Samples in many cell-based experiments are matched/paired but taking this into account does not always increase power of statistical tests for differences in means

Adam Zweifach*
Department of Molecular and Cell Biology, University of Connecticut at Storrs, Storrs, CT 06269-3125

ABSTRACT **Power of statistical tests for differences in means is the probability of obtaining a statistically significant *p* value when means differ. When samples in experimental replicates come from a single cell culture, they are matched or paired because they share between-trials biological variability. This can cause positive correlation between values from conditions in a replicate. Correlation can also be caused in otherwise independent samples by shared technical variability. However, correlation is reduced by noise that affects samples individually. I investigated how to maximize power in experiments with two conditions over a range of correlations. Normalizing data to control increases the rate of false positives, if Student's *t* test is used. Paired *t* tests, theoretically the correct test for matched samples, have higher power than Student's *t* test when correlation is high, but lower power when correlation is low. Testing correlation to select a test for differences in mean can affect the subsequent rate of false positives. Ultimately, components of experimental variability must be considered to choose the most powerful two sample test for differences in mean. This contrasts with experiments with more than two conditions, where random-block ANOVA, a matched samples test, can be used as a default.**

Monitoring Editor
Thomas Pollard
Yale University

Received: May 8, 2023
Revised: Sep 25, 2023
Accepted: Oct 23, 2023

Instructional Resource

SIGNIFICANCE STATEMENT

- Many experiments in cell biology should be analyzed for differences in mean with statistical tests that take matching or pairing into account to maximize statistical power.
- Simply taking cells from the same culture to prepare samples is not sufficient to guarantee matching, and cells from independent sources can be matched or paired when processed in batches using some techniques.
- Researchers can always use matched samples ANOVA, but must choose between Student's *t* test and paired *t* tests when there are only two samples. I provide guidance for how to do this.

INTRODUCTION

A common goal in many experiments is to determine whether the mean value of a parameter of interest is different in two or more experimental conditions. Statistical tests (*t* tests whether there are two conditions or ANOVA followed by a post-hoc test if there are more than two) generate *p* values reflecting the probability random sampling from populations with identical means could have generated differences as large or larger than exist in the experimental samples. When *p* values are small, data are unlikely if population means are the same. Different versions of the tests make additional

This article was published online ahead of print in MBoC in Press (http://www.molbiolcell.org/cgi/doi/10.1091/mbc.E23-05-0159) on November 1, 2023.

*Address correspondence to: Adam Zweifach (adam.zweifach@uconn.edu).

Abbreviations used: ANOVA, analysis of variance; ANOVAn, ANOVA where data have all been normalized to control for that replicate; CV, coefficient of variation; ELISA, enzyme-linked immunosorbent assay; FPR, false positive rate; RBANOVA, random block analysis of variance.

assumptions about the distribution of values in the populations. Student's *t* test and one-way ANOVA, for example, also assume values are distributed identically.

Statistical power of a test for differences in means reflects the probability of getting a statistically significant result when population means are different. When power is low, differences in means are likely to be missed and differences found are more likely either to be false positives or inflated in apparent size (Button *et al.*, 2013; Krzywinski and Altman, 2013; Greenland *et al.*, 2016; Lazic, 2018; Marino, 2018). Some have claimed low power is contributing to a "reproducibility crisis" in science (Button *et al.*, 2013).

Power of tests for differences in means is higher when the *p* value chosen for statistical significance is larger, when the real difference between population means is greater, when scatter or variance in the populations is smaller, and when sample sizes are larger. In cell-based experiments, the number of replicates is often quite small–three is probably the most common choice–and so power tends to be low unless effects are large. This makes it critical to maximize power by optimizing other factors like basic experimental design (Lazic, 2018) and choice of statistical test. For example, Lazic showed analyzing samples acquired at different concentrations of a drug using linear regression can be a more powerful approach than ANOVA because fewer parameters need be estimated (Lazic, 2008). Lew explored the best way to apply analysis of variance (ANOVA) to an experiment comparing the effects of two drugs to control using a cultured cell line in replicates conducted over multiple days (Lew, 2007). There was enough day-to-day variability in the culture that one-way ANOVA followed by Dunnett's test comparing drug-treated means to control did not generate statistically significant results. Lew explained that because the cells used on each day were taken from the same culture, the samples were matched. As will be discussed further, matching can result in positive correlation between values obtained under different conditions in the same replicate. Lew used simulations to show that random block ANOVA (RBANOVA), a form of two-factor ANOVA without within-block replicates (i.e., for each condition in a replicate there is a single value) where one factor is the experimental run and the other is the treatment, had higher power than one-way ANOVA conducted either on raw data or on data normalized by dividing each day's results by that day's control. The latter is commonly used by laboratory scientists to minimize effects of between-runs variability. Finally, Lew showed RBANOVA can be used as a default test because it is only slightly less powerful than one-way ANOVA when there is no correlation between samples.

The type of cultured-cell experiment Lew described is extremely common, used not only to assess the effects of drugs but also genetic manipulations like transient overexpression and knockdown or knockout. It is reasonable to think experiments with only two conditions conducted on cells from a common source should be analyzed with paired *t* tests, which are the two-sample equivalent of RBANOVA. (The F statistic for treatment in ANOVA and RBANOVA is the square of the t statistic for students' *t* test or a paired *t* test, respectively, so the tests yield identical *p* values when applied to experiments with two conditions.) However, matched samples tests are reported far less often in journals that publish results of cell-based laboratory experiments than tests like one-way ANOVA and Student's *t* test that assume samples are independent. This may be because researchers do not know samples are matched/paired when taken from the same source or do not know that they can use paired/matched tests.

I investigated the application of matched/paired samples tests to correlated data and confirmed Lew's results for three or more conditions, but found the situation is more complicated when there are only two conditions. For a given number of replicates, whether Student's *t* test or a paired *t* test has higher power depends on the degree of positive correlation, which is determined by the relative sizes of shared and independent experimental errors. When shared variation is greater than individual variation correlation is high and paired *t* tests have higher power than Student's *t* test. However, Student's *t* test has higher power when correlation is low, which can occur when sources of variation that affect samples individually are larger than shared variation. Thus for a given number of replicates, taking cells from a common source may not always create enough positive correlation to make a paired *t* test more powerful than Student's *t* test. Furthermore, use of techniques that create significant shared variability between replicates could create enough positive correlation that a paired *t* test could be a better choice than Student's *t* test even if samples come from different cultures. Testing for correlation and using the results to choose a test for differences in means is difficult with small samples and increases the false positive rate (FPR) when paired *t* tests are used, so choosing a statistical test for differences in means when there are only two conditions requires careful thought about the sources of variability in the experiment.

RESULTS AND DISCUSSION

Performance of tests for multiple conditions

Lew conducted simulations by generating normally distributed data with a positive correlation of 0.5 between "control" and two "treated" groups. Positive correlation occurs when samples in an experimental replicate share variance or noise, resulting in covariance. In cell-based experiments, one of the most common situations in which correlation occurs is when all samples in an experimental replicate come from the same culture and there is significant day-to-day biological variability in the culture's behavior. When this is the case, all values for different conditions from a replicate might be either higher or lower than values from other replicates (e.g., Figure 1A, right), generating correlation which can be seen when values from replicates of a given treated condition are plotted against their corresponding control values (e.g., Figure 1B, right).

To confirm Lew's results, I used simulations, although many of findings I report would be apparent to someone with a strong statistical background without them. I first simulated data from one "control" and three "treated" conditions generated from normally distributed values with different degrees of correlation. I explored the three different analysis approaches Lew used: 1) one-way ANOVA on raw data (ANOVA); 2) one-way ANOVA on data normalized by dividing all results in a replicate by the "control" for that replicate (ANOVA_n); and 3) random block ANOVA without replication (RBANOVA). Adding the fourth group did not affect the overall results of the simulations compared with Lew's use of three groups. For my simulations, the mean of the "control" condition was set to 100. One "treated" condition had variable mean and the other two had means of 100. The SD of all four conditions was set to 10, which corresponds to 10% coefficient of variation (% CV, 100 × SD/mean), and this was maintained constant as correlation was varied. I often set the number of replicates (n) at three as this is a very common choice in cell biology experiments, but I varied n in some simulations. For ANOVA and ANOVA_n, if the ANOVA's *p* value was < 0.05 I used Dunnett's post-hoc test to compare each experimental group to control. For RBANOVA, I designated "treatment" and "replicate" as factors and if the ANOVA *p* value for "treatment" was <0.05 recorded the *p* values for the three post-hoc comparisons.

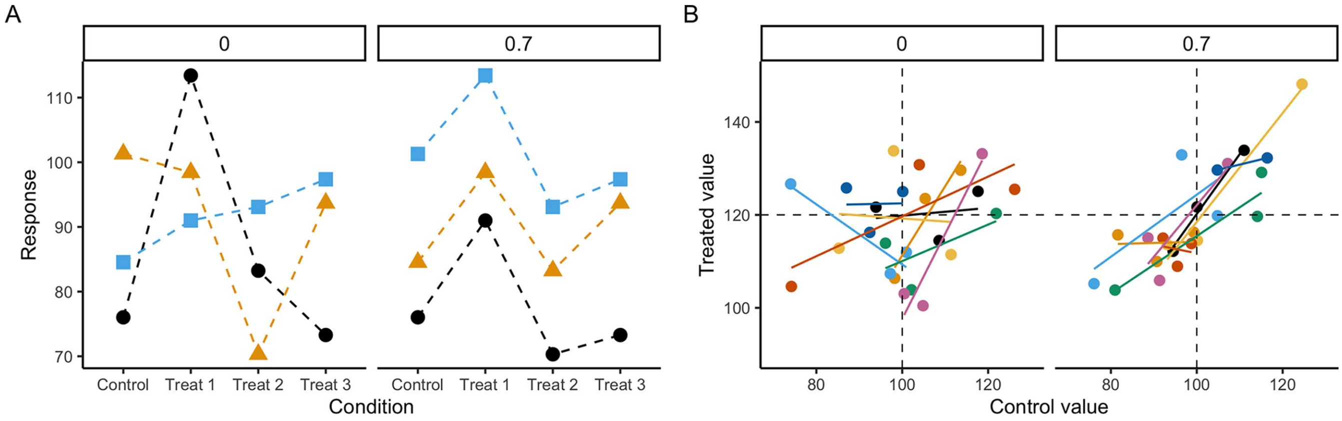

FIGURE 1: Correlation and uncorrelated samples. (A) Simulated representative data from three replicate experiments with correlation of 0 (left) or 0.7 (right). The mean of condition "treat 1" was set to 120 and all the others to 100. Variance for all was 10% CV. In this set of simulated data, random variation resulted in means of 87, 100, 82, and 88, respectively. Replicates are joined by dashed lines and share symbol shape and color. Data generated with 0 correlations are the same values as the sample with 0.7 correlations except the values in each condition were scrambled between replicates. (B) Correlation results in a linear relationship when values of a parameter obtained under different treatments are plotted for each replicate. A total of eight sets of three control and three treated replicates were simulated with means of 100 and 120, respectively, and CV of 10%. Correlation was set to 0 (left) or 0.7 (right). Values for control and treated conditions from each replicate are plotted against each other, and each of the eight sets is coded by color. Linear fits for each set of replicates are shown in the same color as the points. Some sets of replicates in the set generated with 0 correlations appear to be positively correlated.

I first examined the behavior of the tests when condition means were all set to 100, as this allows estimation of the type 1error rate, also known as the FPR (Figure 2A). Because of the way *p* values are defined, the FPR should be ~5% when the cutoff for significance is $p < 0.05$. In the absence of correlation between conditions, each test had a FPR of ~5% for the sum of the four Dunnett's comparisons. As correlation in the simulated data increased, the FPR of ANOVA decreased but the FPR for the other approaches remained constant. This is expected because the true variance of the difference in means is smaller than the F-test denominator.

I next simulated performing the tests when the mean of one "treated" condition was different than the others, increasing correlation from 0 to 0.9 at constant 10% CV (Figure 2B). In the absence of correlation, RBANOVA had power ~90% as high as ANOVA, and as correlation increased RBANOVA became increasingly more powerful than ANOVA. The power of ANOVA$_n$ was substantially lower than ANOVA when correlation was ⩽0.4 but increased with correlation. However, ANOVA$_n$ had lower power than RBANOVA over the entire range of correlation. Power of each test to detect a decrease in mean of a given size was similar to its power to detect an increase of the same size as expected (Supplemental Figure 1).

In the results presented above, effects of correlation were isolated by maintaining total variance at 10% CV. However, correlation is most likely to arise when a process like between trials variability results in addition of shared variance to other errors that affect each sample independently (Figure 2C). This is what happens when time-dependent changes in the responses of a cell culture are combined with other sources of experimental noise. Adding shared noise this way not only increases correlation it also increases total variance, which as mentioned above decreases power. I examined the effects on power of adding shared variance ranging from 0–30% CV to samples generated with independent variance ranging from 1–20% CV. Simulations with 5 and 7.5% independent CV are displayed in Figure 2D. Power of ANOVA decreases as shared noise is added because total variance increases. However, RBANOVA and ANOVA$_n$ are basically unaffected by addition of shared variance. Their power is determined almost entirely by independent sources of error. RBANOVA is more powerful than ANOVA$_n$ over the range of added shared variance.

Performance of tests for two conditions

I next considered four different ways of analyzing data from experiments with only two conditions: 1) Student's *t* test, which assumes independence and equal variance, performed on raw data, 2) Student's *t* test conducted on data normalized to "control" by dividing values from both treated and control conditions in each replicate by the value of that replicate's control, 3) Welch's test (which assumes independence but not equal variance) performed on data normalized to "control" as described for (2), and 4) paired *t* tests. I again first examined the FPR by simulating data in which both control and treated conditions have a mean of 100 (Figure 3A). The FPR for Student's *t* test performed on raw data decreased with increasing correlation, similar to what was observed with ANOVA and for analogous reasons. Student's *t* test performed on data normalized to "control" generated too many false positive *p* values < 0.05. With $n = 3$, the FPR was ~10%. Although the effect depended inversely on the number of replicates–increasing n to 10 decreased the type 1 error rate when applying Student's *t* test to normalized data to ~6% (see Supplemental data file "Data_Fig3A_B_C.csv")–this is one of the few statistical errors apart from pseudoreplication (Lazic, 2010; Eisner, 2021) that can increase the FPR, so the strategy was not investigated further. Performing Student's *t* test on data normalized to control–which appears to be a common means of handling between trials variability–is not a statistically sound practice and should be avoided. A solution is to use Welch's test, which does not assume equal variance and maintains an FPR of ~5% (unpublished data, but see Supplemental data file "Data_Fig3A_B_C.csv").

I next varied the mean of one group, keeping $n = 3$ and variance constant at 10% CV. I checked whether tests, which all assume symmetrical distributions, had similar power to detect increases and decreases in means of a given size (Figure 3B). Welch's test

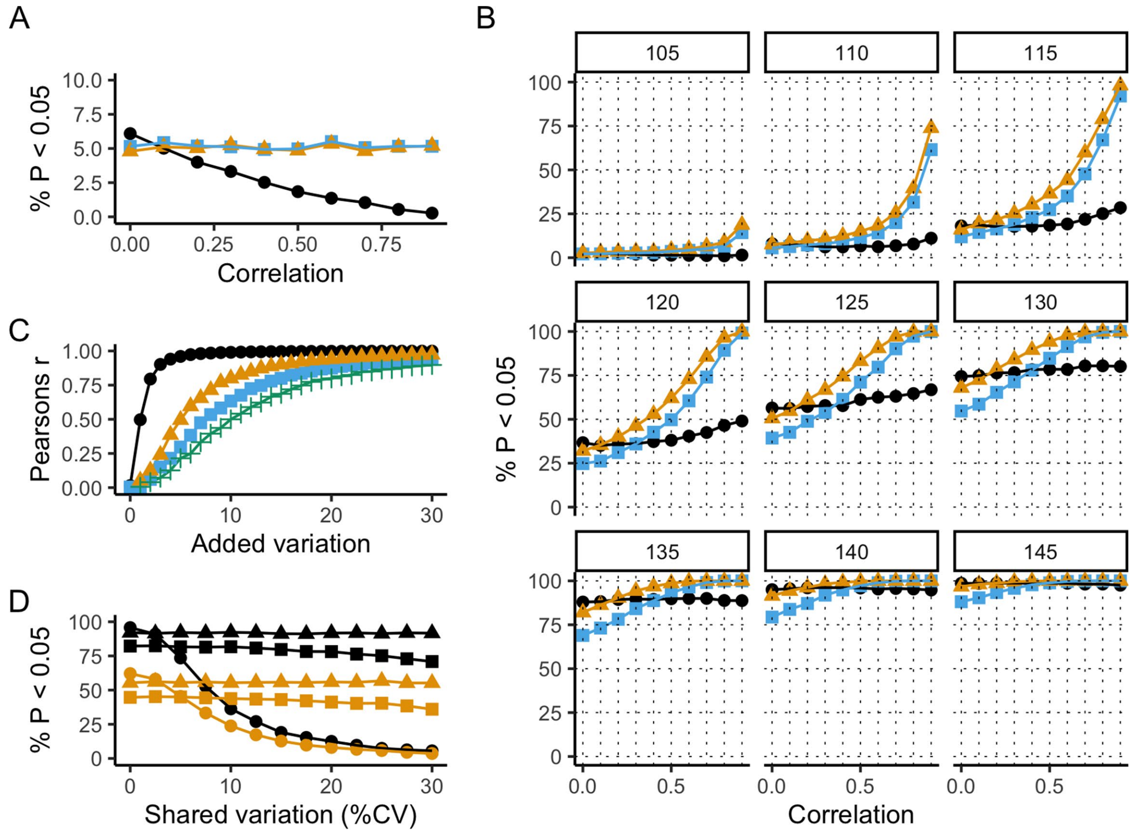

FIGURE 2: Random block ANOVA is best for testing three or more samples. (A) Assessing the FPR at different levels of correlation. The means of four groups were all set to 100 and the SDs to 10. Dunnett's test was performed if the omnibus ANOVA for treatment gave a *p* value < 0.05. The sum of *p* values <0.05 for all three comparisons Dunnett's test makes to control was counted and divided by 100. RBANOVA = orange triangles, ANOVA = black circles and ANOVA$_n$ = sky blue squares. (B) Performance of tests when correlation is generated at constant variance. Numbers over each panel indicate the value of the one mean that was different. RBANOVA = orange triangles, ANOVA = black circles and ANOVA$_n$ = sky blue squares. (C) Adding shared variance creates correlation. Numbers indicate individual variance which was 1 (bluish green circles), 5 (blue triangles), 7.5 (vermillion squares), and 10 (reddish purple crosses) % CV. Correlation is >0.5 when shared variance exceeds individual variance. (D) Performance of tests when correlation is created by adding shared variance to individual variance of 5 (black) and 7.5 (white) % CV. Triangles, RBANOVA; squares, ANOVA$_n$; circles, ANOVA.

performed on normalized data produced 20–30% fewer values <0.05 when the "treated" group's mean was higher than control compared with when it was lower (data are shown for correlation = 0 but the effect is present at all correlations, see Supplemental data file "Data_Fig3A_B_C.csv"). This occurs because normalization results on average in relatively higher variance when the effect is an increase in mean compared with a decrease (see Supplemental Figure 2). Because it seems undesirable to use a test with different power to detect increases versus decreases in mean and because it has lower power than paired *t* tests when the effect is an increase in mean, this strategy was not pursued further.

I next investigated the power of Student's *t* test and paired *t* tests over a range of correlation and numbers of replicates at constant variance (Figure 3C). When n was two replicates, paired *t* tests had lower power than Student's *t* test at all but the most extreme levels of correlation. For other numbers of replicates, though, paired *t* tests had higher power when the correlation was ≳0.5 and Student's *t* tests had higher power when the correlation was ≲0.5. When correlation was generated by combining different amounts of shared and independent variation, paired *t* tests were unaffected by added shared variance, but had ~30% lower power than Student's *t* tests when shared variation was smaller than independent variation (Figure 3D). This effect, which is the same as seen in Figure 3C, occurs because paired *t* tests are performed on differences between three pairs of observations while Student's *t* test is performed on six total observations. As a result, the critical value of the t statistic required for a *p* value > 0.05 is larger for paired *t* tests (Zimmerman, 1997). When correlation is high it is more likely the statistic for paired *t* tests will be large enough to overcome this than when correlation is low. The higher power of Student's *t* test at low correlation was first noted by Pollack and Cohen (Pollak and Cohen, 1981).

Assessing data for correlation before testing increases the conditional Type 1 error rate when paired *t* tests are used

Correlation between conditions can be assessed with Pearson's test, which estimates correlation in the data and returns a *p* value reflecting the probability that this correlation would have been observed if the null hypothesis that there is no correlation is true. I simulated running Pearson's test and performing paired *t* tests whether correlation between conditions in replicates >0.5 was detected with a *p* value < 0.05. Student's *t* test was used otherwise. With *n* = 3, Pearson's test reported correlation 5% of the time when there is no real correlation and only ~30% of the time when correlation was

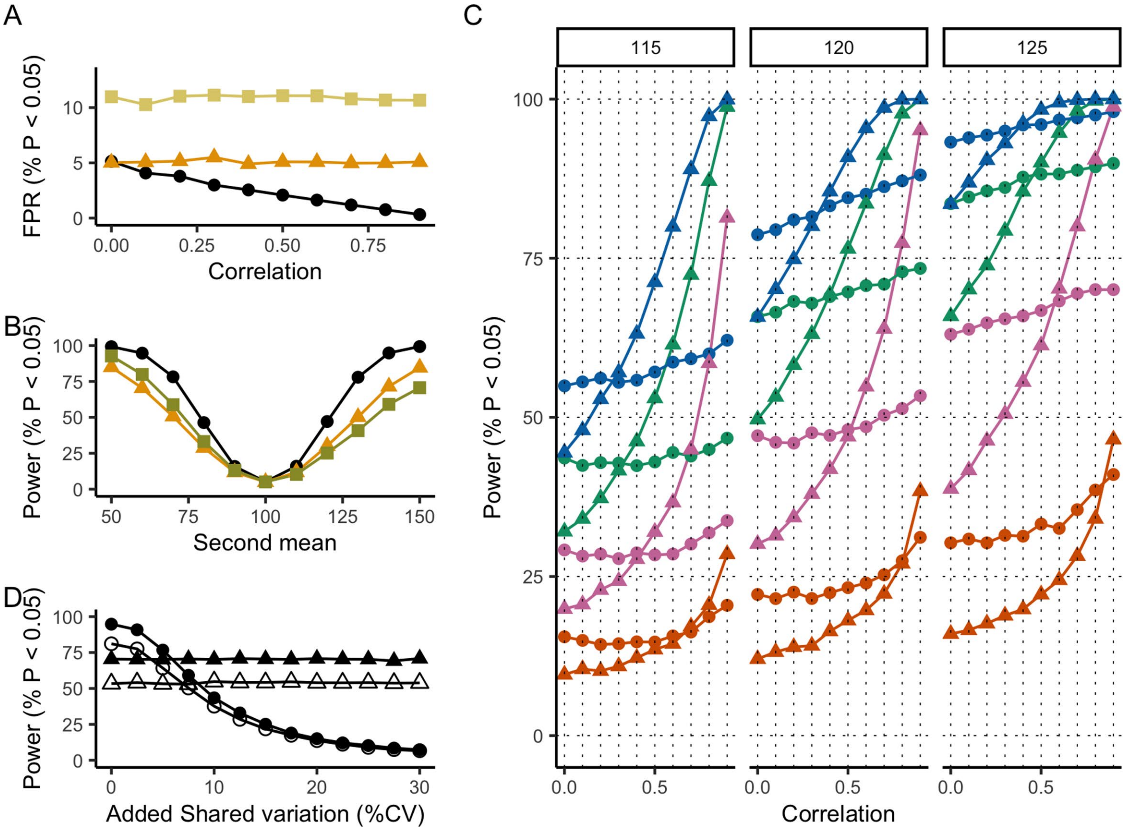

FIGURE 3: Correlation determines which two-sample test has higher power. (A) Assessing the FPR at different levels of correlation. The means of two groups were set to 100 with SDs equal to 10 and the total number of p values < 0.05 was counted and divided by 100 for Student's t test performed on raw data (black circles) or on data normalized to control (sand yellow squares), or paired t tests (orange triangles). (B) Power curves for Student's t test (black circles), Welch's test applied to normalized data (citron squares) and paired t tests (orange triangles) when data are simulated with no correlation. The mean of one group was varied from 50 to 150 while the other was held constant at 100. (C) Power of Student's t test (circles) and paired t tests (triangles) as correlation is varied but variance is held constant. The number of samples per group (n) was 2 (vermillion), 3 (reddish purple), 4 (green), or 5 (blue). (D) Power of Student's t test (circles) and paired t test (triangles) as correlation is created by adding increasing shared variance. Individual variation was 5 (black) or 7.5 (white) % CV.

0.9 (Figure 4A), a reflection of the test's low power when there are only three replicates. Because Student's t test is selected more often, the two-stage testing procedure has the higher power of Student's t test at low correlation, but for the same reason it does not have the full power of paired t tests when correlation is high (Figure 4B). Furthermore, setting the means of the two groups conditions equal to estimate the FPR of the two-stage procedure reveals a subtle but serious problem (Figure 4C). While the overall FPR remains ~5% (Figure 4B), the FPR when paired t tests are used is extremely high. With n = 3, the FPR for paired t tests is ~30% when there is no correlation and ~10% when correlation is 0.9. Because the overall FPR is 5% this may not seem like a problem, but the overall FPR is not relevant when a paired t test has been performed. The bottom line is that the results of paired t tests cannot be trusted when a two stage procedure is used. The effect on FPR of paired t tests occurs because Pearson's test has low power so correlation must be extremely high to be statistically significant. When this occurs, the denominator of the paired t test is too low, inflating the t-statistic and leading to increased FPR. Similar effects on conditional type 1 error rates have been noted previously when pretesting for normality (Hayes and Cai, 2007). Note that the only reason the overall FPR remains constant is that the FPR for Student's t test decreases with correlation (Figure 3A).

Guidance for choosing tests

RBANOVA can be used as a default when there are three or more conditions as it will always have power at least 90% as high as ANOVA and its power will be higher than ANOVA's if there is positive correlation (Figure 2). Terminology surrounding RBANOVA is confusing, and researchers may not know how to perform it, so I have included as a supplement code to do it followed by Dunnett's or Tukey's tests in free R software. However, there is no clear default when there are only two conditions (Figure 3), and data cannot be pretested for correlation without creating a risk of too many false positives when paired t tests are performed (Figure 4). It is important for reproducible science both to maximize experimental power and to ensure consistency whether a particular perturbation is tested in a two-condition experiment or as part of a larger experiment with more than two conditions. Consistency could be achieved by always choosing independent samples tests (ANOVA and Student's t test), but this would mean sacrificing statistical power when there is correlation. As is discussed further below it is likely that data in many common experiments are correlated because when possible experienced researchers generally conduct experiments with samples that come from the same source and so are matched/paired as a way to minimize effects of biological variability. All things being equal, this will tend to result in correlation although as is explained

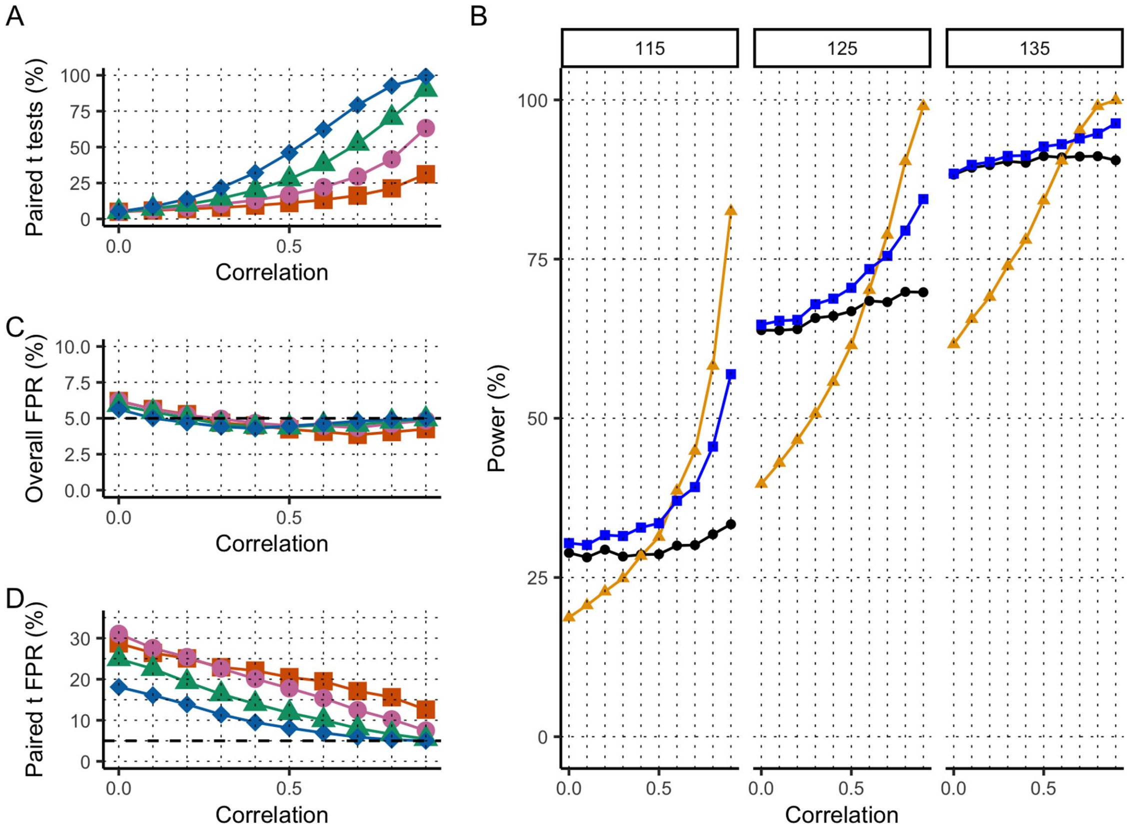

FIGURE 4: Using Pearson's test to choose a two-sample test increases the FPR when paired *t* tests are selected. (A) Percentage of cases in which paired *t* tests were performed because correlation > 0.5 was detected with $p < 0.05$. Correlation was generated at constant variance of 10% CV, and sample size (n) was: 3 (vermillion squares), 4 (reddish purple circles), 5 (green triangles), and 6 (blue diamonds). (B) Power achieved using the two-stage procedure (bright blue squares). Data for Student's *t* test (black circles) and paired *t* tests (orange triangles) are plotted for comparison. For this simulation the mean of one group was 100 and the others were varied as indicated above the sub-panels. (C) Overall FPR for the two-stage procedure. Correlation was generated with constant variance and sample size (n) was 3 (vermillion squares), 4 (reddish purple circles), 5 (green triangles), and 6 (blue purple diamonds). The dashed line is the expected FPR (5%). (D) FPR for paired *t* tests performed because correlation >0.5 was detected with $p < 0.05$. Symbols and colors are as in (C).

below it does not guarantee it. Critically, the benefits of taking cells from the same source are only realized whether a test that takes matching/pairing into account is used. Furthermore, researchers must decide beforehand how much correlation they expect to select the appropriate two-condition test, because simulations show that performing a second test whether the first does not generate a statistically significant *p* value increases the overall FPR from 5 to ~7.5%. Data not included in the test for differences in means could be used to assess correlation, so pilot experiments could be performed. However, large n is needed to reliably detect correlation, and researchers may not be able or willing to devote the necessary resources. An alternative would be to assess correlation in old data sets. Failing this, researchers will have to think about the sources of variation in the experiment and make a judgement about likely correlation. What follows is intended as a guide.

Table 1 outlines common sources of noise in cell-based experiments. As for the simulations, I have expressed errors as percent CV. Total variance in an experiment is the sum of the squared SDs of each contributing factor, and overall SD is the square root of total variance. Noise sources that affect samples individually will tend to decrease correlation while those that affect all samples in a replicate will increase correlation. Noise sources that affect samples individually can include pipetting errors, instrument noise, and averaging error, a kind of variability that occurs when measurements from individual cells are averaged. (As will be discussed below, most measurements in cell biology experiments are averages of the behavior of some number of cells). Individual variability will also include biological variability whether experiments involve comparing different cell lines or different preparations of primary cells, or whether samples in a replicate are taken from different passages or cultures of a single cell line. However, whether cells are from a single source split into groups for treatment, which most experienced researchers probably does, biological variability will contribute to shared variation. There is also a form of technical variability that can be shared when samples are processed together in batches. This is most common in techniques like gel electrophoresis and immunoblotting, PCR, ELISA and flow cytometry, but it can occur in any case where common batches of reagents or instrument settings are used to process all samples in a replicate. Additionally, when common batches of reagents are used; pipetting errors associated with batch preparation will contribute to shared rather than individual variability.

Each of these errors can vary in size. In the discussion that follows if no references are cited the value given is my estimate. Total pipetting error in most experiments is probably ~10% CV. The ISO standard for pipettors themselves is <5% CV even for very small volumes (ISO 8655), but poor technique could increase this, and overall

Source	Size (% CV)	Individual	Shared
Pipetting error	1–10% depending on number of pipetting steps.	Pipetting steps for individual samples.	When reagents used for batches of samples are made.
Instrument noise	1–3%.	Always individual.	———
Sampling error	1–2% or less when an experiment measures the output of many thousands of cells. Can be higher in microscopy.	Always individual.	———
Between trials biological variation*	Up to 50%.	When samples are different cell lines (including stable transfectants), or primary cells prepared from different animals or human subjects.	When a single cell culture or primary cell preparation is split into aliquots then treated (or not). Includes transient transfection.
Shared technical error*	Up to 100%.	———	Occurs when samples are run together using techniques with high inherent between runs variability (e.g., gel electrophoresis, immunoblotting, ELISA, flow cytometry).

*See text for references. Other values are the author's estimate.

TABLE 1: Sources of variability in cell-based experiments.

pipetting error is a function of the total number of pipetting steps in an experiment. An experiment with five pipetting steps that each had an error of 5% CV (which seems to me like a worst-case scenario) would result in total pipetting error with a variance of 5×5^2, corresponding to ~11% CV. Inherent noise for many instruments is probably only a few percent CV. Averaging error can vary a great deal in size because it depends on both the distribution of the parameter in the sample of cells and the number of cells averaged. In experiments that result in a single measurement from a sample (e.g., release of secreted factors from an aliquot of cells or total protein levels in a cell lysate) the measurement is automatically an average of the behavior of many thousands of cells, so sampling error will almost certainly be small, likely only 1–2% CV at most. However, in some kinds of experiments, particularly those that use microscopy, only 10–100 cells may be averaged (individual cells in a dish are not independent biological replicates when they are all treated together, as is too often mistaken to be the case [Lord *et al.*, 2020]) so sampling error might be large.

Apart from sampling error in microscopy, biological variation and shared technical variation are likely to be the largest sources of noise in most common experiments. The sizes of these can with proper replication be estimated separately. Molloy *et al.* (2003) examined variability in two-dimensional gel electrophoresis. Between-runs biological variation in the abundance of different proteins was ~20% CV in multiple different cultured cell lines, including bacterial cells, and was ~50% CV in several primary cell preparations. Technical variation between gels was on the order of ~20% CV. Variation between runs in immunoblotting can be > 20% CV (Butler *et al.*, 2019). In flow cytometry, variation in abundance of different immune cell types ranged from 30– 60% CV between different individuals, while shared technical variation ranged from 12–100% between different cell types (Burel *et al.*, 2017). Lynn *et al.* (1996) reported that CVs for replicate ELISA conducted within labs ranged from < 10 to > 100%.

Taken altogether, consideration of common noise sources suggests the following guidelines for choosing a two-condition test when there are only than two replicates. If each replicate of an experiment involves a single preparation of primary cells or a single culture of transformed cells split into groups and treated differently, whether with a drug or by transient transfection, biological variability will be shared by samples, so shared variation is likely to exceed sources of independent variation. This means that correlation will likely be > 0.5 so paired *t* tests will probably have higher power than Student's *t* test. If replicates are also processed together using a technique such as immunoblotting that contributes to shared variation it will further increase positive correlation and paired *t* tests will be even more likely to have higher power. But the source of cells alone is not enough to guarantee that values are correlated. If between-runs biological variability in a particular cell line is extremely small, or if between runs variability has been defined and eliminated, individual noise might dominate so correlation would likely be low and Student's *t* test would be a better choice. (Trying to eliminate between-trials biological variability so that Student's *t* test could be used with confidence that correlation was minimal would be worth striving for but might be extremely difficult). Microscopy experiments may have relatively high averaging error contributing to independent variation so it is less certain that there will be high correlation even if samples come from a common source. Enough cells should be averaged that a paired *t* test could be used with reasonable certainty it will have higher power than Student's *t* test.

In contrast, when experiments are performed on two different cultured cell lines, a single cell line that has been stably transfected with different mutant proteins multiple passages ago or primary cells isolated from two strains of animal or different individuals, Student's *t* test will likely have higher power because between-trials biological variability will contribute to independent rather than shared variation. (If variance in the data from the two condition is expected to be > 4–5-fold different, Welch's test may a better choice than Student's *t* test, as the FPR of Welch's test is unaffected by unequal variance). Again, though, the source of cells is not enough to guarantee that data are correlated. If between-trials variation in the preparations is relatively low and techniques such as gel electrophoresis, immunoblotting or flow cytometry are used there will be substantial shared technical variability and paired *t* tests could be a better choice.

I mentioned unequal variance above but have not discussed the possibility that samples taken from a common source could have unequal variance. This is because samples taken from a common source should have the same variance unless the treatment affected it. A treatment that changes variance but not mean has had an effect, so a p value < 0.05 is not so much a false positive as a misattribution of the effect's nature. I have also not considered nonparametric tests like the Wilcoxon-Mann-Whitney test for two conditions or the Wilcoxon signed rank test for three or more. Parametric tests like t tests have type 1 error rates close to 5% in many cases even if data are not normally distributed (de Winter, 2019), and nonparametric tests tend to have low power with few replicates–when n is 3, for example, tests will never generate p values < 0.05. They are best reserved for cases where it is certain that the population distribution of a parameter is very different from normal and the number of replicates is large.

A final note: as mentioned above, the combined type 1 error rate of at least one test giving a false positive p value < 0.05 when both tests are performed is elevated, so doing both tests and reporting only one that gives a significant p value would be a form of p-hacking (Head *et al.*, 2015). This does not mean that it is necessarily a bad idea to perform both tests, because if both generate statistically significant p values it would increase confidence in the result. If only one does, though, researchers would have to explain which they favor based on details of the experiment and allow readers to decide for themselves.

MATERIALS AND METHODS

Request a protocol through *Bio-protocol*.

Simulations were performed in R programing language (R Core Team, 2022), version 4.2.2 (2022-10-31). The package faux (DeBruine, 2023) was used to generate data drawn from normal distributions with different levels of correlation at constant variance. To generate data with different levels of shared and independent variation, a parent set of values with n equal to the desired number of replicates was generated using the rnorm function from the R stats package, which generates normally distributed data. The SD selected for this parent set became the shared SD' all derivative samples. Data were finished by adding vectors of normally distributed values to the parent. The means of these were set to 0 (for the control or for data from treated conditions with the same mean as the control) or to other values for data from treated conditions with different mean. The standard deviations of these vectors became the independent variance of the final data. Functions t.test, aov and cor.test functions from the R stats package and package multcomp (Hothorn *et al.*, 2008) were used for tests. R scripts for all simulations are included in a zipped supplemental file.

ACKNOWLEDGMENTS

I would like to thank Dr. Haim Bar of UConn's Department of Statistics for helpful conversations. I would also like to offer my sincerest thanks to the two reviewers of this manuscript. The time they put in to reading the original and revised submissions and the many helpful comments they made helped me clarify my thinking and resulted in a vastly better piece of work.

REFERENCES

Burel JG, *et al.* (2017). An integrated workflow to assess technical and biological variability of cell population frequencies in human peripheral blood by flow cytometry. J Immunol 198, 1748–1758.

Butler TAJ, Paul JW, Chan E-C, Smith R, Tolosa JM (2019). Misleading Westerns: Common Quantification Mistakes in Western Blot Densitometry and Proposed Corrective Measures. BioMed Res Internat 2019, e5214821.

Button KS, Ioannidis JPA, Mokrysz C, Nosek BA, Flint J, Robinson ESJ, Munafò MR (2013). Power failure: why small sample size undermines the reliability of neuroscience. Nature Rev Neurosci 14, 365–376.

DeBruine L (2023). faux: Simulation for Factorial Designs. Zenodo, http://doi.org/10.5281/zenodo.2669586.

de Winter JCF (2019). Using the Student's t-test with extremely small sample sizes. Pract Asses Res Eval 18, 1–12.

Eisner DA (2021). Pseudoreplication in physiology: More means less. J Gen Physiol 153, e202012826.

Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, Altman DG (2016). Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. Eur J Epidemiol 31, 337–350.

Hayes AF, Cai L (2007). Further evaluating the conditional decision rule for comparing two independent means. Brit J Math Stat Psychol 60, 217–244.

Head ML, Holman L, Lanfear R, Kahn AT, Jennions MD (2015). The extent and consequences of P-Hacking in science. PLOS Biology 13, e1002106.

Hothorn T, Bretz F, Westfall P (2008). Simultaneous inference in general parametric models. Biomet J 50, 346–363.

Krzywinski M, Altman N (2013). Power and sample size. Nature Methods 10, 1139–1140.

Lazic SE (2008). Why we should use simpler models if the data allow this: relevance for ANOVA designs in experimental biology. BMC Physiol 8, 16.

Lazic SE (2010). The problem of pseudoreplication in neuroscientific studies: is it affecting your analysis? BMC Neurosci 11, 5.

Lazic SE (2018). Four simple ways to increase power without increasing the sample size. Lab Anim 52, 621–629.

Lew M (2007). Good statistical practice in pharmacology Problem 2. Brit J Pharmacol 152, 299–303.

Lord SJ, Velle KB, Mullins RD, Fritz-Laylin LK (2020). SuperPlots: Communicating reproducibility and variability in cell biology. J Cell Biol 219, e202001064, 1–5.

Lynn F, Reed GF, Meade BD (1996). Collaborative study for the evaluation of enzyme-linked immunosorbent assays used to measure human antibodies to Bordetella pertussis antigens. Clin Diagn Lab Immunol 3, 689–700.

Marino MJ (2018). How often should we expect to be wrong? Statistical power, P values, and the expected prevalence of false discoveries. Biochem Pharmacol 151, 226–233.

Molloy MP, Brzezinski EE, Hang J, McDowell MT, VanBogelen RA (2003). Overcoming technical variation and biological variation in quantitative proteomics. Proteomics 3, 1912–1919.

Pollak M, Cohen J (1981). A comparison of the independent-samples t-test and the paired-samples t-test when the observations are nonnegatively correlated pairs. J Stat Plan Inference 5, 133–146.

R Core Team (2022). R: A Language and Environment for Statistical Computing, Vienna, Austria: R Foundation for Statistical Computing. https://cran.r-project.org/web/packages/report/vignettes/cite_packages.html.

Zimmerman DW (1997). Teacher's corner: a note on interpretation of the paired-samples *t* test. J Educ Behav Stat 22, 349–360.