E-Jurnal Matematika Vol. 15(4), November 2026, pp. 262-276 ISSN: 2303-1751
DOI: https://doi.org/10.24843/MTK.2026.v15.i04.p530

COMPARATIVE PERFORMANCE OF PARAMETRIC AND NONPARAMETRIC TESTS UNDER VARYING SAMPLE SIZES AND DISTRIBUTIONS

Alfred Ayo Ayenigba [1§], James Serumun Ivande[2], Wasilat Odunola Yekeen[3]

[1]Department of Mathematical Sciences, Ajayi Crowther University, Oyo, Nigeria
[Email: aa.ayenigba@acu.edu.ng]
[2]Department of Statistics, Joseph Sarwuan Tarka University, Makurdi, Nigeria.
[Email: ivande.james@uam.edu.ng]
[3]Department of Mathematical Sciences, Ajayi Crowther University, Oyo, Nigeria
[Email: yekeenwasilat422@gmail.com]
[§]*Corresponding Author*

ABSTRACT

The choice between parametric and nonparametric tests remains a persistent methodological question in educational research. A Monte Carlo study compared the t test with the Mann–Whitney U test, and one-way ANOVA with the Kruskal–Wallis H test, across five distributions and six sample sizes (n = 10 to 200 per group), with 20,000 replications per condition at α = .05. A second experiment added unequal group sizes, unequal variances and a zero-inflated, grade-like distribution, together with Welch's t test. With balanced groups and equal variances, both families kept Type I error near the nominal level; parametric tests held a small power advantage for symmetric, light-tailed data, whereas nonparametric tests were substantially more powerful for skewed and heavy-tailed data. With unequal sizes and variances, Student's t test became unreliable (Type I error from .002 to .236), Welch's test stayed between .047 and .066, and the Mann–Whitney test was not a dependable substitute. In 395 student grades, a zero-inflated outcome, the school-support comparison gave p = .100 (Student), .020 (Welch) and .003 (Mann–Whitney). Test choice should therefore follow sample size, group balance, variance and distributional shape.

Keywords: *parametric test, nonparametric test, Monte Carlo simulation, Type I error, statistical power, Welch's t test*

1. INTRODUCTION

Statistical hypothesis testing occupies a central position in empirical research across the social, biological, and educational sciences (Ayenigba & Ajao, 2026). Whenever a researcher seeks to determine whether an observed difference between groups reflects a genuine population effect rather than sampling variability, a formal statistical test is required. Two broad families of tests are available for this purpose: parametric procedures, such as the independent-samples *t* test and one-way analysis of variance (ANOVA; Fisher, 1925), which assume the sampled data arise from a specific distributional family, most commonly the normal distribution; and nonparametric, rank-based procedures, such as the Mann–Whitney *U* test and the Kruskal–Wallis *H* test (Kruskal & Wallis, 1952; Mann & Whitney, 1947), which make comparatively weaker distributional assumptions. Parametric tests are, under ideal conditions, more statistically efficient than their nonparametric counterparts (Chicco et al., 2025), but real data, particularly in educational and social-science settings, frequently depart from normality through skewness, heavy tails, or bounded ranges.

This raises an enduring and practically consequential question: under what combinations of sample size and distributional shape can a parametric test still be trusted to deliver a valid and efficient decision, and when does a nonparametric alternative become the more defensible choice? An incorrect choice carries two distinct costs. First, if a parametric test is applied when its assumptions are seriously

violated, the nominal significance level may no longer reflect the true Type I error rate, undermining the validity of the reported conclusion. Second, even where Type I error remains approximately controlled, a poorly matched test can sacrifice statistical power, increasing the risk of a Type II error and causing a genuine effect to go undetected. Vrbin (2022) notes that this decision is one of the most frequently encountered, and frequently mishandled, analytic choices in applied research, particularly when comparing samples with *t* tests, the Mann–Whitney *U* test, ANOVA, and the Kruskal–Wallis test.

Recent simulation evidence continues to demonstrate that the comparative performance of these two test families is neither fixed nor intuitive. Şimşek (2023), simulating Likert-type and ordinal educational data, found that the Wilcoxon–Mann–Whitney test was superior to Welch's *t* and Student's *t* tests, particularly for small and unequal group sizes, although the parametric *t* test achieved comparable power once sample sizes reached 200 or more under balanced conditions. Working with a three-group design, Clark et al. (2023) compared Monte Carlo Kruskal–Wallis and ANOVA power studies across a wide range of simulated and real medical datasets and found that Monte Carlo Kruskal–Wallis power estimation might perform better than its ANOVA equivalent in many settings and remained usable when ANOVA assumptions were invalid. Most recently, Oladugba and Ikebuife (2026), simulating a randomized block design under normal, Cauchy and logistic errors, found a randomization test to be robust in controlling Type I error and, for most conditions, more powerful than the classical *F* test and the Friedman test. These recent findings motivate a systematic, high-precision re-examination that spans both the two-group and three-group comparison designs jointly, across a graded range of sample sizes and distributional shapes, and that is validated against genuine, nonsimulated educational data.

The present study addresses this need directly. Using a Monte Carlo simulation with 20,000 replications per condition, a substantially larger number than most classical studies in this literature, we compare the independent-samples *t* test with the Mann–Whitney *U* test, and one-way ANOVA with the Kruskal–Wallis *H* test, across five population distributions and six sample sizes ranging from 10 to 200 observations per group. Because the real data prove to be unbalanced, heteroscedastic and zero-inflated, a second experiment then adds unequal group sizes, unequal variances and a zero-inflated distribution, and brings Welch's *t* test into the comparison. The simulation-derived recommendations are then validated against a real, publicly documented dataset of secondary-school academic performance (Cortez & Silva, 2008), comparing final grades by sex, receipt of school educational support, and maternal education level. The specific objectives of the study are to: (i) compare the empirical statistical power of the parametric and nonparametric tests across distributions and sample sizes; (ii) compare their empirical Type I error rates; (iii) examine how distributional shape moderates their relative performance; (iv) examine how sample size moderates this performance; (v) assess the effect of unequal group sizes, unequal variances and zero-inflation on the *t* test, Welch's *t* test and the Mann–Whitney *U* test; and (vi) validate the simulation findings against real educational data.

2. LITERATURE REVIEW

Parametric tests rest on the assumption that the sampled observations arise from a population distribution belonging to a specified parametric family, fully characterized by a finite set of parameters, typically the population mean and variance. The independent-samples *t* test evaluates whether two independent population means differ, referring the standardized mean difference to a *t* distribution with degrees of freedom fixed by the combined sample size; one-way ANOVA extends this logic to three or more groups via the *F* ratio of between-group to within-group variance. Both procedures are, in principle, the most powerful available tests for detecting a genuine location shift when their assumptions of normality and homogeneity of variance are satisfied (Chicco et al., 2025).

Nonparametric, rank-based tests were developed specifically to relax these distributional requirements. The Mann–Whitney *U* test ranks all observations from two combined groups and evaluates whether one population is stochastically larger than the other, without assuming a specific parametric form (Mann & Whitney, 1947; Wilcoxon, 1945); the Kruskal–Wallis *H* test extends this rank-based logic to three or more groups as the nonparametric analog of ANOVA. Because these tests operate

on ranks rather than raw values, they are considerably less sensitive to outliers, skewness, and heavy tails, and remain valid across a broader class of population distributions, at some potential cost in efficiency when the parametric assumptions genuinely hold (Chicco et al., 2025; Vrbin, 2022).

A substantial and continuing body of simulation evidence has examined how this trade-off manifests in finite samples. Şimşek (2023) conducted a Monte Carlo comparison of the Wilcoxon–Mann–Whitney, Welch's *t*, and Student's *t* tests for Likert-type educational data, manipulating sample size, group-size ratio, and effect size; the nonparametric test was found to be superior for small and unequal-sized samples, while the *t* test only matched its power once samples reached approximately 200 per group under balanced conditions. In the three-group setting, Clark et al. (2023) compared Bernstein-fit-based Kruskal–Wallis power estimation with Monte Carlo and analytical ANOVA power studies across simulated Pearson-type populations and three real medical datasets, concluding that Monte Carlo Kruskal–Wallis power estimation might perform better than the ANOVA equivalent in many settings and remained usable when ANOVA assumptions were invalid, although the ANOVA equivalent was closer to measured power in one of three medical datasets. Nwobi and Akanno (2021) reached a closely related conclusion in a large factorial simulation crossing normal and nonnormal distributions with equal and unequal variances and group sizes: the Kruskal–Wallis test outperformed the *F* test in 85 of 92 nonnormal cases, while the power advantage of the *F* test under normality was not statistically significant in any of the 92 corresponding normal-distribution cases. Concerning the two-sample case specifically, Delacre et al. (2017) argued that Welch's *t* test should be used by default instead of Student's *t* test, because the pooled-variance test can misstate its Type I error when group sizes and variances differ, a condition that the extended experiment of the present study examines directly.

Beyond the two-sample and k-sample independent-group designs considered directly in this study, related recent work confirms that rank-based and randomization procedures continue to offer robustness advantages in more complex designs. Oladugba and Ikebuife (2026), simulating a randomized complete block design under normal, Cauchy, and logistic errors, found a randomization test to be robust in controlling Type I error and, for most conditions, more powerful than the ANOVA *F* test and the Friedman test, with slight limitations under Cauchy errors in some settings. Kamath et al. (2025) examined a closely related methodological question, comparing thirteen normality tests, including the Shapiro–Wilk test used in the present study as a diagnostic tool, under Monte Carlo-simulated skewness and kurtosis, and found that test performance depended jointly on sample size and the specific combination of skewness and kurtosis present, echoing the central finding of the present study that no single decision rule is adequate in isolation.

Within statistics and data-science education specifically, Asare (2026) applied the Kruskal–Wallis *H* test, in preference to one-way ANOVA, to compare the statistics achievement of information technology education students across four class cohorts after Shapiro–Wilk and Kolmogorov-Smirnov tests indicated significant nonnormality in every cohort ($p < .05$), reporting a large and statistically significant effect ($H = 20.68, \varepsilon^2 = .491$) together with Dunn's post hoc pairwise comparisons. This application, on genuine educational assessment data structurally similar to the dataset employed for validation in the present study, illustrates directly why the joint consideration of sample size, distributional shape, and test selection is not merely a simulation-based abstraction but a recurring, practically consequential decision in contemporary educational research.

Taken together, this recent literature converges on three points that motivate the present study. First, both test families generally maintain reasonable control of the Type I error rate, so the more decision-relevant criterion is statistical power (Clark et al., 2023; Nwobi & Akanno, 2021). Second, the power advantage of nonparametric tests under nonnormality is not uniform but depends jointly on sample size, group-size balance, and the specific form of nonnormality present (Kamath et al., 2025; Şimşek, 2023). Third, very few of these studies combine a systematically graded simulation, spanning both two-group and three-group designs, with validation on a real, contemporary educational dataset within a single, internally consistent study, a gap the present study is designed to address.

3. RESEARCH METHOD

This study combines a Monte Carlo simulation with a real-data validation component. The simulation constitutes a factorial computational design crossing five population distributions with six sample sizes, under both a null condition (used to estimate Type I error) and an alternative condition (used to estimate power), for two independent test-comparison designs: a two-group design (*t* test vs. Mann–Whitney *U* test) and a three-group design (one-way ANOVA vs. Kruskal–Wallis *H* test). This yields $5 \times 6 \times 2 \times 2 = 120$ simulation cells, each evaluated with 20,000 independent Monte Carlo replications, a considerably larger number than in most classical studies in this literature (cf. Nwobi & Akanno, 2021, *N* = 1,000 per cell), which narrows the Monte Carlo margin of error surrounding the reported rates. A second, extended experiment, described below, adds unbalanced allocation, unequal variances and zero-inflated data to the two-group design.

Five distributions were simulated, each standardized to a population mean of 0 and standard deviation of 1 prior to any location shift: the Normal distribution, $N(0,1)$; the Uniform distribution, $U(-\sqrt{3},\sqrt{3})$, symmetric and light-tailed; the Exponential distribution, strongly right-skewed; the Lognormal distribution ($\sigma = 0.75$ on the log scale), more severely right-skewed than the Exponential; and the Student *t* distribution with three degrees of freedom, symmetric but strongly heavy-tailed. Six sample sizes per group were examined (*n* = 10, 15, 30, 50, 100, 200), spanning the small, medium, and large sample regimes typical of undergraduate, postgraduate, and larger-scale research. For the two-group design, a location shift of Cohen's *d* = 0.5 (a conventional medium effect size; Cohen, 1988) was applied to the second group under the power condition; for the three-group design, a graded shift of 0, d, and 2d was applied across the three groups, respectively.

The extended experiment concerned the two-group design and compared Student's *t* test, Welch's *t* test and the Mann–Whitney *U* test. Total sample sizes of $N = 64, 128$ and 256 were divided between the groups in the ratios $1{:}\,1, 1{:}\,3$ and $1{:}\,7$, the last being close to the ratio of the school-support comparison in the real data ($51\ vs.\ 344$). The standard deviation of one group was set at twice that of the other, either in the smaller group or in the larger group, alongside an equal-variance reference; with equal group sizes only one unequal-variance condition arises. Two parent distributions were used. The first was the Normal distribution. The second was a zero-inflated distribution calibrated on the real final grades: a point mass of 9.6% at zero (38 of 395 grades) combined with a Normal component with mean 11.52 and standard deviation 3.23, rounded to integers between 1 and 20. Both were standardized to mean 0 and standard deviation 1 and then scaled about the common mean, so that the two population means were equal under the null hypothesis. Under the alternative, the second group was shifted by 0.5 standard deviations of the first group. This gives $2 \times 3 \times 8 = 48$ cells, each run under the null and the alternative hypothesis with 20,000 replications. Power is comparable across tests within a cell but not across cells that differ in variance pattern.

Real-data validation used the Student Performance (Mathematics) dataset assembled by Cortez and Silva (2008) and distributed through the UCI Machine Learning Repository (Cortez, 2014), comprising 395 students from two Portuguese secondary schools. The final grade contains 38 zeros (9.6%), a feature that the extended experiment reproduces. The final grade (G3, 0–20 scale) served as the outcome variable. Three grouping variables instantiated the same two-group and three-group designs examined in the simulation: sex (male vs. female), receipt of school educational support, schoolsup (yes vs. no; a markedly unbalanced comparison, 51 vs. 344 students), and maternal education level, Medu, recoded into three ordered categories (Low, Medium, High).

Five inferential tests and one diagnostic test were used throughout. The independent-samples (Student's) *t* test evaluates the null hypothesis of equal population means using the pooled-variance statistic in Equation 1, referred to a *t* distribution with $n_1 + n_2 - 2$ degrees of freedom. Welch's *t* test replaces the pooled variance by the two separate sample variances, as in Equation 3, and uses the Welch–Satterthwaite approximation for the degrees of freedom, so that it does not require equal variances (Delacre et al., 2017). The Mann–Whitney *U* test ranks the pooled observations and evaluates the resulting rank-sum statistic against its known sampling distribution. One-way ANOVA evaluates the *F* ratio of between- to within-group variance against an *F*

distribution with $(k-1, N-k)$ degrees of freedom. The Kruskal–Wallis H test performs the analogous rank-based comparison for k independent groups, with H approximately chi-square distributed with $(k-1)$ degrees of freedom under the null hypothesis (Chicco et al., 2025). The Shapiro–Wilk test (Shapiro & Wilk, 1965) was used descriptively to characterize the distributional shape of both simulated and real samples, consistent with recent evidence that its performance, like that of other normality tests, depends jointly on sample size, skewness, and kurtosis (Kamath et al., 2025), and was not used as a pre-test decision rule for choosing between test families.

The independent-samples t test statistic is given by:

$$t = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \quad (1)$$

and the Kruskal–Wallis H statistic, used for the three-group nonparametric comparison, is given by:

$$H = \frac{12}{N(N+1)}\sum_{j=1}^{k}\frac{R_j^2}{n_j} - 3(N+1) \quad (2)$$

Welch's t statistic, which does not assume equal variances, is given by:

$$t_W = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}} \quad (3)$$

All computations were performed in Python 3.11 using NumPy 2.4 for vectorized pseudo-random number generation and SciPy 1.17 for test statistics, with a fixed random seed ensuring exact reproducibility. For each of the 120 cells, two samples (or three, for the k-sample design) were drawn from the specified distribution, both the parametric and nonparametric test were applied to the same generated data, and the empirical Type I error rate or power was computed as the proportion of 20,000 replications for which the two-sided p value fell below α = .05. The Monte Carlo standard error of a simulated proportion, $SE(\hat{p}) = \sqrt{[\hat{p}(1-\hat{p})/R]}$ with R = 20,000, does not exceed approximately .0035 for any reported rate, and was used to judge whether an observed difference between the parametric and nonparametric rate in a given cell exceeded simulation noise (approximately two standard errors). The extended experiment used the same software, replication count and two-sided decision rule, with an independent random stream for every cell spawned from the master seed. For the real-data component, Welch's t test and the Mann–Whitney U test were applied to the two-group comparisons (with Student's t test also reported for the school-support comparison, where group sizes and variances differ), and ANOVA and the Kruskal–Wallis H test to the three-group comparison, together with Cohen's d and the rank-biserial correlation for two-group comparisons, and eta-squared and epsilon-squared for the three-group comparison.

4. RESULTS AND DISCUSSION

Table 1 reports the empirical Type I error rates for the two-group design (t test vs. Mann–Whitney U test; Figure 1) and Table 2 for the three-group design (ANOVA vs. Kruskal–Wallis H test; Figure 2), each based on 20,000 replications per cell.

Table 1. Empirical Type I Error Rates for the Two-Sample Design

Distribution	Test	*n* = 10	*n* = 15	*n* = 30	*n* = 50	*n* =100	*n* =200
Normal	*t* test	.050	.052	.051	.050	.049	.054
	Mann–Whitney *U*	.044	.047	.049	.051	.048	.052
Uniform	*t* test	.050	.051	.050	.049	.052	.052
	Mann–Whitney *U*	.044	.046	.048	.049	.052	.051
Exponential	*t* test	.045	.045	.048	.047	.052	.048
	Mann–Whitney *U*	.045	.044	.047	.047	.050	.048
Lognormal	*t* test	.041	.044	.045	.048	.049	.051
	Mann–Whitney *U*	.043	.045	.047	.049	.048	.052
Student *t*(3)	*t* test	.045	.045	.045	.048	.049	.049
	Mann–Whitney *U*	.044	.045	.046	.049	.053	.053

Note. Entries are proportions of 20,000 replications per cell in which the null hypothesis was rejected at α = .05. n = sample size per group; Mann–Whitney *U* = Mann–Whitney *U* test.

Table 2. Empirical Type I Error Rates for the Three-Sample Design

Distribution	Test	*n* = 10	*n* = 15	*n* = 30	*n* = 50	*n* =100	*n* =200
Normal	ANOVA	.051	.054	.050	.050	.048	.049
	Kruskal–Wallis *H*	.046	.052	.048	.049	.048	.050
Uniform	ANOVA	.052	.054	.049	.051	.052	.051
	Kruskal–Wallis *H*	.045	.051	.048	.051	.051	.051
Exponential	ANOVA	.042	.044	.046	.047	.049	.049
	Kruskal–Wallis *H*	.046	.047	.048	.047	.049	.048
Lognormal	ANOVA	.039	.039	.043	.048	.047	.049
	Kruskal–Wallis *H*	.049	.047	.047	.050	.048	.051
Student *t*(3)	ANOVA	.042	.042	.046	.049	.046	.048
	Kruskal–Wallis *H*	.046	.046	.050	.049	.049	.052

Note. Entries are proportions of 20,000 replications per cell in which the null hypothesis was rejected at α = .05. n = sample size per group; Kruskal–Wallis *H* = Kruskal–Wallis *H* test.

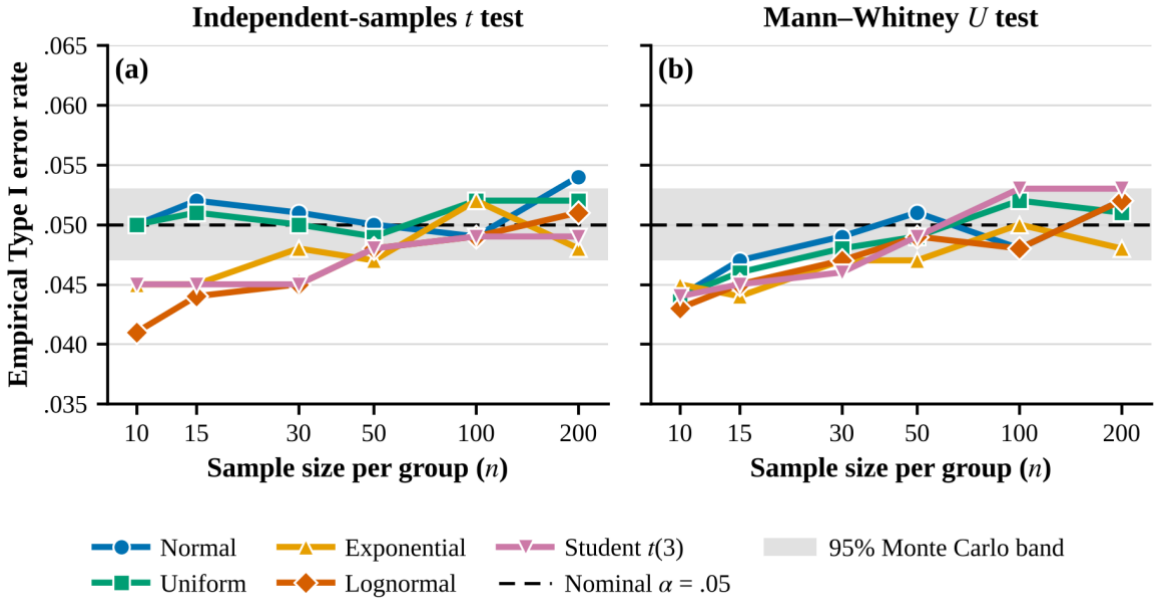


Figure 1. Empirical Type I Error Rates by Distribution and Sample Size in the Two-Sample Design

Note. The dashed line marks the nominal level α = .05 and the shaded band its 95% Monte Carlo interval (.05 ± .003 for 20,000 replications). Each point is the proportion of rejections in 20,000 replications.

Across all conditions, the empirical Type I error rate remained close to the nominal .05 level, with estimates falling between approximately .039 and .054. Given a Monte Carlo standard error of at most approximately .0016 at these replication counts, no rate deviated from the nominal level in a manner suggesting systematic miscalibration, with one qualified exception: under the Exponential and Lognormal distributions at the smallest sample sizes ($n = 10, 15$), both the *t* test and ANOVA showed a mild, consistent conservative tendency (rates as low as .039–.045), rather than an inflated error rate. This pattern is consistent with the classical finding that the *t* test's Type I error behavior under skewness tends toward conservatism, rather than anti-conservatism, for equal, small sample sizes, and directly parallels Nwobi and Akanno's (2021) finding that the *F* test's power advantage under normality, though numerically present, was not statistically distinguishable from that of the Kruskal–Wallis test, underscoring that the practically important criterion for choosing between these tests is statistical power rather than Type I error control.

Table 3 and Figure 3 report empirical statistical power for the two-group design at a medium effect size ($Cohen's\ d = 0.5$); Table 4 and Figure 4 report the corresponding results for the three-group design.

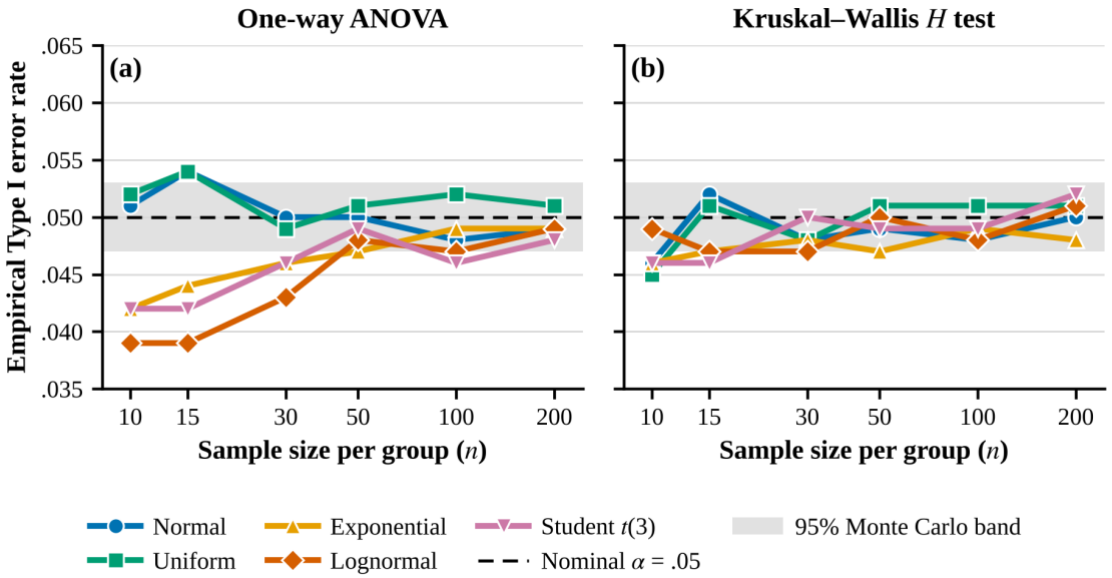


Figure 2. Empirical Type I Error Rates by Distribution and Sample Size in the Three-Sample Design

Note. The dashed line marks the nominal level α = .05 and the shaded band its 95% Monte Carlo interval (.05 ± .003 for 20,000 replications). Each point is the proportion of rejections in 20,000 replications.

Table 3. Empirical Statistical Power for the Two-Sample Design

Distribution	Test	*n* = 10	*n* = 15	*n* = 30	*n* = 50	*n* =100	*n* =200
Normal	*t* test	.191	.263	.478	.696	.942	.999
	Mann–Whitney *U*	.165	.237	.454	.673	.931	.998
Uniform	*t* test	.176	.250	.470	.693	.937	.999
	Mann–Whitney *U*	.153	.224	.431	.645	.912	.997
Exponential	*t* test	.224	.306	.503	.702	.939	.998
	Mann–Whitney *U*	.306	.454	.759	.934	.999	1.000
Lognormal	*t* test	.263	.348	.534	.719	.931	.997
	Mann–Whitney *U*	.360	.531	.842	.972	1.000	1.000
Student *t*(3)	*t* test	.256	.345	.558	.748	.937	.994
	Mann–Whitney *U*	.281	.415	.718	.916	.997	1.000

Note. Entries are proportions of 20,000 replications per cell in which a false null hypothesis was rejected at α = .05 (location shift of Cohen's *d* = 0.5). n = sample size per group.

Table 4. Empirical Statistical Power for the Three-Sample Design

Distribution	Test	*n* = 10	*n* = 15	*n* = 30	*n* = 50	*n* =100	*n* =200
Normal	ANOVA	.457	.655	.935	.995	1.000	1.000
	Kruskal–Wallis *H*	.421	.624	.921	.993	1.000	1.000
Uniform	ANOVA	.449	.641	.939	.996	1.000	1.000
	Kruskal–Wallis *H*	.391	.573	.897	.990	1.000	1.000
Exponential	ANOVA	.511	.677	.930	.993	1.000	1.000
	Kruskal–Wallis *H*	.676	.870	.995	1.000	1.000	1.000
Lognormal	ANOVA	.570	.711	.923	.988	1.000	1.000
	Kruskal–Wallis *H*	.771	.929	.999	1.000	1.000	1.000
Student *t*(3)	ANOVA	.579	.730	.930	.985	.998	1.000
	Kruskal–Wallis *H*	.684	.870	.996	1.000	1.000	1.000

Note. Entries are proportions of 20,000 replications per cell in which a false null hypothesis was rejected at α = .05 (group means shifted by 0, d, and 2d, with *d* = 0.5). n = sample size per group.

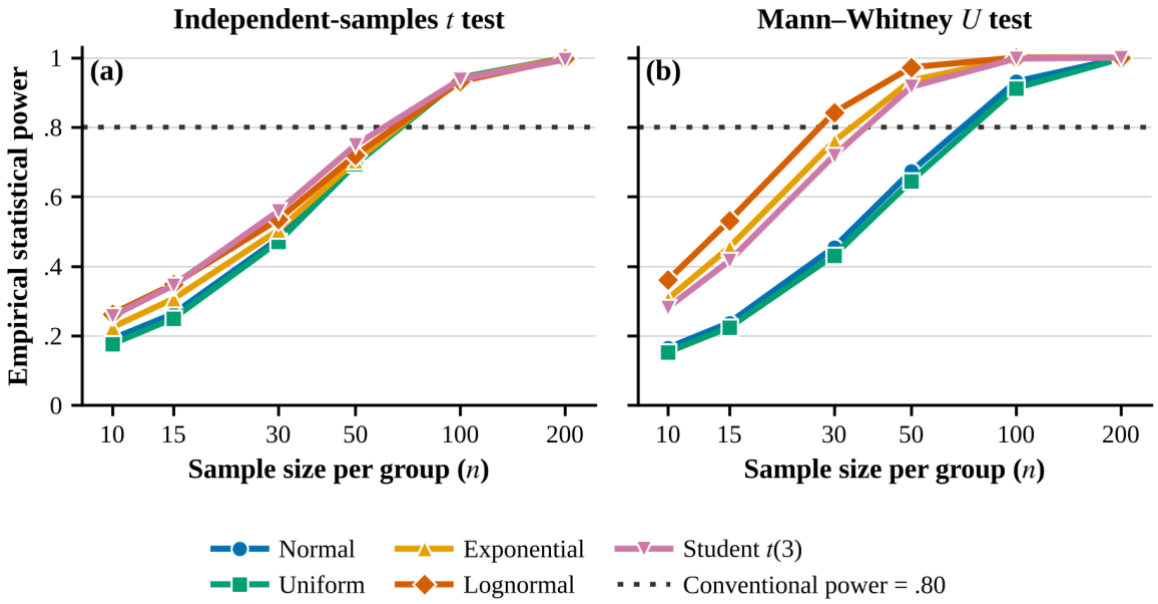


Figure 3. Empirical Statistical Power by Distribution and Sample Size in the Two-Sample Design

Note. Location shift of $d = 0.5$. The dotted line marks the conventional power of .80.

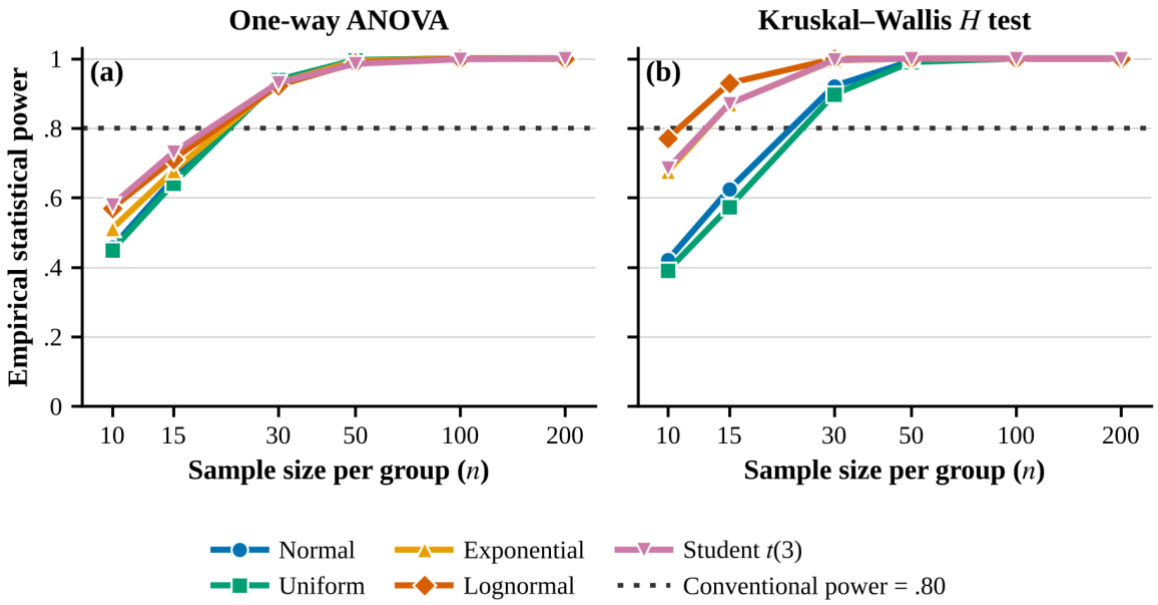


Figure 4. Empirical Statistical Power by Distribution and Sample Size in the Three-Sample Design

Note. Group means shifted by 0, d, and 2d, with $d = 0.5$. The dotted line marks the conventional power of .80.

Under the Normal and Uniform distributions, the t test held a modest but consistent power advantage over the Mann–Whitney U test at every sample size (Table 3 and Figure 3; for example, .478 versus .454 at $n = 30$ under the Normal distribution), narrowing in absolute terms as both tests approached the power ceiling at larger samples. Under the skewed and heavy-tailed distributions, this pattern reversed, often dramatically: under the Lognormal distribution at $n = 30$, the Mann–Whitney U test achieved a power of .842 against only .534 for the t test, a difference exceeding 30 percentage points, and the advantage remained visible even at $n = 100$ (.999 *vs* .931). This pattern directly replicates, under a substantially larger number of replications, the central finding of Şimşek (2023) that the Wilcoxon–Mann–Whitney test dominates for small-to-moderate, nonnormal samples, converging with the t test only once

samples are large and balanced. The three-group results (Table 4, Figure 4) mirror this pattern: Kruskal–Wallis held a clear power advantage over ANOVA under the Exponential, Lognormal, and Student $t(3)$ distributions at every sample size examined, most pronounced at small-to-moderate n, in line with Clark et al.'s (2023) finding that Kruskal–Wallis power estimation might perform better than the ANOVA equivalent in many settings.

Figure 5 summarizes the power differential (Mann–Whitney U minus t test) across sample size for the three most strongly nonnormal distributions. The advantage is not constant but peaks at small-to-moderate sample sizes (broadly $n = 15 - 50$) before narrowing as both tests' power saturates near 1.0 at the largest samples, indicating that the practical cost of selecting the less powerful test is greatest precisely in the sample-size range that characterizes much undergraduate and applied educational research, echoing the joint sample-size-by-distribution dependency documented by Kamath et al. (2025) for normality-test performance more generally.

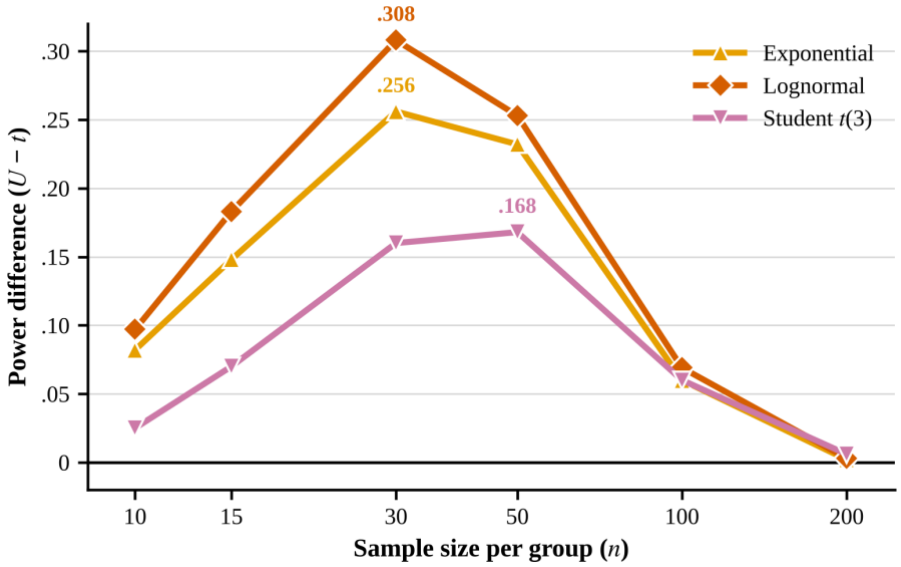


Figure 5. Power Difference Between the Mann–Whitney U Test and the t Test in the Two-Sample Design

Note. Positive values favor the Mann–Whitney U test. Labels give the largest difference for each distribution. The Normal and Uniform distributions are omitted because the Mann–Whitney U test is slightly less powerful than the t test there (Table 3).

The real data differ from the main simulation in three respects: the groups are unequal in size, their variances differ, and the final grade is zero-inflated and not smoothly skewed. The extended experiment was designed to separate these features. Table 5 gives the empirical Type I error rates at $N = 128$ and Table 6 the corresponding power. The results at $N = 64$ and $N = 256$ led to the same conclusions, with the one exception noted below.

With equal standard deviations, all three tests held their size at every allocation under the Normal distribution, with rates between .046 and .053. Under the zero-inflated distribution, Student's and the Mann–Whitney tests stayed between .047 and .052, and Welch's test rose to between .056 and .066 only when the smaller group had 8 or 16 observations. Unequal standard deviations changed the picture sharply for Student's t test, and the direction of the distortion depended on which group was the more variable. When the smaller group was more variable, Student's test was liberal, with rates between .149 and .236, three to five times the nominal level. When the larger group was more variable, it was severely conservative, with rates between .002 and .010. Welch's test remained between .047 and .066 in all 48 cells.

The Mann–Whitney U test was not a safe alternative. Under the Normal distribution it was liberal (.098 to .123) when the smaller group was more variable and conservative (.007 to .023) when the larger group was, a milder version of the pattern shown by Student's test. With zero-inflated data its rejection rate under equal means ranged from .017 to .202, and for equal group

sizes with a 2:1 ratio of standard deviations it rose with the sample size, from .088 at $N = 64$ to .200 at $N = 256$. This does not indicate a fault in the test on its own terms. The Mann–Whitney test evaluates whether one population is stochastically larger than the other, and two asymmetric distributions that share a mean but differ in spread are not stochastically equal, so rejection becomes more likely as the sample grows. It does mean that the test cannot be used as a remedy for unequal variances when the question concerns means.

Table 5. Empirical Type I Error Rates With Unequal Group Sizes and Unequal Variances

Allocation	Standard deviations	Normal			Zero-inflated		
		Student	Welch	MWU	Student	Welch	MWU
64:64	Equal	.051	.051	.049	.052	.052	.052
	Unequal (first group 2 times second)	.051	.051	.060	.051	.051	**.128**
32:96	Equal	.051	.050	.050	.050	.052	.049
	Smaller group 2 times larger	**.153**	.051	**.100**	**.155**	.055	**.151**
	Larger group 2 times smaller	**.008**	.050	**.023**	**.010**	.051	.069
16:112	Equal	.048	.049	.046	.048	.060	.048
	Smaller group 2 times larger	**.233**	.050	**.123**	**.230**	.064	**.146**
	Larger group 2 times smaller	**.002**	.051	**.008**	**.002**	.051	.028

Note. $N = 128$; $\alpha = .05$; 20,000 replications per cell. Allocation is the ratio of the smaller to the larger group. SD × 2 indicates that the group has twice the standard deviation of the other group. Student, Welch, and MWU denote Student's *t* test, Welch's *t* test, and the Mann–Whitney *U* test. Values outside .025 to .075 are in bold. The Mann–Whitney rate is the rejection rate when the population means are equal.

Table 6. Empirical Power With Unequal Group Sizes and Unequal Variances

Allocation	Standard deviations	Normal			Zero-inflated		
		Student	Welch	MWU	Student	Welch	MWU
64:64	Equal	.799	.799	.782	.805	.805	.940
	Unequal (first group 2 times second)	.943	.942	.925	.958	.957	.805
32:96	Equal	.681	.672	.655	.676	.684	.866
	Smaller group 2 times larger	.896	.751	.802	.913	.783	.676
	Larger group 2 times smaller	.202	.445	.291	.216	.444	.584
16:112	Equal	.460	.433	.437	.451	.410	.642
	Smaller group 2 times larger	.779	.457	.606	.782	.435	.477
	Larger group 2 times smaller	.056	.342	.126	.072	.331	.332

Note. $N = 128$; shift of 0.5 standard deviations of the first group; 20,000 replications per cell. Power is shown for all cells but is interpretable only where the Type I error rate in Table 5 is close to .05.

The power results (Table 6) are informative only where the size of the test is controlled. With equal variances, the pattern of the main simulation was reproduced. Under the Normal distribution, Student's and Welch's tests were slightly more powerful than the Mann–Whitney test (.799 vs. .782 for balanced groups), whereas under the zero-inflated distribution the Mann–Whitney test was far more powerful (.940 vs. .805), a gain of 13.5 percentage points that is comparable to those seen for the skewed distributions in Table 3. Allowing for unequal variances cost little when variances were in fact equal: at $N = 128$, Welch's power was within

.041 of Student's in every equal-variance cell. Where the larger group was more variable, Welch's test was clearly more powerful than Student's (.342 vs. .056 at a 1:7 allocation under the Normal distribution, and .331 vs. .072 with zero-inflated data), while the Mann–Whitney test fell between them for Normal data (.126) and matched Welch's test for zero-inflated data (.332). Where the smaller group was more variable, Student's test showed the highest power (.779 at 1:7 under the Normal distribution) only because it rejects far too often, so that apparent advantage is not genuine.

The simulation-derived patterns were next examined against the real dataset of 395 secondary-school students (Cortez & Silva, 2008). The final grade (G3) departed significantly from normality overall (Shapiro–Wilk $W = .929$, $p < .001$; skewness = −0.73), but the departure is of a different kind from the right-skewed Exponential and Lognormal shapes simulated in Tables 1 to 4. As Figure 6 shows, G3 is zero-inflated: 38 of the 395 grades (9.6%) equal zero, forming a spike that is detached from the remaining 357 grades. These range from 4 to 20, are roughly symmetric (median 11, skewness = 0.21) and have a mean of 11.52 and a standard deviation of 3.23. The negative overall skewness is produced by the spike and not by a long right tail. The zeros are also unevenly spread, making up 10.8% of the students without school support (37 of 344) but only 2.0% of those receiving it (1 of 51). G3 is therefore better described as a mixture than as a continuous skewed variable, and the comparisons below are read against the zero-inflated scenario of the extended experiment (Tables 5 and 6) as well as against the skewed scenarios of Tables 3 and 4.

Table 7. Comparison of Final Grade (G3) by Sex and School Educational Support

Comparison	Group	*n*	*M*	*SD*	Test	Statistic	*p*
Sex	Male	187	10.914	4.495	Welch *t* test	$t = 2.065$	.040
	Female	208	9.966	4.622	Mann–Whitney *U*	$U = 21{,}758$	.041
School support	Yes	51	9.431	2.865	Welch *t* test	$t = -2.370$	.020
	No	344	10.561	4.770	Mann–Whitney *U*	$U = 6{,}507$	.003
					Student *t* test	$t = -1.647$	.100

Note. G3 = final grade (0–20 scale). Degrees of freedom were 390.6 for the Welch t test by sex, 97.1 for the Welch t test by school support, and 393 for Student's *t* test. Welch's and Student's *t* tests compare means, whereas the Mann–Whitney *U* test compares the two distributions.

Table 7 compares G3 by sex. Despite confirmed nonnormality in both subgroups (Shapiro–Wilk $p < .001$), Welch's *t* test ($t(390.6) = 2.065$, $p = .040$, $d = 0.208$) and the Mann–Whitney *U* test ($U = 21{,}758$, $p = .041$, rank-biserial correlation = −.119) agreed closely. This agrees with the extended experiment, in which the three tests behaved alike when group sizes and spreads were nearly equal, even with zero-inflated data (Table 5, first row); here the groups are 187 and 208 students with standard deviations of 4.50 and 4.62.

Table 7 also compares G3 by receipt of school educational support, a comparison that is unbalanced (51 vs. 344 students, a ratio of about 1:6.7) and heteroscedastic (standard deviations 2.87 and 4.77). The three tests gave materially different answers. Student's pooled *t* test did not reject ($t(393) = -1.647$, $p = .100$), Welch's *t* test gave $t(97.1) = -2.370$, $p = .020$ ($d = -0.247$), and the Mann–Whitney *U* test gave $p = .003$ (rank-biserial correlation = .258). This ordering is what the extended experiment anticipates for a configuration in which the larger group is the more variable one. The pooled test is then severely conservative (a Type I error rate of .002 at the 1:7 allocation in Table 5) and loses power (Table 6), whereas Welch's test holds its size. The smaller Mann–Whitney *p* value should not be read as a power gain over the Welch test of the same hypothesis, because the two tests address different hypotheses and the groups differ in spread and in shape, a difference to which the zero grades, concentrated in the larger group, contribute. When the 38 zero grades were set aside, the evidence was much stronger and the tests agreed (Welch $t(76.9) = -5.46$, $p < .001$; Mann–Whitney $p < .001$), with students receiving support scoring lower in both analyses (mean 9.62 vs. 11.83). The school-support comparison therefore illustrates the consequences of unequal variances and zero-inflation, and not of skewness alone, and it is not direct confirmation of the nonparametric power advantage reported in Tables 3 and 4.

Table 8. Comparison of Final Grade (G3) by Mother's Education Level

Group	*n*	*M*	*SD*	Test	Statistic	*p*	Effect size
Low	62	8.887	4.402	One-way ANOVA	$F = 10.383$	<.001	$\eta^2 = .050$
Medium	202	10.010	4.627	Kruskal–Wallis *H*	$H = 21.007$	<.001	$\varepsilon^2 = .048$
High	131	11.763	4.268				

Note. G3 = final grade (0–20 scale). Degrees of freedom were 2 and 392 for the ANOVA F test and 2 for the Kruskal–Wallis *H* test. η^2 = eta squared; ε^2 = epsilon squared.

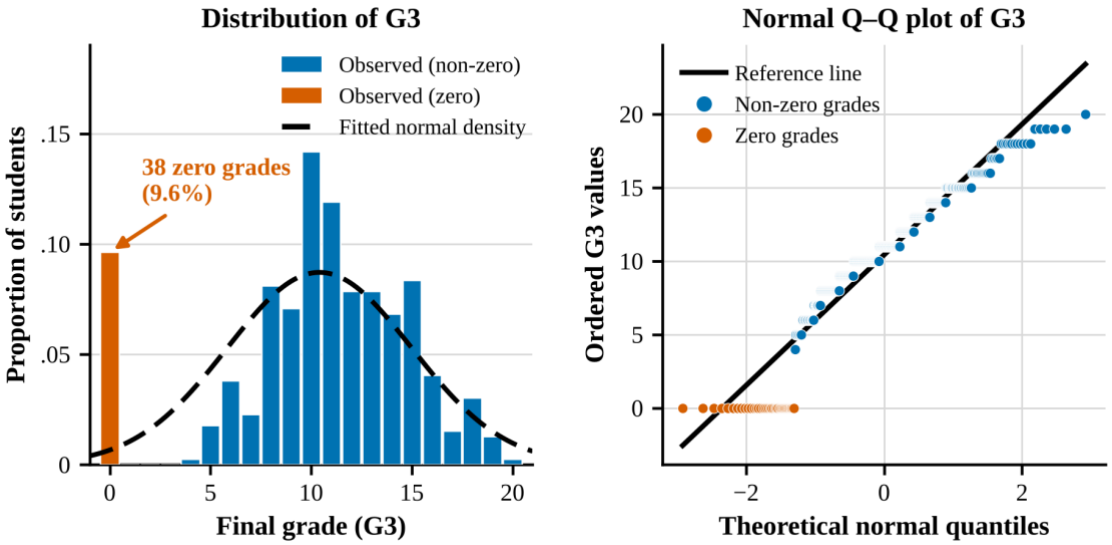


Figure 6. Distribution and Normal Q–Q Plot of Final Grade (G3), N = 395

Note. Panel (a) shows the observed proportion of students at each grade and a normal density with the sample mean and standard deviation. Panel (b) plots the ordered grades against normal quantiles; the 38 zero grades form the horizontal run at the lower left.

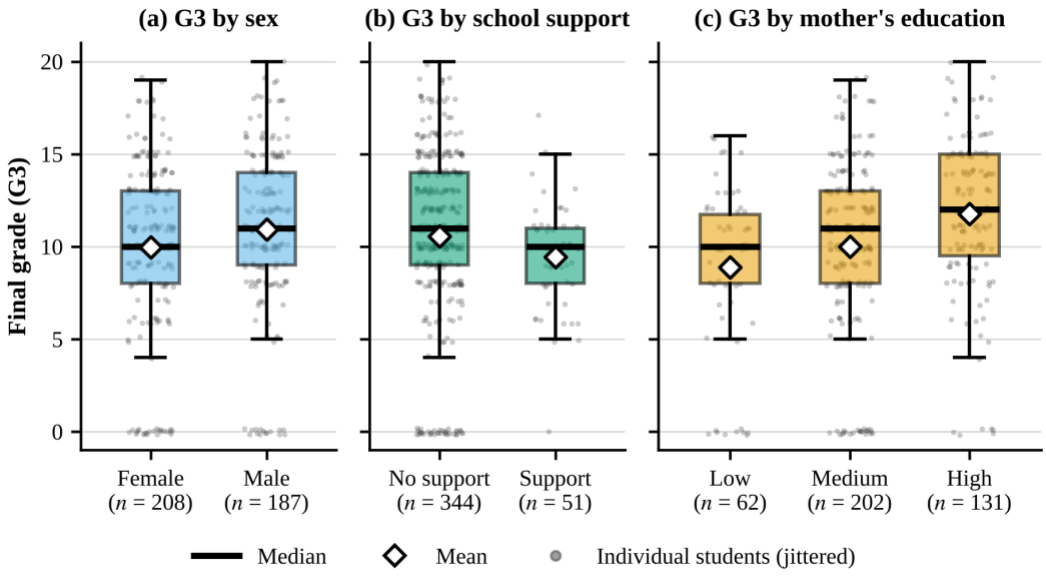


Figure 7. Boxplots of Final Grade (G3) by Sex, School Support, and Mother's Education

Note. Boxes show the interquartile range, whiskers extend to the most extreme values within 1.5 times the interquartile range, and the bar and diamond mark the median and mean. Mother's education: Low = none or primary education (up to 4th grade), Medium = 5th to 9th grade or secondary education, High = higher education.

Table 8 compares G3 across three ordered categories of maternal education, with the group distributions shown in Figure 7. One-way ANOVA ($F(2, 392) = 10.383$, $p < .001$, $\eta^2 = .050$) and the Kruskal–Wallis H test ($H(2) = 21.007$, $p < .001$, $\varepsilon^2 = .048$) agreed closely in both significance and effect-size magnitude. This is plausibly because the groups are comparatively large and well populated (62, 202 and 131 students) and have similar standard deviations (4.40, 4.63 and 4.27), a regime in which the simulation results indicate that the gap between the two test families narrows. Such close agreement, in an educational application structurally similar to Asare's (2026) comparison of information-technology students' statistics achievement across cohorts with the Kruskal–Wallis test, shows that the nonparametric test can be applied to nonnormal educational data without departing from the parametric conclusion when samples are adequately sized and balanced.

The findings, taken together, extend the recent methodological literature in three respects. First, using 20,000 replications per condition, considerably more than the 1,000 used by Nwobi and Akanno (2021) or the 1,000–2,000 typical of many prior comparisons, the present results narrow the Monte Carlo uncertainty surrounding the reported Type I error and power estimates, lending greater confidence to the subtle conservative-error finding under skewness at small n. Second, by crossing six sample sizes with five distributions within one internally consistent simulation for both a two-group and three-group design, the study traces the continuous evolution of the parametric-nonparametric power gap across the small-to-large sample spectrum, revealing that this gap is not monotonic but peaks at small-to-moderate sample sizes (Figure 5), extending Şimşek's (2023) and Clark et al.'s (2023) distribution-specific findings into a more complete, jointly parameterized picture. Third, by adding unequal allocation, unequal variances and zero-inflation to the design, the extended experiment shows that departures from equal variances can matter more than departures from normality. Student's t test was the only procedure whose Type I error was grossly distorted by unequal variances (.002 to .236), Welch's t test remained close to the nominal level in every cell, and the Mann–Whitney test, which addresses a different hypothesis, was not a dependable substitute. The school-support comparison illustrates the point on real data. The three tests gave p values of .100, .020 and .003, a spread that reflects unequal group sizes, unequal variances and zero grades concentrated in the larger group, and not skewness alone. This supports Vrbin's (2022) recommendation and the argument of Delacre et al. (2017) that test selection should rest on an explicit assessment of distributional shape, sample size and group variances and not on convention.

These findings carry direct practical relevance for educational and social-science research conducted with the modest sample sizes typical of undergraduate and postgraduate projects in many institutions, including within Nigerian and other West African universities. Research projects in this setting routinely compare examination scores, questionnaire-derived indices, or attendance and performance records across two or three groups defined by sex, school type, or socioeconomic background, precisely the kind of comparison simulated here, and frequently with sample sizes in the small-to-moderate range where the present results show the parametric-nonparametric divergence to be largest. Supervisors and methods instructors may therefore find it useful to require an explicit report of distributional shape (skewness, kurtosis, or a normality test) alongside the group sizes, standard deviations and chosen inferential test, so that the test-selection decision is made transparently rather than by default, consistent with the broader call for more transparent statistical reporting practices in Vrbin (2022) and Kamath et al. (2025).

5. CONCLUSION AND SUGGESTIONS

This study conducted a high-replication Monte Carlo simulation, validated against real educational data, to compare the performance of parametric and nonparametric statistical tests under varying sample sizes and data distributions. Both the independent-samples t test and one-way ANOVA maintained empirical Type I error rates close to the nominal .05 level across virtually all balanced, equal-variance conditions, with only a mild conservative tendency under pronounced skewness at the smallest sample sizes. With respect to statistical power, the parametric tests held a modest, consistent advantage under symmetric, light-tailed distributions, while the nonparametric tests demonstrated a substantial, and at small-to-moderate sample sizes often dramatic, power

advantage under skewed and heavy-tailed distributions. A second experiment showed that unequal group sizes combined with unequal variances distorted Student's *t* test far more than nonnormality did (Type I error from .002 to .236), whereas Welch's *t* test stayed between .047 and .066 and the Mann–Whitney test, which does not test equality of means, was not a dependable substitute (.007 to .202). In the real data of 395 secondary-school students, the tests agreed closely for the balanced comparisons (sex and mother's education) but diverged for the unbalanced, heteroscedastic school-support comparison ($p = .100$, .020 and .003 for Student's *t*, Welch's *t* and Mann–Whitney), a divergence explained by unequal group sizes, unequal variances and the zero-inflation of the final grade, and not by skewness alone.

On this basis, the choice between a parametric and nonparametric test should not be resolved by convention, software default, or an isolated normality pre-test, but by a joint, explicit consideration of sample size, group balance, variance and distributional shape. Researchers working with small-to-moderate samples (broadly $n < 50$ per group) and skewed or heavy-tailed data are advised to prefer the Mann–Whitney *U* test or the Kruskal–Wallis *H* test over their parametric counterparts, given the substantial power advantage demonstrated here. When the target is a difference in means and group sizes or variances differ, Welch's *t* test should replace Student's *t* test, and the Mann–Whitney test should be chosen because its stochastic-ordering hypothesis is of substantive interest and not as a remedy for unequal variances. Parametric tests remain the more efficient choice under approximately symmetric, light-tailed data with balanced groups and similar variances.

Some limitations should be acknowledged. The simulation examined only two experimental designs (independent two-group and three-group comparisons) and a single medium effect size (Cohen's $d = 0.5$), and the extended experiment covered only the two-group design, with one variance ratio (2:1), three allocation ratios and one zero-inflated distribution calibrated on a single dataset; paired designs, factorial structures, and a continuous range of effect sizes were not examined. The five simulated distributions, while spanning symmetric, skewed, and heavy-tailed shapes, do not exhaust every distributional form encountered in practice, such as multimodal data common in survey research. The real-data validation drew on a single, though genuinely nonnormal and appropriately sized, educational dataset; the specific empirical relationships it revealed should be read as illustrative of the statistical-testing question under study rather than as generalizable substantive findings about any particular educational system. The meaning of the zero grades could not be established from the data, so they were treated as observed values.

Future studies could extend the present simulation framework to paired and repeated-measures designs (the paired *t* test vs. the Wilcoxon signed-rank test), to additional distributional families such as multimodal distributions, to unbalanced and heteroscedastic three-group designs, to two-part models that treat the zeros of a zero-inflated outcome separately, and to a continuous range of effect sizes to construct fuller power curves. The real-data validation could also be extended to additional, ideally Nigerian or West African, educational datasets to assess whether the specific patterns of parametric-nonparametric agreement and divergence documented here generalize across educational systems and measurement contexts.

DECLARATIONS

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflict of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability

The real dataset analyzed in this study (Cortez, 2014; Cortez & Silva, 2008) is publicly available through the UCI Machine Learning Repository. The Monte Carlo simulation code, together with the code used for the real-data analysis, is available from the corresponding author upon reasonable request.

Author Contributions

All authors contributed to the conception and design of the study, the implementation of the Monte Carlo simulation, the interpretation of

results, and the drafting and critical revision of the manuscript. All authors read and approved the final manuscript.

Acknowledgements

The authors are grateful to the anonymous reviewers for their constructive comments on an earlier version of this manuscript.

REFERENCES

Asare, B. (2026). A comparison of information technology education students' achievement in statistics using Kruskal-Wallis H-test statistic: A nonparametric analysis. *Frontiers in Education, 11*, Article 1873889. https://doi.org/10.3389/feduc.2026.1873889

Ayenigba, A. A., & Ajao, O. M. (2026). A review of chi-square test of hypothesis and its applications in real-life situations. *Ajayi Crowther Journal of Pure and Applied Sciences, 5*(1), 1–10. https://doi.org/10.56534/acjpas.v5i1.187

Chicco, D., Sichenze, A., & Jurman, G. (2025). A simple guide to the use of Student's t-test, Mann-Whitney U test, chi-squared test, and Kruskal-Wallis test in biostatistics. *BioData Mining, 18*, Article 56. https://doi.org/10.1186/s13040-025-00465-6

Clark, J. S. C., Kulig, P., Podsiadło, K., Rydzewska, K., Arabski, K., Białecka, M., Safranow, K., & Ciechanowicz, A. (2023). Empirical investigations into Kruskal-Wallis power studies utilizing Bernstein fits, simulations and medical study datasets. *Scientific Reports, 13*, Article 2352. https://doi.org/10.1038/s41598-023-29308-2

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

Cortez, P. (2014). *Student performance* [Data set]. UCI Machine Learning Repository. https://doi.org/10.24432/C5TG7T

Cortez, P., & Silva, A. M. G. (2008). Using data mining to predict secondary school student performance. In A. Brito & J. Teixeira (Eds.), *Proceedings of the 5th Annual Future Business Technology Conference* (pp. 5–12). EUROSIS-ETI.

Delacre, M., Lakens, D., & Leys, C. (2017). Why psychologists should by default use Welch's t-test instead of Student's t-test. *International Review of Social Psychology, 30*(1), 92–101. https://doi.org/10.5334/irsp.82

Fisher, R. A. (1925). *Statistical methods for research workers*. Oliver & Boyd.

Kamath, A., Poojari, S., & Varsha, K. (2025). Assessing the robustness of normality tests under varying skewness and kurtosis: A practical checklist for public health researchers. *BMC Medical Research Methodology, 25*, Article 206. https://doi.org/10.1186/s12874-025-02641-y

Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association, 47*(260), 583–621. https://doi.org/10.2307/2280779

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics, 18*(1), 50–60. https://doi.org/10.1214/aoms/1177730491

Nwobi, F. N., & Akanno, F. C. (2021). Power comparison of ANOVA and Kruskal-Wallis tests when error assumptions are violated. *Advances in Methodology and Statistics, 18*(2), 53–71. https://doi.org/10.51936/ltgt2135

Oladugba, A. V., & Ikebuife, U. V. (2026). A Monte Carlo simulation study of the type-I-error and power of ANOVA F-test, Friedman test and randomization test for randomized complete block design. *Quality & Quantity, 60*(1), 2781–2803. https://doi.org/10.1007/s11135-025-02374-6

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika, 52*(3–4), 591–611. https://doi.org/10.1093/biomet/52.3-4.591

Şimşek, A. S. (2023). The power and type I error of Wilcoxon-Mann-Whitney, Welch's t, and Student's t tests for Likert-type data. *International Journal of Assessment Tools in Education, 10*(1), 114–128. https://doi.org/10.21449/ijate.1183622

Vrbin, C. M. (2022). Parametric or nonparametric statistical tests: Considerations when choosing the most appropriate option for your data. *Cytopathology, 33*(6), 663–667. https://doi.org/10.1111/cyt.13174

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin, 1*(6), 80–83.