn(m+n-2)}{m+n}} \cdot \frac{\bar{X} - \bar{Y}}{\sum_{i=1}^{m}(X_i - \bar{X})^2 + \sum_{j=1}^{n}(Y_i - \bar{Y})^2}$$

but the critical region and p-values are different from the usual t-test. They can be determined from the distribution of T for the values on the boundaries between the null and alternative hypotheses. In the case of a symmetric equivalence interval, i.e., ε_1

$=\varepsilon_2=\varepsilon$, the critical region of the test can be given explicitly (Formula 6.6. in Wellek 2010, p. 121) and by the same reasoning, the *p*-value is given by

$$\sqrt{F_{1,n+m-2,\lambda^2}(T^2)}$$

where $F_{p,q,\lambda 2}$ is the distribution function of the noncentral F-distribution with p and q degrees of freedom and non-centrality parameter

$$\lambda^2 = \frac{mn\varepsilon^2}{m+n}$$

The null hypothesis is then rejected in favour of the alternative if the *p*-value is smaller than the chosen significance level α. In this case, the conclusion is that the two treatments are equivalent. It should be noted that – as with any other test – failure to reject the null hypothesis does not prove the null hypothesis, i.e., a non-significant result does not imply that the treatments are not equivalent. We use the implementation of the UMPI equivalence test available in package equivUMP for R.

With the described methodology, the equivalence of two treatments can be established statistically. The method can also be combined with the classical t-test as suggested by Dinno (2014). The combination of the tests for difference and equivalence yields four possible conclusions; see Table 1 (adapted from Dinno, 2014). This can help especially when putting significant or non-significant results from t-tests in context. For example, a significant result from a t-test is often (wrongly) taken to imply a practically meaningful effect, but if the equivalence test also yields a significant result, we know that while we have established the existence of an effect, it is too small to be practically relevant.

	t-test *significant*	t-test *not significant*
Equivalence test *significant*	Conclude *Trivial effect*	Conclude *Equivalence*
Equivalence test *not significant*	Conclude *Difference*	Conclude *Inconclusive*

Table 1. Combining results for tests for difference and equivalence (adapted from Dinno, 2014)

Note that in contrast to the stronger claim made by Dinno (2014), in the case of a significant result for the t-test and a non-significant equivalence test (lower left field of the table), we cannot conclude that the effect is necessarily relevant but only that a

relevant effect cannot be ruled out, i.e., the effect may either be larger than the equivalence bounds or it may be smaller but the equivalence test has not enough power to detect equivalence.

Study Context FLEX

The following section uses these considerations and applies them to a use case. The [university anonymous] launched the new study format for flexible learning (FLEX) in 2015 as part of a comprehensive e-learning strategy. The Bachelor's degree in Banking and Finance (BSc B&F), a successful and established program, was selected as the first FLEX study program. The BSc B&F is already being run as a full-time (FT) and part-time (PT) program. Accordingly, the FLEX format is the third study format for this degree program. The first part of the curriculum consists of an assessment level worth 60 ECTS (European Credit Transfer System). Here, together with other areas of specialization at the school, basic knowledge of business economics is taught. In the main section of the program, 66 ECTS are awarded for specialization in Banking & Finance and 54 ECTS for general business management topics. In the part-time program, lessons are held on one weekday and a maximum of two evenings and/or Saturday mornings. Full-time programs are normally conducted in six semesters while part-time and FLEX programs cover eight semesters. For part-time studies (including FLEX), a maximum vocational employment level of 60% is recommended.

The main objective of the newly introduced FLEX format was to offer students the best possible opportunities to combine their work or private responsibilities with a flexible study program. Regarding the number and distribution of face-to-face lessons over the 14-week term, compatibility with a distant place of residence was the guiding principle, e.g., up to how many out-of-home overnight stays are acceptable for potential students. At the same time, regular physical face-to-face meetings should foster reflection of the course content developed during the online phases. As a result of these considerations, on-site classroom teaching for FLEX was reduced by about half compared to the part-time program and replaced with online sessions over periods of three weeks. This means that FLEX students attend the university every three weeks for two days and the interject self-study phase allows students to learn flexibly in terms of time and place and to follow their preferred learning path. The selected 49% face-to-face time corresponds to the current state of empirical knowledge regarding blended learning, namely that with an online ratio of one third to one half, learning success is

higher than for blended learning with a smaller online proportion (Owston & York, 2018).

With regard to the dimensions of flexible learning according to Chen (2003), the FLEX format offers greater flexibility in terms of time, place, pace, learning style, and learning path than the conventional study format, but not in terms of assessment and content, which are identical in the FLEX and conventional study formats. As a result, FLEX students take exactly the same examinations as students in the part-time program and at the same time, which allows for a comparison of the exam results with high empirical significance.

The research design consists of the experimental group FLEX (Cohort 15, N = 28; Cohort 16, N = 28) with students attending all courses in the new FLEX format and a control group part-time (PT) (Cohort 15, N = 100; Cohort 16, N = 117). The design is tightly controlled for a long-term field study in an educational area, firstly because the framework conditions are comparable with the same learning objectives and identical assessment, and secondly because the presence of a control group means that a quasi-experimental design is available (see also Fraenkel, Wallen, & Hyun, 2015). Care was also taken to ensure that the experimental and control groups were taught by the same lecturer wherever possible.

From a study context, it is clear that the central objective of implementing FLEX is not to improve the test performance of students, but that they should be able to achieve equivalent exam results despite more flexible study conditions and a lower proportion of face-to-face meetings.

Performance Testing in FLEX

The following Tables 2 and 3 list the exam results of Cohort 15 (beginning in fall semester 2015) and Cohort 16 (beginning in fall semester 2016) for the assessment level (Semesters 1-3). The exam results of the FLEX students (FLEX) are compared with those of the B&F students from the part-time program (PT). The assessment is identical in all courses, and the exams are not corrected by the lecturer of the respective class but by an independent pool of lecturers. The FLEX and PT samples are independent, and the sample size and histograms of the test results do not indicate a violation of the requirements of normal distribution and uniformity of variance.

Courses (Semester)	*FLEX format (FLEX)*			*Part-Time format (PT)*			*d*	*p* t-test	*p* E-test
	N	*M*	*SD*	*N*	*M*	*SD*			
Business Administration (1)	27	4.24	0.53	93	4.17	0.67	0.12	0.598	0.037*
Mathematics 1 (1)	27	4.19	0.90	92	4.11	0.76	0.10	0.660	0.029*
Business Law (1)	28	4.23	0.88	92	4.15	0.90	0.10	0.659	0.028*
Marketing (1)	28	4.18	0.56	94	4.29	0.50	-0.22	0.310	0.096
Mathematics 2 (2)	21	4.31	0.73	81	4.23	0.83	0.09	0.706	0.040*
Business English 1 (2)	18	4.50	0.64	83	4.33	0.73	0.24	0.350	0.160
Financial Accounting (2)	20	4.08	0.78	79	4.25	0.79	-0.22	0.385	0.128
Strategy (3)	21	4.83	0.53	78	4.82	0.68	0.02	0.937	0.008**
Communication (3)	20	4.20	0.66	76	4.11	0.65	0.15	0.564	0.074
Microeconomics (3)	21	3.71	0.73	74	3.74	0.76	-0.04	0.877	0.016*
Business English 2 (3)	19	4.58	0.51	75	4.43	0.60	0.26	0.313	0.174

Note: E-Test = Equivalence-Test. * significant at $\alpha = 0.05$ (two-tailed), ** significant at $\alpha = 0.01$ (two-tailed)

Table 2. Statistical analysis of course grades FLEX and PT assessment level, cohort 15

The results for Cohort 15 (see Table 2) show that the mean values differ only slightly. The direction is indicated by the effect size (Cohen's *d*); in eight of the 11 courses examined, the mean values of the FLEX cohort are higher than those of the PT students (positive sign, range of marks from 1-6, where 6 is the best performance, and all grades below 4 are unsatisfactory). The results of the t-test do not show any significant differences in the exam results between FLEX and PT students, but in six cases the equivalence test is significant. For our purposes, we set the equivalence boundaries to $\varepsilon = 0.5$, i.e., we only regard standardized mean differences as relevant if they are larger than 0.5 in absolute value. Thus, the H_0 hypothesis can be rejected for these courses: The exam results for the two groups with experimental design (FLEX) and traditional design (PT) can be regarded as statistically equivalent.

To consider a possible bias at the entry competence level of the first FLEX cohort, the exam results of the second year (Cohort 16) were also analysed (see Table 3). Cohort 16 of the FLEX format also has higher mean values than the control group of PT students (in 8 out of 11 courses). For the three courses Business Law ($t(139) = 2.23$, $p = 0.028$, with effect size Cohen's $d = 0.47$), Business English 1 ($t(117) = 2.04$, $p = 0.044$, $d = 0.47$), and Business English 2 ($t(110) = 2.07$, $p = 0.041$, $d = 0.48$) significant differences can be observed. FLEX students have achieved significantly better exam results in these courses compared to the PT Students. The results for three courses (Business Administration, Marketing, and Microeconomics) show very few differences and the equivalence test is significant.

Courses (Semester)	*FLEX format (FLEX)*			*Part-Time format (PT)*					
	N	*M*	*SD*	*N*	*M*	*SD*	*d*	*p* t-test	*p* E-test
Business Administration (1)	28	4.23	0.74	117	4.25	0.70	-0.03	0.894	0.006**
Mathematics 1 (1)	28	4.04	0.82	108	3.79	0.79	0.31	0.141	0.190
Business Law (1)	28	4.34	0.72	113	3.98	0.78	0.47	0.028*	0.442
Marketing (1)	28	4.14	0.54	110	4.20	0.67	-0.09	0.677	0.023*
Mathematics 2 (2)	23	3.98	0.70	96	3.68	1.02	0.31	0.183	0.208
Business English 1 (2)	24	4.71	0.78	95	4.31	0.87	0.47	0.044*	0.438
Financial Accounting (2)	22	4.50	0.67	91	4.26	0.92	0.28	0.248	0.172
Strategy (3)	23	4.70	0.42	90	4.41	0.67	0.45	0.056	0.416
Communication (3)	22	4.14	0.47	90	4.01	0.67	0.21	0.388	0.107
Microeconomics (3)	22	4.07	0.54	87	4.08	0.97	-0.01	0.955	0.005**
Business English 2 (3)	23	4.59	0.56	89	4.24	0.76	0.48	0.041*	0.471

Note: E-Test = Equivalence-Test. * significant at $\alpha = 0.05$ (two-tailed), ** significant at $\alpha = 0.01$ (two-tailed)

Table 3. Statistical analysis of course grades FLEX and PT assessment level, cohort 16

In total, out of the 22 courses for the first two cohorts examined for the assessment level, three courses can be designated as *Difference* (with better results in FLEX), nine courses as *Equivalence* (statistical equivalent), and ten courses as *Inconclusive* (no statement possible, see Table 4).

	t-test *significant*	t-test *not significant*
Equivalence test *significant*	Conclude *Trivial effect* (0 courses)	Conclude *Equivalence* (9 courses)
Equivalence test *not significant*	Conclude *Difference* (3 courses)	Conclude *Inconclusive* (10 courses)

Table 4. Combining FLEX test results for difference and equivalence

When comparing the two cohorts, it is noticeable that in Cohort 15, six of 11 courses are statistically equivalent while in Cohort 16 there are only three, but in addition, the FLEX students from Cohort 16 achieve significantly better exam results in three courses. This means that in both years more than half the courses are at least equivalent to conventional part-time teaching, while in the other courses a statistically verified statement is not possible.

In summary, it can be concluded that students in the assessment level of the FLEX format achieve exam results at least equivalent to students in the part-time

format. These FLEX results confirm previous findings regarding blended learning (see, e.g., Bernard, Borokhovski, Schmid, Tamim, & Abrami, 2014; Means, Toyama, Murphy, & Baki, 2013) which have shown that students in blended learning courses achieve at least equivalent or even slightly better exam results compared to students on face-to-face courses.

In the case described here, proof that students in a blended learning study program with reduced face-to-face time achieve equivalent results led to the continuation of the new study program and the transition of further study programs into a blended learning FLEX format.

Conclusion

Innovation in teaching is labour-intensive and time-consuming. Additionally, educational institutions have a great responsibility towards the learner; they must ensure that learning design produces the best possible learning outcomes and is as efficient as possible. This is perhaps particularly true in the area of higher education, where the cost of studying is not insignificant, whether at the expense of the student or the public sector.

The transition from traditional learning methods to new forms of teaching must, therefore, at best, be empirically justified. In the context of the introduction of flexible learning formats, a blended learning programme is often compared to one with a traditional format. This comparison is carried out statistically using a t-test or an ANOVA, and it is checked whether the two formats differ significantly from each other. At this point, however, the goal of the learning innovation and the applied statistical procedure often do not fit together.

Teaching innovations such as the introduction of blended learning formats are not primarily aimed at improving learning performance, but rather at optimizing learning and teaching conditions for students and lecturers. For example, students are to be given greater flexibility in terms of time or space, which will allow them to combine study and work more effectively than before. This flexibilization of study programs is intended to reduce dropout rates and address new target groups. It is not a question of improving the learning performance of the students themselves, but of keeping this performance at least at the same level but with greater organizational flexibility. The aim is, therefore, to prove similarity and rather than difference.

In such contexts, conventional statistical tests as typically used in educational research, which test for differences, reach their limits. A wrong conclusion is often drawn from a non-significant t-test (or ANOVA, Mann-Whitney-U-test, etc.); if no difference can be found, the groups are declared equal. However, as many others have already said, “the absence of evidence is not evidence of absence.” A non-significant result in a t-test is a clue but it is not evidence, since failure to reject the null hypothesis may be due to low power, for example, because of small sample sizes.

It is, however, entirely possible to statistically establish absence of a practically relevant difference within the classical frequentist hypothesis testing framework if the right hypothesis is tested. As described above, both Bayesian counterparts and approaches based on confidence intervals are also available. The equivalence test can be used to make statistically valid judgments about the equivalence of study formats and can avoid the frequent misinterpretation of an insignificant *p*-value in a difference test (e.g., t-test or ANOVA) as evidence of equivalence. The equivalence test is relatively easy to employ and follows the same test logic as the t-test (basically switching H_0 and H_1).

However, we consider it useful to combine the t-test (or similar methods) with the equivalence test. To do so, a classical t-test is performed first, followed by an equivalence test on the same sample. The results of the two test runs can be located in a four-field matrix: (1) statistically equivalent and not different, (2) not equivalent and statistically different, (3) statistically equivalent and statistically different, and (4) not equivalent and not different. Failure to reject the null hypothesis may always be due to low power, and especially case (4) where the groups can neither be shown to be different nor equivalent strongly indicates low power (usually due to small samples or high variance).

With a combination of the tests, statistically confirmed statements concerning 12 out of 22 courses could be made in this case study of the FLEX study program. With the t-test alone, such a statement would have been possible for only three courses. Nevertheless, it was possible to prove for the new study program format that in 55% of the courses increased flexibility was not at the expense of the learning outcome. The other courses were statistically neither equivalent nor different.

The fact that the equivalence test has not played a role in empirical pedagogical research so far seems to have mainly historical reasons. The equivalence test is virtually unknown in psychological research, which feeds into pedagogical research.

Furthermore, it is not taught on statistical courses, not mentioned in relevant textbooks, and not provided in SPSS.

Despite this, several different add-on packages are available for the open-source statistic software R. Our results were obtained using equivUMP (Mildenberger, 2019). The UMPI test used here is also available in EQUIVNONINF (Wellek, & Ziegler 2017), although this implementation calculates critical regions instead of p-values. The widely used TOST approach is available in TOSTER (Lakens, 2018). Bayesian procedures based on ROPE are implemented in BEST (Kruschke, & Meredith, 2018), Bayes factors can be calculated using BayesFactor (Morey, & Rouder, 2018). For STATA, Dinno (2018) provides code for TOST and other equivalence tests, although the UMPI test for standardized mean differences described here is not included. In addition, the TOST procedure for unstandardized mean differences could always be carried out manually using *any* statistical software that can perform the two one-sided t-tests separately or calculate the corresponding confidence interval.

The real challenge when testing equivalence – regardless of whether the UMPI tests described here, a confidence interval or a Bayesian methods are used – is setting equivalence limits. If there are no theoretical considerations, these can be based on the benchmarks for small, medium, and large effect sizes as a starting point. The application of equivalence tests in pedagogical research would lead to a stronger differentiation of boundaries in the future.

Ultimately, the application of the equivalence test extends the methodological repertoire for evidence-based pedagogical research, enables more reliable statements to be made when one teaching innovation is compared with another, and helps decision-making in higher education institutions.

References

Bernard, R. M., Borokhovski, E., Schmid, R. F., Tamim, R. M., & Abrami, P. C. (2014). A meta-analysis of blended learning and technology use in higher education: from the general to the applied. *Journal of Computing in Higher Education, 26*(1), 87-122. doi: 10.1007/s12528-013-9077-3

Brown, M. G. (2016). Blended instructional practice: A review of the empirical literature on instructors' adoption and use of online tools in face-to-face teaching. *The Internet and Higher Education, 31*(Supplement C), 1-10. doi: https://doi.org/10.1016/j.iheduc.2016.05.001

Chen, D.-T. (2003). Uncovering the provisos behind flexible learning. *Educational Technology & Society, 6*(2), 25-30.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155-159.

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis.* New York, NY: Routledge.

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. Frontiers in Psychology (5), Article 781, 1-17. doi: 10.3389/fpsyg.2014.00781

Dinno, A. (2014). Comment on "The Effect of Same-Sex Marriage Laws on Different-Sex Marriage: Evidence From the Netherlands". *Demography, 51*(6), 2343-2347. doi: 10.1007/s13524-014-0338-1

Dinno, A. (2018). *tost: Two One-Sided Tests of Equivalence.* STATA package version 3.0.2. [Computer software]. Retrieved from https://www.alexisdinno.com/stata/tost.html

Foster, C. (2018). Developing mathematical fluency: comparing exercises and rich tasks. *Educational Studies in Mathematics 97*, 121-141

Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (2015). *How to design and evaluate research in education* (Ninth edition ed.). New York: McGraw-Hill.

Graham, C. R. (2006). Blended learning systems: definition, current trends, and future directions. In C. J. Bonk & C. R. Graham (Eds.), *The handbook of blended learning: global perspectives, local designs* (pp. 3-21). San Francisco: Wiley & Sons.

Kass, R. E., Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773-795

Kruschke, J. K., Meredith, M. (2018*). BEST: Bayesian Estimation Supersedes the t-Test.* R package version 0.5.1. [Computer software]. Retrieved from https://CRAN.R-project.org/package=BEST

Lakens, D. (2017). Equivalence Tests: A Practical Primer for t Tests, Correlations, and Meta-Analyses. *Social Psychological and Personality Science, 8*(4), 355-362. doi: 10.1177/1948550617697177

Lakens, D. (2018*).TOSTER: Two One-Sided Tests (TOST) Equivalence Testing.* R package version 0.3.4. [Computer software]. Retrieved from https://CRAN.R-project.org/package=TOSTER

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence Testing for Psychological Research: A Tutorial. *Advances in Methods and Practices in Psychological Science, 1*(2), 259-269. doi:10.1177/2515245918770963

Liao, J.G., Midya, V., & Berg, A. (2019). *Connecting Bayes factor and the Region of Practical Equivalence (ROPE) Procedure for testing interval null hypothesis.* Preprint. Retrieved from https://arxiv.org/abs/1903.03153v2

Means, B., Toyama, Y., Murphy, R., & Baki, M. (2013). The Effectiveness of Online and Blended Learning: A Meta-Analysis of the Empirical Literature. *Teachers College Record, 115*(3), 1-47. Retrieved from: http://www.tcrecord.org/Content.asp?ContentId=16882

Meyners, M. (2012). Equivalence tests – A review. *Food Quality and Preference, 26*(2), 231-245. doi: https://doi.org/10.1016/j.foodqual.2012.05.003

Mildenberger, T. (2019*). equivUMP: Uniformly Most Powerful Invariant Tests of Equivalence.* R package version 0.1.1. [Computer software]. Retrieved from https://CRAN.R-project.org/package=equivUMP

Morey, R. D., & Rouder, J.N. (2011). Bayes Factor approaches for testing interval null hypotheses. *Psychological Methods, 16*(4), 406-419.

Morey, R. D., & Rouder, J.N. (2018). *BayesFactor: Computation of Bayes Factors for Common Designs.* R package version 0.9.12-4.2. [Computer software]. Retrieved from https://CRAN.R-project.org/package=BayesFactor

Owston, R., & York, D. N. (2018). The nagging question when designing blended courses: Does the proportion of time devoted to online activities matter? *The Internet and Higher Education, 36*(Supplement C), 22-32. doi: https://doi.org/10.1016/j.iheduc.2017.09.001

Rogers, K, Howard, I., Vessey, J. T. (1993). Using Significance Tests to Evaluate Equivalence Between Two Experimental Groups. *Psychological Bulletin 113*(3), 553-565

Schmidt, F. L., & Hunter, J. E. (1997). Eight Common But False Objections to Discontinuation of Significance Testing in the Analysis of Research Data. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 37-64). Mahwah, N.J.: Lawrence Erlbaum Associates Publishers.

Schuirmann, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinet Biopharm, 15*(6), 657-680.

Tucker, R., & Morris, G. (2012). By Design: Negotiating Flexible Learning in the Built Environment Discipline. *Research in Learning Technology, 20*(1), n1. doi: 10.3402/rlt.v20i0.14404

Vo, H. M., Zhu, C., & Diep, N. A. (2017). The effect of blended learning on student performance at course-level in higher education: A meta-analysis. *Studies in Educational Evaluation, 53*(Supplement C), 17-28. doi: https://doi.org/10.1016/j.stueduc.2017.01.002

Wellek, S. (2010). *Testing statistical hypotheses of equivalence and noninferiority* (2. ed. ed.). Boca Raton: CRC Press.

Wellek, S., & Ziegler, P. (2017). *EQUIVNONINF: Testing for Equivalence and Noninferiority.* R package version 1.0. [Computer software]. Retrieved from https://CRAN.R-project.org/package=EQUIVNONINF

Supplemental material

The file equivalence.R contains example code that shows how the UMPI equivalence test can be performed in R using the implementation in the equivUMP package (Mildenberger, 2019). All settings are the same as the ones used in the paper, and the artificial data set is similar to the actual data used in the study.