■ *Research paper*

Urška Komatar, Tinkara Perme

Linear Mixed Models as an Alternative to Paired Samples *t*-Test in Cases of Missing Data: A Simulation Study

Abstract. In our study, we explore the equivalence and performance of a paired samples *t*-test and linear mixed models (LMMs) in statistical analysis. While the equivalence holds under the assumption of paired data with complete cases, their performance differs in scenarios with missing values. Methods were compared based on their assumptions, test size, and power. In order to test different scenarios, we generated data with varying sample sizes and different percentages of incomplete cases, to highlight the advantages of LMMs over the paired sample *t*-test in handling missing data. Another impact considered in the study is the correlation between paired measurements. Based on our conclusions, we propose a user's guide to help researchers determine when each test is most appropriate or equivalent, providing a practical framework for selecting the most suitable statistical method for their data.

Key words: linear mixed models; paired samples *t*-test; independent samples *t*-test; missing data; correlation.

Linearni mešani modeli kot alternativa parnemu testu *t* v primeru manjkajočih podatkov: simulacijska študija

Povzetek. V naši študiji smo raziskali enakovrednost in učinkovitost parnega testa *t* in linearnih mešanih modelov (LMM) v statistični analizi. Enakovrednost med metodama velja pod predpostavko parnih podatkov brez manjkajočih vrednosti, a njuno delovanje se zelo razlikuje v primeru kršenja predpostavk. Metode smo primerjali glede na njihove predpostavke, velikost testa in moč. Za testiranje različnih scenarijev smo generirali podatke z različnimi velikostmi vzorcev in različnimi odstotki nepopolnih primerov, da bi poudarili prednosti LMM v primerjavi s parnim testom *t* pri obravnavi podatkov z manjkajočimi vrednostmi. Drugi vpliv, ki smo ga upoštevali v študiji, je korelacija med parnimi meritvami. Na podlagi naših zaključkov predlagamo vodilo za uporabnike, ki bo raziskovalcem pomagalo določiti, kateri test je v določenem primeru najprimernejši ali enakovreden. S tem zagotovimo izbiro najustreznejše statistične metode za njihove podatke.

Ključne besede: linearni mešani modeli; parni test *t*; test *t* za neodvisne vzorce; manjkajoči podatki; korelacija.

■ **Infor Med Slov** 2024; 29(2): 1-9

Institucija avtorjev / Authors' institution: Faculty of Electrical Engineering, University of Ljubljana.

Kontaktna oseba / Contact person: Urška Komatar, Faculty of Electrical Engineering, Tržaška cesta 25, 1000 Ljubljana, Slovenia.
E-pošta / E-mail: urskak2000@gmail.com.

Prispelo / Received: 10. 7. 2024. Sprejeto / Accepted: 26. 11. 2024.

Introduction

In various scientific disciplines, researchers frequently encounter paired data, where two variables are dependent and correlated within each pair. This type of data is particularly prevalent in medicine, where the same individuals may be measured multiple times over a period or under different conditions, often resulting in "before and after" comparisons.[1,2]

To determine whether an intervention has had any effect on subjects – specifically, whether there is any difference between measurements taken at two different time points – the paired sample *t*-test is the most commonly used statistical method. It assesses whether the average difference between paired observations is significantly different from zero, while considering the variance.[3]

However, the paired samples *t*-test has certain limitations due to its underlying assumptions, discussed in detail in the next section. One key assumption is that the differences between paired observations are normally distributed. Many researchers have addressed violations of this assumption and have modified the original paired samples *t*-test to enhance its robustness.[3,4] In this work, we focus on missing values, specifically the situation where some subjects have only one of the two measurements. In such cases, all incomplete cases are excluded from the data, thereby reducing statistical power of a paired *t*-test.

Apart from modified versions of the paired samples *t*-test, other methods, such as a corrected *z*-test have been proposed, specifically for dealing with data including both correlated and independent measurements. The latter method was shown to perform equal or better than a paired t-test, except in cases of very high correlation ($\varrho = 0.9$).[5] A more complex approach has been introduced in the form of a permutation test, which uses a test statistic that considers both the proportion of incomplete cases and the correlation coefficient between complete pairs.[6]

Another alternative, explored in more detail in this study, are linear mixed models (LMMs). In cases of paired data without missing values, a paired *t*-test and LMMs yield the same results.[7]

We investigate the conditions under which this equivalence no longer holds and assess which method performs better by testing various percentages of incomplete cases, different levels of correlation between pairs, and varying sample sizes. In cases of larger percentages on incomplete cases, the two variables lose their dependence and become less correlated, thus resembling independent samples. For this reason, an independent samples *t*-test was also included in the analysis. The performance of each method was evaluated by test size and test power calculations.

Finally, we explore the dilemma of which approach is more advantageous: using a paired samples *t*-test, even if it means excluding a substantial portion of the data, or opting for an independent samples *t*-test, which ignores the information about correlation between paired observations. Both options were compared with a linear mixed model that uses all data and should hence have the highest power. These insights can help researchers make faster and more informed decisions when choosing the appropriate statistical test for their data.

Assumptions and Properties

Paired Samples t-Test

For the proper application of a paired t-test, data should be organised into pairs, where observations within each pair are correlated, but observations between different pairs are independent. It is assumed that the correlation is the same across all pairs. Unlike an independent sample t-test, which compares the means of two independent groups, a paired sample t-test compares the mean of differences between pairs to zero. The only assumption is that these differences are normally distributed.1 Because this test relies on the information about differences, it cannot use incomplete data cases where one of the measurements in a pair is missing.

Independent Samples *t*-Test

The data within a sample should be independent, and the data from two different samples should also be independent of each other[1]. The test compares the means of two independent groups while assuming that the data in each group are normally distributed. The basic version of an independent samples *t*-test assumes homogeneity of variance.[8] However, when this assumption is not met, Welch's *t*-test, which accounts for unequal variances, can be used instead.

In our study, in the extreme case where one value was deleted from every pair of data, we effectively created independent samples, for which an independent samples *t*-test is typically used.

Because an independent samples *t*-test retains more data in the analysis than a paired samples *t*-test (it does not exclude incomplete cases, as it does not treat the data as paired), it exhibits greater power. However, when comparing the power of the two *t*-tests on the same data, the unpaired samples *t*-test only has an advantage when correlation within pairs is low.[4] It has been shown that the power of a paired *t*-test surpasses that of an independent samples *t*-test when the correlation is at least 0.25.[4,9] It should however be explored in more detail how incomplete cases influence this statement.

Linear Mixed Models

Linear Mixed Models (LMMs) extend basic linear regression models by enabling the modelling of correlations among observations. In addition to the fixed effects familiar from linear regression, LMMs introduce random effects[2]. They share several assumptions with basic linear models: the values of the response variable should be uncorrelated (independent) from each other,[2,10] the relationship between the predictor and response variable should be linear, residuals are expected to be homogeneous, meaning their variability remains consistent across all levels of the independent variable, and the residuals should be independently normally distributed with a constant variance and the expected value of zero.[10]

In addition to those assumptions, LMMs also make assumptions about the random effect coefficients and errors, which should be independent and identically distributed. These coefficients and errors are expected to be independent and constant within groups of observations. LMMs are more flexible and can use different distributions, although normal distributions are most commonly used,[10] as they provide more flexibility in modelling.[2]

Differences in Use

Apart from paired data, in medical sciences, we often encounter non-independent data with more than two longitudinal measurements at the level of individuals as well as patches, cohorts, or measuring batches.[10] While a paired sample *t*-test is suitable for simpler experimental designs with a maximum of two longitudinal measures, LMMs are particularly useful for more complex data. They enable the analysis of hierarchical/multi-level data by allowing the inclusion of multiple fixed and/or random effects. Another significant advantage of LMMs is their ability to handle incomplete cases, thereby retaining more data in the analysis.

In conclusion, to satisfy the assumptions of both methods, we need paired data with normally distributed differences between measurements and no missing values.

Materials and Methods

Software and Tools

All simulations and graphic representations were conducted using R programming language (version 4.2.1). The following libraries were utilised: “tidyr”, “dplyr”, “ggplot2”, “knitr”, “kableExtra”, “reshape2”, “nlme”, “lattice”, “xtable”, and “MASS”. Simulations can be reproduced using a seed of 123.

Data Generation

Data were generated from a bivariate normal distribution with specified mean vector and covariance matrix parameters. For the purpose of test size calculations, both means were set to 0. For test power calculations, the second mean was set to 0.3. The variances of both variables were consistently set to 1. Covariances were set to either 0.2 or 0.8. In this specific case, because the variances are set to 1, the covariances are equivalent to the correlations between the two variables.

Test Size and Power Calculations

To assess the validity of a test, the test size should be calculated. The test size indicates the proportion of times the null hypothesis is rejected when it is actually true. Ideally, the test size should match the significance level *a*, which is typically set at 0.05. Once an appropriate test size is confirmed, we can proceed with calculating the test power, an effective measure for comparing the performance of different statistical tests. Statistical power reflects the proportion of correctly rejected null hypotheses when the null hypothesis is indeed false.

Test size and power were calculated across seven different sample sizes (10, 20, 50, 100, 150, 200, and 500). For each sample size, scenarios with six different percentages of incomplete cases (0 %, 20 %, 40 %, 60 %, 80 %, and 100 %) were examined. After data generation, a specified percentage of the data was randomly deleted. The deletion process ensured that half of the missing data were removed from the first column and the other half from the second column, with no data being deleted from both columns simultaneously, producing a desired percentage of incomplete cases.

The data generation process and the subsequent calculations for test size and power were repeated 1000 times for each scenario. Three statistical tests were applied to the generated variables: a paired *t*-test, an independent *t*-test, and a linear mixed model. The default R function was used for the *t*-tests, while linear mixed models were built using the “lme” function from the “nlme” library. To make sure incomplete cases were not removed from the analysis, we set the parameter “na.action” to “na.exclude” within the “lme” function. If a test indicated that the two variables were significantly different (p-value < 0.05), the case was assigned a value of 1; otherwise, it was assigned a value of 0. Finally, the proportion of statistically significant iterations was calculated to represent test size (for data generated from a bivariate distribution with the same means, both set to 0) or test power (for data generated from a bivariate distribution with different means, set to 0 and 0.3).

Results

Results of Test Sizes

Figures 1-3 display the test size for each method: paired samples t-test, independent samples t-test, and linear mixed model (LMM) at different sample sizes and percentages of incomplete cases across three scenarios: low correlation ($\varrho = 0.2$), moderate correlation ($\varrho = 0.5$), and high correlation ($\varrho = 0.8$). We observe that the LMM and paired samples t-test remain closer to the 0.05 line across all levels of missing data, while the test size for the independent samples t-test approaches almost 0 as correlation increases. As the percentage of incomplete cases increases, the difference between the LMM and independent samples t-test decreases, especially in scenarios with lower correlation. In all three scenarios, the graph with 100 % incomplete cases is the same for both the LMM and independent samples t-test.

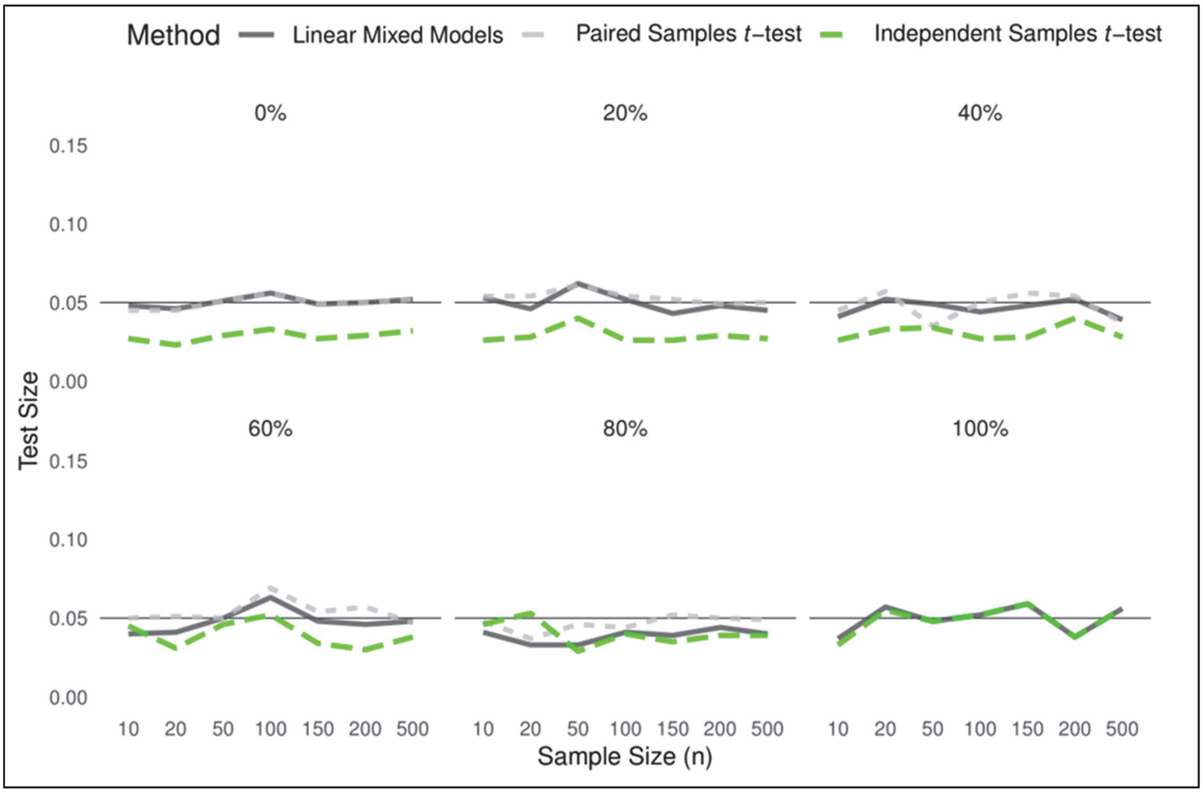

Figure 1 Test sizes across different sample sizes and percentages of incomplete cases at $\varrho = 0.2$ (thin black lines mark the significance level $a = 0.05$).

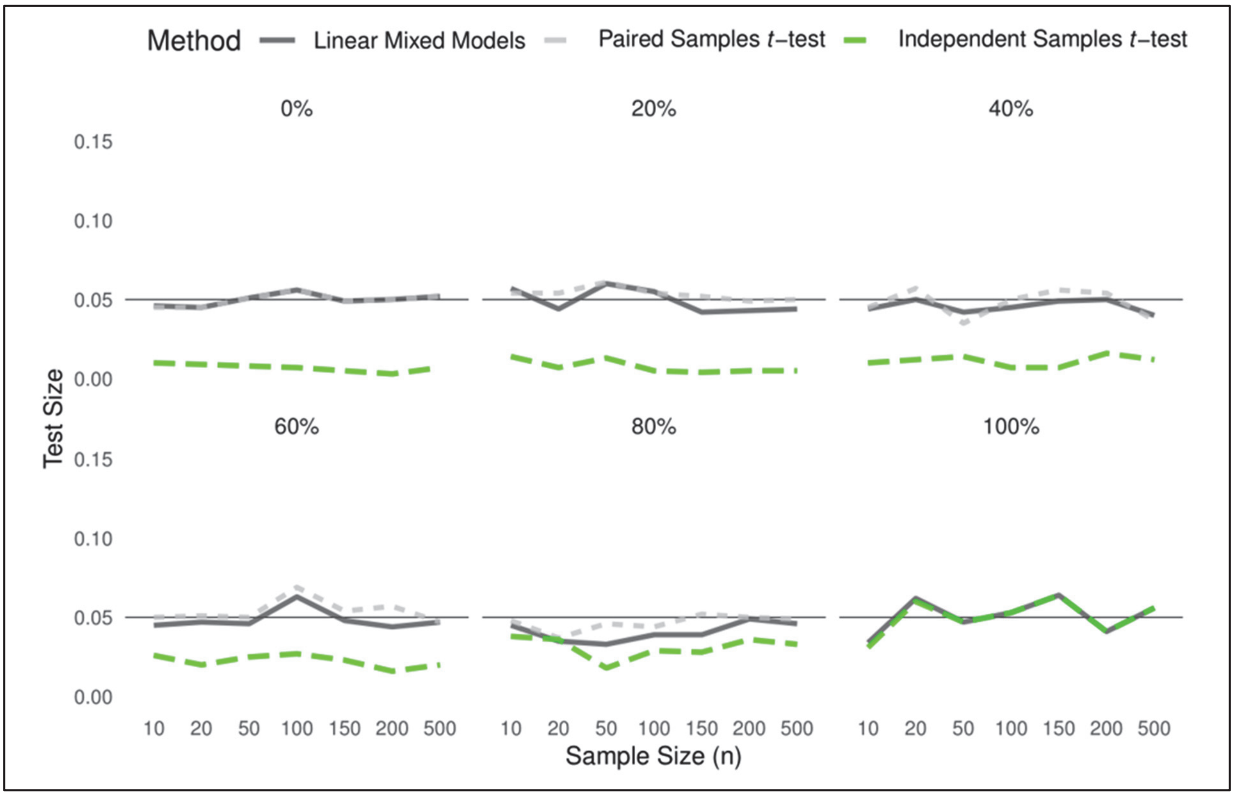

Figure 2 Test sizes across different sample sizes and percentages of incomplete cases at $\varrho = 0.5$ (thin black lines mark the significance level $a = 0.05$).

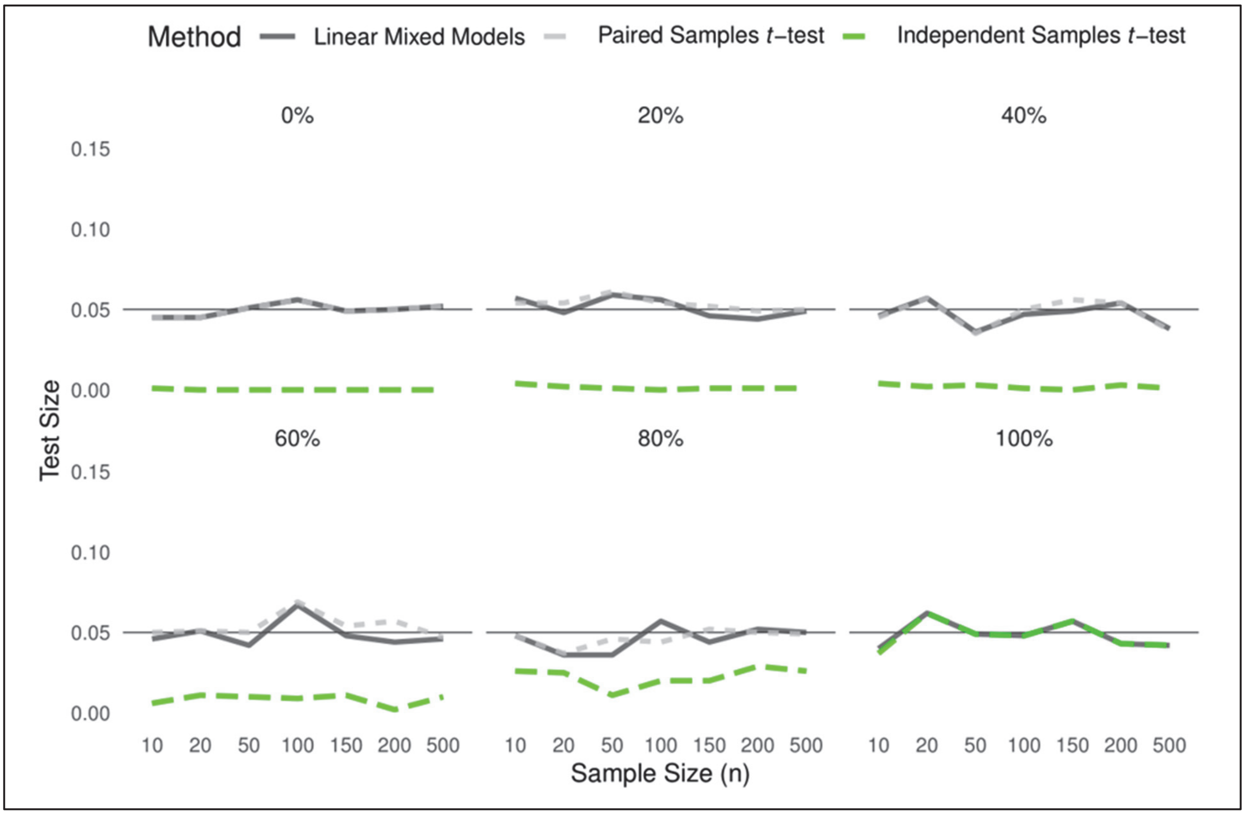

Figure 3 Test sizes across different sample sizes and percentages of incomplete cases at $\varrho = 0.8$ (thin black lines mark the significance level $a = 0.05$).

Results of Test Powers

Figures 4-6 display the power of each method across different sample sizes and percentages of incomplete cases in three scenarios: low correlation ($\varrho = 0.2$), moderate correlation ($\varrho = 0.5$), and high correlation ($\varrho = 0.8$). The independent samples *t*-test may outperform the paired samples *t*-test in terms of power, especially when the data are incomplete – specifically in cases with 40 % or more incomplete data and higher correlation. As correlation increases, the power of both the independent samples *t*-test and the paired samples *t*-test decreases. The LMM exhibits the highest power across all three scenarios, regardless of sample size or percentage of incomplete cases. The paired samples *t*-test performs as well as the LMM when the percentage of incomplete cases is low, but its power diminishes as the percentage of incomplete cases increases. Similar to the test size results, in all three scenarios the graph with 100 % incomplete cases is the same for both the LMM and the independent samples *t*-test.

The Effect of Sample Size

Regardless of the level of correlation, sample size significantly impacts the performance of all methods. This effect is more pronounced in the results for test power than in test size where the value is constantly around 0.05, regardless the sample size. To clearly illustrate the variations in power across different methods and sample sizes, the difference in means when generating data under the alternative hypothesis was intentionally set to a small value (0.3).

When one test has a clear advantage over others, it is evident that it requires a smaller sample size to achieve adequate power.

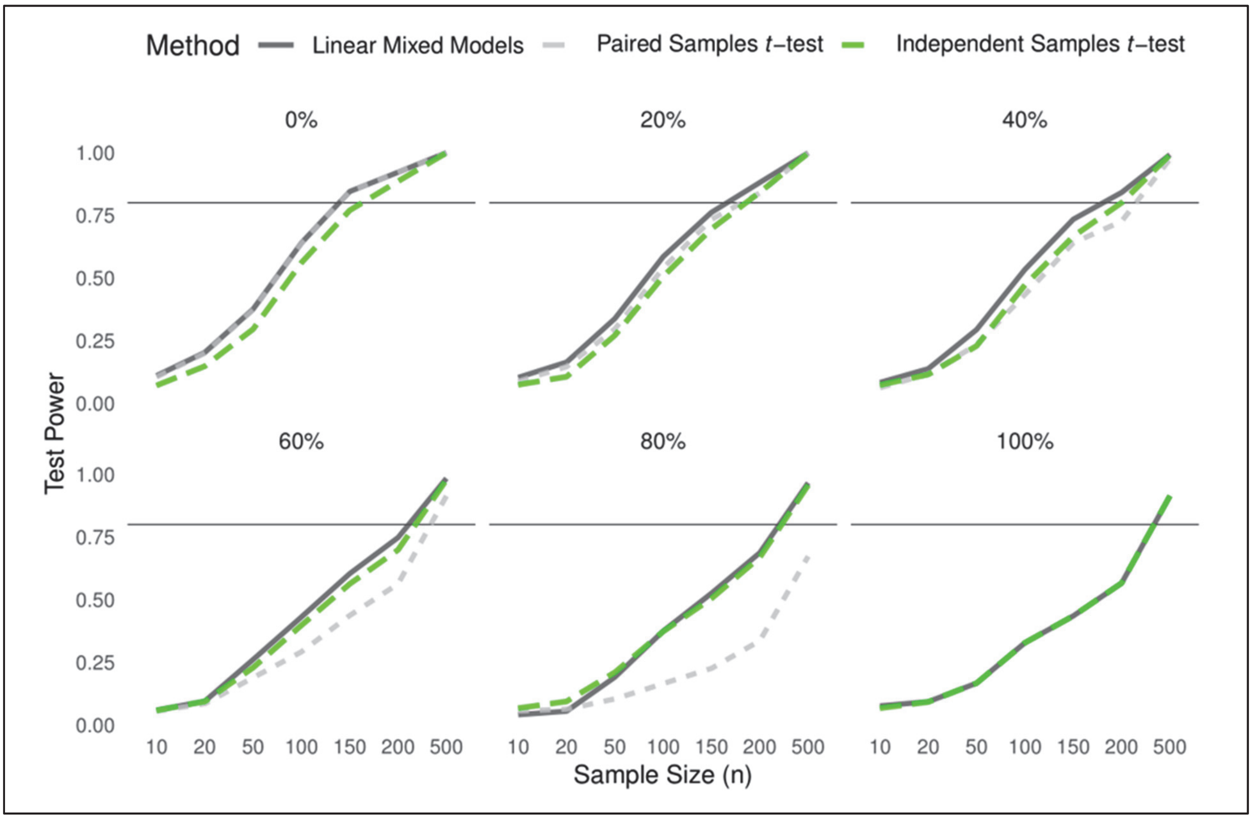

Figure 4 Test powers across different sample sizes and percentages of incomplete cases at $\varrho = 0.2$ (thin black lines mark the desired power of 0.8).

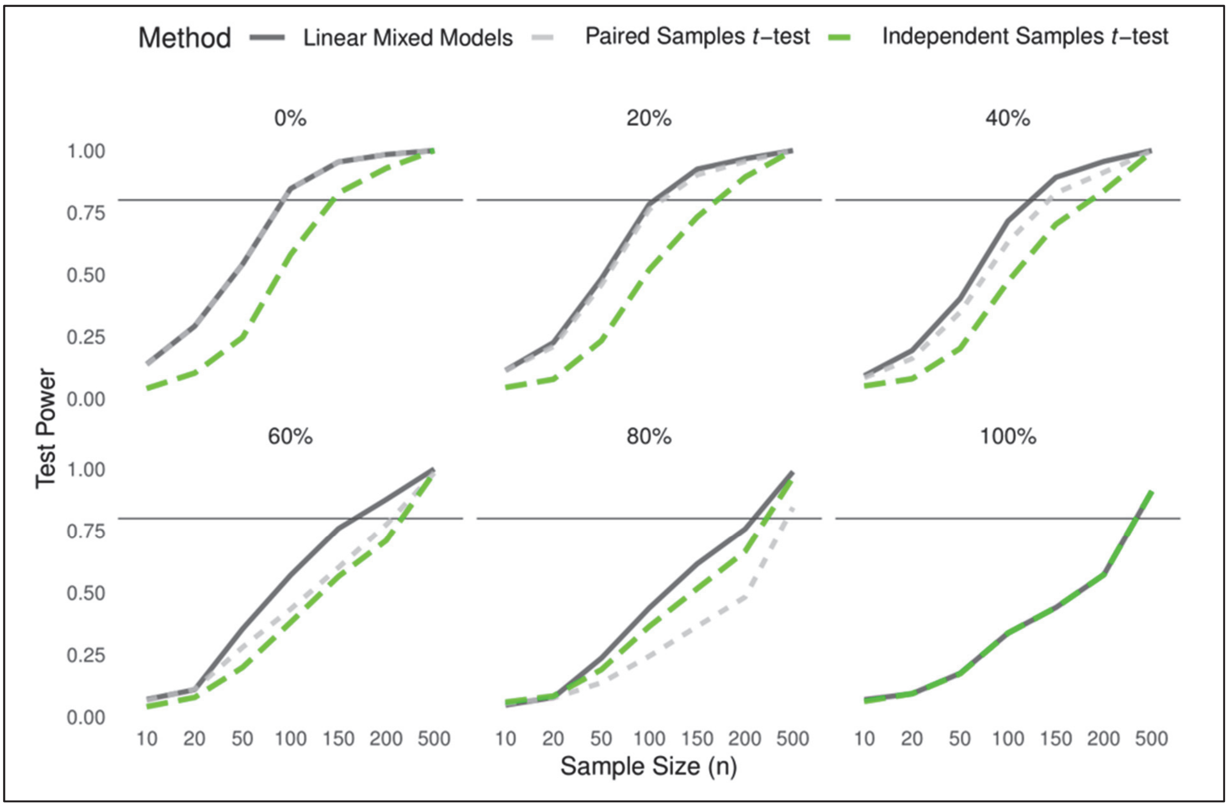

Figure 5 Test powers across different sample sizes and percentages of incomplete cases at $\varrho = 0.5$ (thin black lines mark the desired power of 0.8).

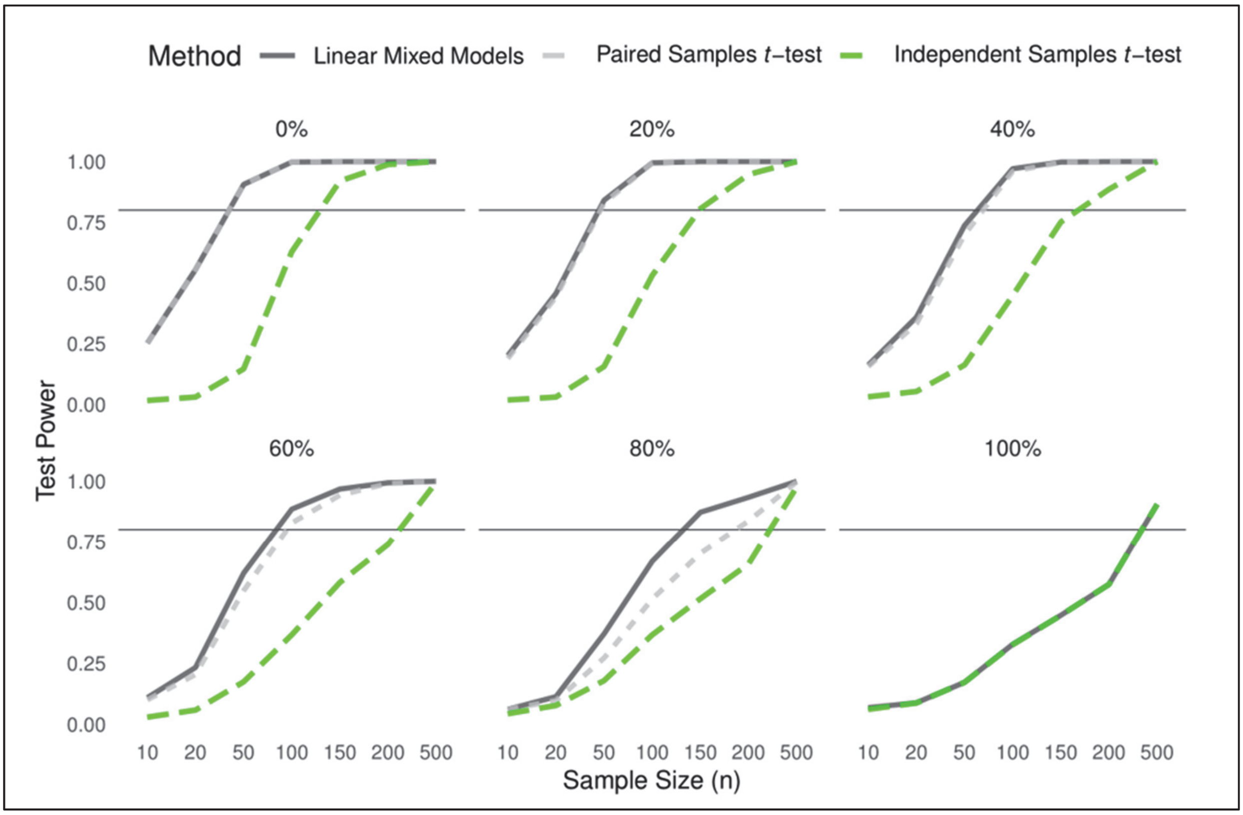

Figure 6 Test powers across different sample sizes and percentages of incomplete cases at $\varrho = 0.8$ (thin black lines mark the desired power of 0.8).

The Effect of Incomplete Cases

While paired samples *t*-tests and linear mixed models perform equivalently with complete data, the presence of incomplete cases – where one of the paired values is missing – significantly impacts the performance of the paired samples *t*-test. The paired samples *t*-test discards all rows with any missing values, leading to a substantial loss of data and, consequently, a reduction in statistical power. In contrast, linear mixed models can still utilise the remaining data, making them more powerful in the presence of incomplete cases.

This effect is evident in both low and high correlation scenarios. When correlation is low, it becomes noticeable at a lower percentage of incomplete cases. In such cases, slight differences in the performance of the two tests begin to emerge with as little as 20 % incomplete cases, with a more pronounced difference observed at 60 % incomplete cases.

In the case of high correlation, however, the power of the two tests remains comparable until the proportion of incomplete cases reaches 60 %.

A large percentage of incomplete data essentially reduces the dependence between the paired observations, resulting in less correlated data. In the extreme case of 100 % incomplete cases, the data effectively becomes two independent samples. For this reason, an independent samples *t*-test was also included in the analysis and is discussed in more detail in the section focusing on correlation.

In this extreme case, a paired samples *t*-test cannot be used, as it eliminates all rows of data, leaving no data to analyse. As a result, the paired samples *t*-test does not appear on the graph for this scenario.

The Effect of Correlation

When comparing the test sizes, we observe that the independent samples *t*-test has a test size close to zero, particularly in cases of high correlation with complete data. As the percentage of incomplete cases increases and the correlation decreases, the independent samples *t*-test becomes less conservative, resulting in a test size comparable to the other methods. A highly conservative test typically results in lower test power because it sets a stricter threshold for rejecting the null hypothesis.

It is evident that with data of lower correlation, an independent samples *t*-test outperforms a paired samples *t*-test, particularly when there is a large percentage of incomplete cases. In contrast, when the correlation is higher, the paired samples *t*-test retains greater power even with a high percentage of incomplete cases, though the difference in power between the two *t*-tests narrows.

The difference in test size and power between the two *t*-tests arises from the way variance is calculated in each test. With a higher positive correlation, the paired samples *t*-test uses a smaller variance[1]. As a result, the standard error (SE) used in the calculation of the test statistic is also smaller. This makes it easier to reject the null hypothesis compared to the independent samples *t*-test.

This is why, as demonstrated by our results, the paired samples *t*-test consistently performs better in cases of high correlation: it accounts for the correlation and therefore uses the appropriate variance. In contrast, the independent samples *t*-test overestimates the variance, and hence has lower size and power.

In the extreme scenario of 100 % incomplete cases, resulting in two independent samples, both the independent samples *t*-test and linear mixed models produce equivalent results.

Discussion

In this section, we summarise how different factors influence the performance of linear mixed models, paired samples *t*-tests, and independent samples *t*-tests, differentiating between four scenarios.

When dealing with paired data, it is advisable to first check the correlation between the measurements from the first and second assessments. Additionally, it is important to verify that there are enough cases with both measurements available. Based on the level of correlation and the percentage of incomplete cases, we identify four distinct scenarios:

- Low correlation, few incomplete cases: In this scenario, a larger sample size is required to achieve adequate test power compared to situations with higher correlation. Although all three tests yield similar results, the linear mixed model and paired samples t-test tend to perform better. Either of those can be chosen.
- Low correlation, many incomplete cases: Using a paired samples t-test is no longer appropriate because of the limited amount of available data. A linear mixed model should be chosen instead. If the correlation is very low and there is a large percentage of incomplete cases, an independent samples t-test will perform equivalently.
- High correlation, few incomplete cases: Either the paired samples t-test or the linear mixed model can be used effectively. However, we strongly

recommend against using the independent samples t-test.

- High correlation, many incomplete cases: Use either the paired samples t-test or the linear mixed model. We still do not recommend using the independent samples t-test in this scenario.

While linear mixed models offer a universal solution for data including both correlated and uncorrelated samples (incomplete cases), our study could be expanded to include other methods, previously shown to perform well on this type of data. It could be explored in more detail how, for example, a corrected *z*-test[5] and a permutation approach[6] perform in comparison to not only a paired samples *t*-test but also the LMMs. While a corrected *z*-test outperformed a paired samples *t*-test in most scenarios in the original simulation study, there was a case where a paired samples *t*-test exhibited greater power, which in contrast, never happened in our comparison to LMMs. Equivalently to LMMs, in the case of completely independent samples, a corrected *z*-test maintained power equal to that of an independent samples *t*-test.[5]

Conclusion

Our study highlights the importance of selecting the appropriate statistical test for paired data based on the specific characteristics of the dataset, particularly the level of correlation and the presence of incomplete cases. We found that while paired samples *t*-tests and linear mixed models perform similarly with complete data, the paired samples *t*-test loses its effectiveness when faced with lower correlations or high percentages of incomplete cases. In contrast, linear mixed models maintain robustness under these conditions, making them a preferable choice in many scenarios.

The independent samples *t*-test, although less powerful in cases of high correlation, can outperform the paired samples *t*-test when correlation is low, particularly in the presence of missing data. These findings offer useful guidance for researchers in selecting the most suitable analytical approach based on their specific data characteristics.

Acknowledgment

This article would not have been possible without the guidance and help of Professor Maja Pohar Perme, who contributed her valuable assistance in the preparation and completion of this study.

References

1. Xu M, Fralick D, Zheng JZ, Wang B, Tu XM, Feng C: The differences and similarities between two-sample *t*-test and paired *t*-test. *Shanghai Arch Psychiatry* 2017; 29(3): 184-188.
2. Jiming J: *Linear and generalized linear mixed models and their applications.* New York 2007: Springer. https://doi.org/10.1007/978-0-387-47946-0
3. Kim H, Park C, Wang M: Paired *t*-test based on robustified statistics. In: *Fall Conference, Korean Institute of Industrial Engineers, Seoul, Korea, 2018*; 2347-2353. https://www.researchgate.net/publication/329024164_Paired_t-test_based_on_robustified_statistics
4. Fradette K, Keselman HJ, Lix L, Algina J, Wilcox RR. Conventional and robust paired and independent-samples *t* tests: type I error and power rates. *J Mod Appl Stat Methods* 2003; 2(2): 481-496. https://doi.org/10.22237/jmasm/1067646120
5. Looney SW, Jones PW: A method for comparing two normal means using combined samples of correlated and uncorrelated data. *Statist Med* 2003; 22(9): 1601-1610. https://doi.org/10.1002/sim.1514
6. Einsporn RL, Habtzghi D: Combining paired and two-sample data using a permutation test. *J Data Sci* 2013; 11(4): 767-779. https://doi.org/10.6339/JDS.2013.11(4).1164
7. Brauer M, Curtin JJ: Linear mixed-effects models and the analysis of nonindependent data: a unified framework to analyze categorical and continuous independent variables that vary within-subjects and/or within-items. *Psychol Methods* 2018; 23(3): 389-411. https://doi.org/10.1037/met0000159
8. Kim TK, Park JH: More about the basic assumptions of *t*-test: normality and sample size. *Korean J Anesthesiol* 2019; 72(4): 331-335. https://doi.org/10.4097/kja.d.18.00292
9. Vonesh EF: Efficiency of repeated measures designs versus completely randomized designs based on multiple comparisons. *Commun Stat-Theor M* 1983; 12(3): 289-301. https://doi.org/10.1080/03610928308828458
10. Schielzeth H, Dingemanse NJ, Nakagawa S. et al.: Robustness of linear mixed-effects models to violations of distributional assumptions. *Methods Ecol Evol* 2020; 11(9): 1141-1152. https://doi.org/10.1111/2041-210X.13434