Journal of Research in Didactical Sciences
ISSN: 2997-5905 (Online)
VOL. 5, ISSUE 1, 2026
Pages: 133-31

RESEARCH ARTICLE

OPEN ACCESS

Mann-Whitney U Test for Two Independent Groups

Bright Asare*[1], Yarhands Dissou Arthur[1], Benjamin Adu Obeng[1], Sadri Alija[2], Senad Orhani[3]

[1] University of Skills Training and Entrepreneurial Development (USTED), Kumasi, Ghana
[2] Faculty of Business and Economics, South East European University, North Macedonia
[3] Faculty of Education, University of Prishtina, Prishtina, Kosovo
* Corresponding author: asarebrigth6592@gmail.com

Abstract: The Mann-Whitney U test is commonly used to compare two independent groups, especially when the data do not meet the assumptions required for parametric tests. Although it is widely applied in fields such as education, psychology, and health sciences, it is often misunderstood or incorrectly interpreted. This paper provides a clear and practical guide to the Mann-Whitney U test by explaining its key concepts, assumptions, and step-by-step procedures. It also highlights common mistakes, particularly the tendency to interpret the test as a direct comparison of medians, and stresses the importance of reporting effect sizes alongside p-values. Using simple examples and software illustrations, the paper shows how the test can be properly applied in real research situations. In all, the study aims to help researchers use the Mann-Whitney U test more accurately and confidently in their analyses

Keywords: *Mann-Whitney Test, Independent Group, Wilcoxon Rank-Sum test, Statistical software*

INTRODUCTION

In many real-world studies, researchers are forced to analyze data that violate key statistical assumptions such as normality and homogeneity of variance. Despite the widespread use of parametric tests like the t-test, the inappropriate application of these methods can lead to misleading conclusions. This challenge has increased the relevance of nonparametric alternatives such as the Mann-Whitney U test, particularly in education, health, and social sciences.

The Mann-Whitney U test was developed by Henry Mann and Donald Whitney in 1947 as a nonparametric alternative to the independent samples t-test. It is also called the Wilcoxon Rank-Sum Test, because it is mathematically equivalent to the method introduced by Frank Wilcoxon in 1945 (Wilcoxon, 1945). This test was an important advancement in statistics, allowing researchers to compare groups when data did not satisfy parametric requirements. Over time, the test has become widely adopted in many research areas, including psychology, education, medicine, and social sciences (Nachar, 2008; Okeh, 2009). Its continued use reflects its flexibility and robustness, making it a valuable tool for researchers who work with real-world data that often do not meet ideal conditions.

The main purpose of the Mann-Whitney U test is to determine whether two independent groups differ in a meaningful way when the data are ordinal or non-normal. The test works by pooling all the observations from both groups, ranking them from smallest to largest, and then calculating the U statistic, which represents how often values from one group precede values from the other. A larger difference in ranks indicates a bigger difference between groups. The Mann-Whitney U statistic is closely related to the Wilcoxon Rank-Sum statistic, so both tests essentially measure the same thing for two independent samples (Lin et al., 2021). This approach makes the Mann-Whitney U test a simple yet powerful method for comparing groups when parametric methods are not suitable. This paper provides

© 20236 The Author(s). This article is distributed under the Creative Commons Attribution 4.0 International License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction, as long as the original author(s) and source are credited and any modifications are indicated. The data and materials are also dedicated to the public domain under the CC0 1.0 waiver, unless otherwise stated.

a comprehensive and practical guide to the Mann-Whitney U test by integrating theoretical foundations, computational procedures, effect size interpretation, and real-world applications, with a focus on improving its correct usage in applied research.

Theoretical Background

The Mann-Whitney U test is a nonparametric statistical procedure developed by Mann and Whitney (1947) to compare differences between two independent groups when the assumptions of parametric tests are not met. It is commonly used as an alternative to the independent samples *t*-test, particularly when data are ordinal, skewed, or contain outliers. The theoretical foundation of the test is based on rank ordering rather than raw data values, which makes it robust against violations of normality and homogeneity of variance. By converting observations into ranks, the Mann-Whitney U test focuses on the relative position of scores within the combined sample, thereby reducing the influence of extreme values (Pérez et al., 2015). From a theoretical perspective, the Mann-Whitney U test examines whether two independent samples come from the same population distribution. Rather than directly comparing means, the test evaluates differences in the central tendency of rank distributions between groups. When the shapes of the distributions are similar, the test is often interpreted as a comparison of medians (Wilcox, 2005). The U statistic is derived from the sum of ranks assigned to each group, and its distribution under the null hypothesis is well established, allowing for the calculation of exact or asymptotic p-values depending on sample size (Liu & Xie, 2020).

In terms of hypothesis testing, the Mann-Whitney U test is based on the following assumptions: the null hypothesis (H_0) states that there is no significant difference between the two independent groups, meaning their distributions (or medians, if the shapes are similar) are equal. The alternative hypothesis (H_1) states that there is a significant difference between the two groups, meaning their distributions (or medians) are not equal.

The importance of the Mann-Whitney U test in theory lies in its flexibility and wide applicability across disciplines such as education, psychology, health sciences, and social research. It aligns with distribution-free statistical theory, meaning it does not rely on strict population parameters, which enhances its usefulness in real-world research settings where ideal conditions are rarely met. As a result, the test is theoretically grounded in nonparametric inference and continues to be a reliable tool for analyzing group differences when traditional parametric assumptions are untenable.

Assumptions of the Mann-Whitney U Test

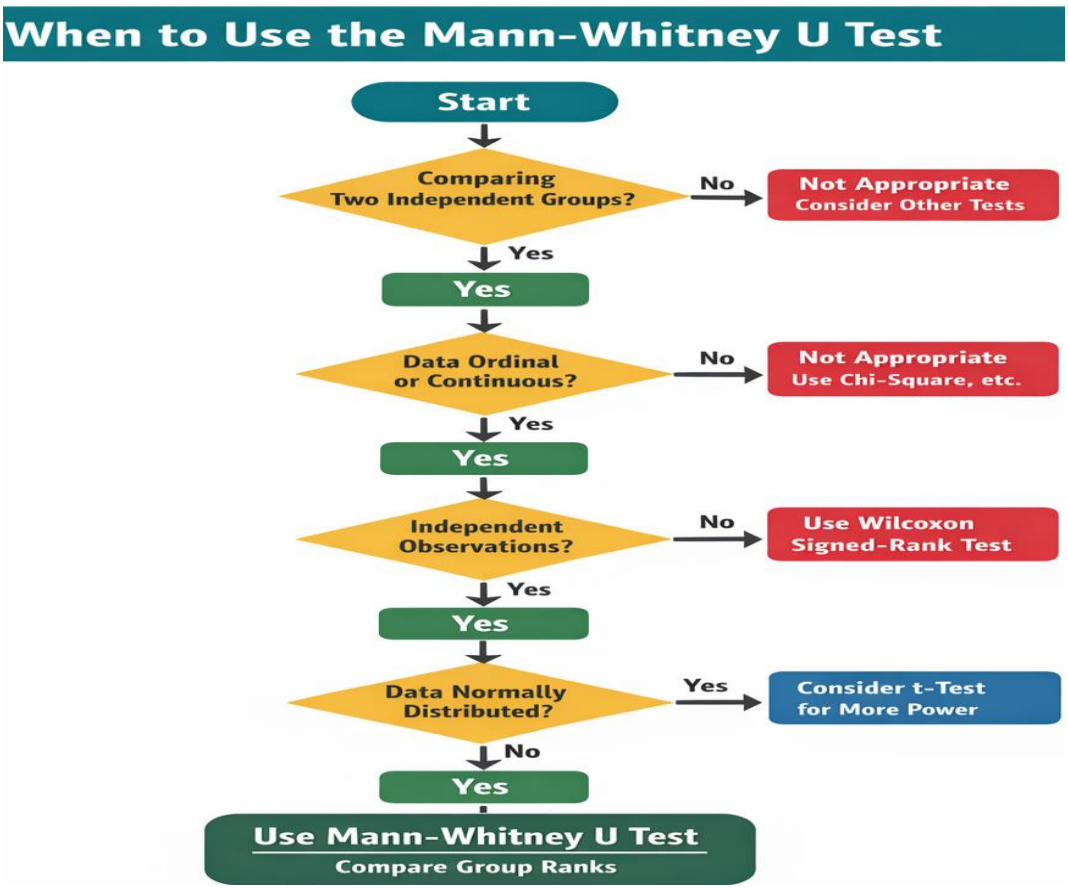
The Mann-Whitney U test is a widely used nonparametric statistical method designed to compare two independent groups, but its validity depends on several key assumptions (Sedgwick, 2015). First, the samples being compared must be independent. This means that the selection of participants or observations in one group should not influence the other group in any way. If the samples are related, such as measurements taken from the same participants at different times or paired observations, the results of the Mann-Whitney test may be misleading. In such cases, alternative tests like the Wilcoxon signed-rank test are more appropriate. Ensuring true independence helps maintain the integrity of the ranking procedure that underpins the U statistic. Another crucial assumption is the type of data that can be analyzed. The Mann-Whitney U test is suitable for ordinal, interval, or ratio data. Ordinal data, such as Likert-scale responses, allow the researcher to rank observations meaningfully, even if the distances between ranks are not equal (Rutkowski, 2025). Interval and ratio data, which have consistent measurement scales, also meet the requirements of the test. Using nominal data or data without a logical order can violate this assumption, rendering the ranking process invalid and the resulting statistical conclusions unreliable (Norman, 2010). Therefore, careful consideration of the measurement scale is essential when choosing to apply this test. A final assumption often overlooked is that the distributions of the two groups should have a similar shape, especially when the goal is to compare medians. If the distributions differ substantially, the Mann-Whitney test may detect differences in shape or spread rather than central tendency, leading to misinterpretation of results. Additionally, random sampling

from the population is assumed to ensure that the groups represent the larger population accurately. Violations of these assumptions, such as skewed distributions or non-random sampling, can bias the U statistic and p-values, potentially resulting in incorrect conclusions. Researchers should therefore check for distributional similarities and ensure sampling procedures are rigorous to preserve the validity of the analysis.

Flowchart on When to Use the Mann-Whitney U Test

The process starts, as shown in *Figure 1*, by asking a simple but important question: are you trying to compare two independent groups? If the answer is no, then the Mann-Whitney U test is not the right choice, and you should look for another statistical method that better fits your study design. However, if you are comparing two separate groups, the next step is to examine the type of data you have. The test works well with continuous data, such as scores, or ordinal data, such as rankings. If your data are categorical instead, then a different test, like Chi-square, would be more appropriate. After confirming the data type, you also need to ensure that your observations are independent, meaning each value comes from a different subject or unit. If the data are related or paired, the Mann-Whitney U test is not suitable, and a Wilcoxon signed-rank test would be a better alternative. Another key consideration is whether your data follows a normal distribution. If they do, a t-test is often preferred because it has greater statistical power. But if the data are not normally distributed, or if the sample size is small, the Mann-Whitney U test becomes a strong option since it does not rely on normality assumptions. When all these conditions are met, two independent groups, an appropriate data type, independent observations, and a non-normal distribution, you can confidently use the Mann-Whitney U test to compare the groups and determine whether there is a meaningful difference between them.

Figure 1. Flowchart on When to Use Mann-Whitney U Test (Source: Authors' Creation, 2026)



Data Requirements and Research Design

The Mann-Whitney U test is designed for research situations involving two independent groups and requires data that are measured at ordinal or continuous levels. Unlike parametric tests, it does not assume that the dependent variable is normally distributed, making it appropriate for skewed data, small sample sizes, or data containing outliers. The key data requirement is that observations must be independent, meaning each participant or case contributes only one score to the dataset and belongs to only one group. According to Mann and Whitney (1947), the test compares the relative positions of scores between groups by ranking the combined data, which allows researchers to examine differences in distributions rather than relying on mean values.

From a research design perspective, the Mann-Whitney U test is commonly applied in experimental, quasi-experimental, and non-experimental (survey or observational) studies where random assignment may not be feasible. The independent variable must consist of two distinct and unrelated groups, such as male versus female students or treatment versus control groups, while the dependent variable reflects an outcome of interest. Dehaene et al. (2021) emphasizes that the test is particularly useful in social science and educational research where assumptions of homogeneity of variance and normality are often violated, yet meaningful group comparisons are still required.

In terms of sample size and distributional considerations, the Mann-Whitney U test is flexible and performs well with both small and moderate samples, though larger samples improve statistical power and allow for normal approximation of the test statistic. The shape of the distributions in the two groups should be similar if the test is interpreted as a comparison of medians; otherwise, it is more accurately interpreted as a comparison of overall distributions (Sihombing, 2025). This makes the test well-suited for applied research contexts where data quality and measurement conditions vary, reinforcing its value as a robust alternative to the independent samples t-test.

Procedure for Conducting the Test

The first step in conducting a Mann-Whitney U test is to combine and rank the data from both groups. All observations from the two independent groups are pooled together and ranked from the smallest to the largest value. If two or more values are tied, they are assigned the average of their ranks. This ensures that the ranking is fair and accounts for ties in the data.

Next, the rank sums for each group are calculated. This involves adding together all the ranks that belong to the first group to obtain R_1, and all the ranks for the second group to obtain R_2. These rank sums provide the basis for computing the U statistic and reflect the relative standing of the groups in the combined ranking. After calculating rank sums, the U-statistics for each group are determined. The formulas are

$$U_1 = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - R_1$$

for the first group and

$$U_2 = n_1 n_2 + \frac{n_2(n_2 + 1)}{2} - R_2$$

For the second group, where n_1 and n_2 are the sample sizes of the groups? The smaller of the two U values is typically used for testing significance. Once the U statistic is computed, significance testing is performed. For small sample sizes (less than 20 per group), critical values of U from reference tables are used. For larger samples, the U statistic can be converted to a Z-score using the formulas

$$\mu_U = \frac{n_1 n_2}{2}$$

And

$$\sigma_U = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$$

The Z-score is then compared to the standard normal distribution to determine the p-value and test whether the difference between groups is statistically significant. Finally, the effect size is calculated to assess the magnitude of the difference between groups. The effect size $r$is computed as

$$r = \frac{Z}{\sqrt{N}}$$

where N is the total number of observations in both groups. Cohen's guidelines are often used for interpretation, with $r = 0.1$ indicating a small effect, $r = 0.3$ a medium effect, and $r = 0.5$ a large effect. In reporting the results, the rank sums, the U statistic, the Z-score, the p-value, and the effect size should all be included. This provides a complete picture of the comparison between the two groups, showing both statistical significance and the strength of the observed difference.

Applications in Research

The Mann-Whitney U test has been widely employed across multiple research fields due to its robustness in handling nonparametric data and small sample sizes. In education, for example, researchers frequently use the test to compare student performance between different teaching methods or intervention groups, particularly when test scores or survey responses do not meet the assumptions required for parametric tests. Similarly, in health sciences, the test is applied to evaluate the effectiveness of treatments or interventions where clinical measurements or patient-reported outcomes are skewed or measured on an ordinal scale. Within psychology, the Mann-Whitney U test is often utilized to examine differences in behavioral outcomes, cognitive scores, or attitudes between independent groups, offering a reliable alternative when data deviate from normality or sample sizes are limited.

One of the primary reasons for the widespread use of the Mann-Whitney U test is its ability to handle data that violate parametric assumptions. Non-normal distributions are common in social science research, business studies, and clinical trials, where extreme values or skewed data can distort the results of traditional t-tests (Vickers, 2005). Additionally, the test is particularly advantageous when working with small sample sizes, where parametric methods might produce unreliable estimates. It also allows for the analysis of ordinal data, such as Likert-scale survey responses, enabling researchers to make valid comparisons between groups without assuming equal intervals between scale points. This flexibility makes the test especially valuable in exploratory studies and pilot research where data characteristics are uncertain.

Empirical studies across various disciplines demonstrate the practical application of the Mann-Whitney U test. For instance, in education research, Khan et al. (2017) found that the Whitney U test has been used to compare student engagement scores between traditional and online learning groups, highlighting meaningful differences despite small class sizes. In health sciences, it has been applied to assess changes in pain levels or quality-of-life measures between control and treatment groups in clinical trials. In psychology, studies have examined differences in stress or anxiety levels between experimental and control groups, revealing significant distinctions that might be missed using parametric methods. Across business and social science research, the test is utilized to compare customer satisfaction ratings, employee engagement levels, or community survey responses between independent groups, providing a reliable statistical approach when assumptions for parametric tests are not met.

Strengths of the Mann-Whitney U Test

One of the key strengths of the Mann-Whitney U test lies in its robustness to outliers (Bakker & Wicherts, 2014). Unlike parametric tests, such as the independent samples t-test, which can be heavily influenced by extreme values, the Mann-Whitney test relies on the ranking of data rather than the raw

numbers themselves. This ranking process minimizes the effect that unusually high or low values have on the outcome, making the test particularly reliable when datasets include anomalies or skewed measurements. Researchers often encounter datasets in social sciences, healthcare, and education where extreme observations naturally occur, and in such cases, the Mann-Whitney U test provides a more stable and trustworthy comparison between groups (Wonu & Anaekwe, 2023). Another significant advantage of the Mann-Whitney U test is that it does not require the data to follow a normal distribution. Many traditional parametric methods assume normality, which is often violated in real-world research, especially with small sample sizes or ordinal data (Olvera, 2020). The Mann-Whitney test is nonparametric, meaning it can effectively handle data that are skewed, heteroscedastic, or measured on an ordinal scale, such as survey ratings or rankings. Its flexibility in accommodating ordinal measurements further broadens its applicability, allowing researchers to compare groups even when precise interval-level data are unavailable. This combination of robustness and minimal assumptions makes the Mann-Whitney U test a powerful and practical tool for analyzing a wide range of datasets.

Limitations and Misconceptions

While the Mann-Whitney U test is widely used for comparing two independent groups, it has several limitations that researchers should carefully consider. One key constraint is that the test is sensitive to differences in the shape of the distributions between the groups (Emerson, 2023). If the distributions are not similarly shaped, the test may not accurately reflect differences in central tendency, and interpreting the results as a direct comparison of medians can be misleading. Additionally, the Mann-Whitney U test is specifically designed for independent samples, and therefore cannot be applied to paired or matched data. In situations involving repeated measures or paired observations, alternatives like the Wilcoxon Signed-Rank test are more appropriate. Researchers must also be cautious when working with small sample sizes or highly skewed data, as these conditions can affect the reliability of the test statistics.

Another area of concern lies in common misconceptions surrounding the interpretation of the test. Many researchers mistakenly assume that the Mann-Whitney U test compares group means directly, when in reality it primarily evaluates whether one group tends to have higher or lower values than the other (Chicco et al., 2025). Similarly, it is often misapplied to Likert-scale data without considering whether treating ordinal responses as interval measurements is justified, which can compromise the validity of conclusions (Abreh et al., 2025). Furthermore, effect sizes are sometimes ignored, leading to an incomplete understanding of the practical significance of findings, even when statistical significance is achieved. Awareness of these misconceptions is essential for conducting robust analyses and reporting results accurately.

Comparison with Alternative Tests

When comparing the Mann-Whitney U test with alternative statistical methods for analyzing two independent groups, several options emerge, each with specific strengths and limitations. The independent samples t-test is the parametric counterpart to the Mann-Whitney test and is most suitable when the data are continuous, normally distributed, and variances across groups are roughly equal (Gibbons et al., 1991). When these assumptions are violated, the Welch's t-test offers a more flexible alternative by adjusting for unequal variances, making it robust in situations where group variability differs significantly. The Kolmogorov-Smirnov test, in contrast, focuses on the overall difference between cumulative distributions rather than comparing central tendencies, making it useful for detecting general distributional differences but less intuitive for median or mean comparisons (Zarrin et al., 2020). For cases where the shapes of distributions differ substantially, the Brunner-Munzel test offers a more robust nonparametric alternative to the Mann-Whitney test, as it does not rely on the assumption of similar distributions and provides a more accurate assessment of stochastic dominance between groups. Finally, permutation tests offer a flexible, assumption-free approach by empirically

generating the sampling distribution of the test statistic through repeated resampling, which can provide highly accurate p-values, especially in small or irregular samples.

Despite the availability of these alternative methods, the Mann-Whitney U test remains a preferred choice under specific conditions. It is particularly advantageous when dealing with small sample sizes or ordinal data, where parametric assumptions of normality or equal variance cannot be justified. Its simplicity, computational efficiency, and robustness to outliers make it a practical tool for many applied research scenarios. However, it becomes less suitable when the distributions of the two groups differ markedly in shape, as it may then misrepresent differences in central tendency. In such situations, researchers might prefer the Brunner-Munzel test or permutation-based approaches to obtain more reliable results. Generally, the choice between these tests depends on the nature of the data, the sample size, and the specific research question, with the Mann-Whitney U test striking a balance between simplicity, interpretability, and applicability in nonparametric contexts (Wang et al., 2025).

Reporting Guidelines

When reporting the results of a Mann-Whitney U test in research articles, it is essential to provide a clear and comprehensive description of the data and analysis (Wonu & Anaekwe, 2023). Authors should begin by detailing the sample sizes of the two independent groups and include relevant group characteristics such as age, gender, or other variables that may affect interpretation. Providing these details helps readers assess the comparability of groups and the reliability of the results. The core statistical outcomes of the test, specifically the U statistic and the p-value, should be clearly stated. The U statistic summarizes the rank differences between the two groups, while the p-value indicates whether the observed difference is statistically significant (White et al., 2022). It is equally important to report effect sizes, which quantify the magnitude of the difference between groups and give practical context to the statistical findings. Common measures for effect size in Mann-Whitney analyses include the rank-biserial correlation and Cliff's delta, both of which allow readers to understand the strength of the observed effect beyond p-values alone.

In addition to test statistics, authors should provide descriptive summaries of the data, particularly the medians and interquartile ranges (IQRs) for each group. Because the Mann-Whitney test is nonparametric and often used with non-normally distributed data, these measures give a more accurate picture of central tendency and variability than means and standard deviations. It is also important to discuss assumption checks, such as confirming the independence of samples and examining whether the distributions of the groups are similarly shaped, as violations can affect the interpretation of the test. For example, a proper reporting statement could read: *"A Mann-Whitney U test was conducted to compare test scores between Group A (n = 25, median = 78, IQR = 12) and Group B (n = 27, median = 85, IQR = 10). The test indicated a significant difference, U = 210, p = 0.03, with a moderate effect size (rank-biserial correlation = 0.38), suggesting that Group B scored higher than Group A."* This approach ensures transparency, reproducibility, and clarity in conveying statistical results to the reader.

Example 1: Health Study: Comparing Pain Scores

A Mann-Whitney U test was performed to compare pain scores between patients receiving Treatment X (n = 30, median = 6, IQR = 2) and Treatment Y (n = 28, median = 4, IQR = 3). The analysis revealed a significant difference in scores, U = 280, p = 0.012, with a small-to-moderate effect size (rank-biserial correlation = 0.32), indicating that patients in the Treatment Y group experienced lower pain levels than those in Treatment X.

Example 2: Education Study: Comparing Exam Performance

To assess differences in exam performance between students taught using Method A (n = 22, median = 82, IQR = 6) and Method B (n = 24, median = 88, IQR = 5), a Mann-Whitney U test was conducted. Results indicated a statistically significant difference, U = 160, p = 0.004, with a moderate effect size (Cliff's delta = 0.41), suggesting that students in Method B performed better overall.

Example 3: Social Sciences: Comparing Satisfaction Levels

A Mann-Whitney U test compared job satisfaction ratings between employees in Department 1 (n = 35, median = 4.0, IQR = 0.8) and Department 2 (n = 33, median = 3.5, IQR = 1.0). The test showed no statistically significant difference, U = 540, p = 0.178, with a small effect size (rank-biserial correlation = 0.15), indicating similar satisfaction levels across departments.

Software Implementation

Software implementation of the Mann-Whitney U test is widely supported in statistical packages such as SPSS, R, Stata, Python (SciPy), and Jamovi, enabling efficient comparison of two independent groups without assuming normality. These programs automatically perform rank ordering, compute the U statistic, adjust for ties, and generate exact or asymptotic p-values, thereby minimizing computational errors and enhancing analytical accuracy (Mann & Whitney, 1947). The Mann-Whitney U test is carried out by first pooling the data from the two independent groups and arranging all the values from the smallest to the largest. Each value is given a rank, and when values are the same, they share an average rank. The ranks belonging to each group are then added up and used to compute the U statistic. This statistic is converted into a p-value, which helps the researcher decide whether the difference between the two groups is statistically significant. The test is especially useful when the data do not meet the assumptions of normality or are measured on an ordinal scale (Mann & Whitney, 1947).

Interpretation of Results

Interpretation of results in the Mann-Whitney U test begins with examining the U statistic, Z value, and the associated p-value, which indicate whether there is a statistically significant difference between the two independent groups. The p-value is compared with the chosen level of significance (commonly $\alpha = 0.05$). If the p-value is less than α, the null hypothesis, that the distributions of the two groups are equal, is rejected, suggesting a meaningful difference between groups. Unlike parametric tests that focus on mean differences, the Mann-Whitney U test evaluates differences in rank distributions, making it more appropriate for ordinal or non-normally distributed data (Mann & Whitney, 1947). Beyond statistical significance, proper interpretation requires assessing group ranks and central tendencies, such as median scores and mean ranks, which help explain the direction of the difference. Researchers should compare the median or mean rank of each group to determine which group tends to have higher or lower values on the outcome variable. For example, a higher mean rank indicates that observations in that group generally have larger values. Reporting medians rather than means is recommended, as medians better represent central tendency in nonparametric analysis (Lotsch & Ultsch, 2020). Finally, interpretation should include an evaluation of effect size, which reflects the practical importance of the observed difference. Commonly reported effect sizes for the Mann-Whitney U test include the rank-biserial correlation or the standardized effect size (r), calculated from the Z statistic. According to Cohen et al. (2007) effect size values provide insight into whether the difference is small, moderate, or large, regardless of sample size. Also, Dennis and Somerville (2023) emphasis that incorporating effect size alongside p-values strengthens result interpretation and aligns with best practices in statistical reporting

Working Example: How to Perform the Mann-Whitney U Test Using SPSS

First and foremost, to perform the Mann-Whitney U test analysis, you must have two variables' data, for example, considering data of a researcher who compares mathematics performance between two independent groups of students (see *Table 3*):

Group A: Students taught the circle theorem using GeoGebra

Group B: Students taught the circle theorem using traditional methods

Table 3. Scores of Students taught using ethnomathematics and traditional methods

Group A (Students taught circle theorem using GeoGebra)	*Group B (students taught the circle theorem via the traditional method)*
78	65
85	70
88	68
90	72
76	66
84	74
73	82
73	62
82	77
95	37
84	75
85	78

a. State the null and alternative hypotheses

b. Rank all the scores together

c. Compute the Mann-Whitney U statistic

d. Interpret the result at $\alpha = 0.05$

Manual Calculation

a. Hypotheses

Null hypothesis (H0): There is no difference in mathematics performance between students taught using GeoGebra and those taught using the traditional method.

Alternative hypothesis (H1): There is a difference in mathematics performance between the two groups.

b. First, combine and rank all 24 scores (smallest = rank 1). For tied values, assign average ranks as shown in Table 4.

c. Maa-Whitey U Statistics

Sample size for individual groups

$$n_A = n_B = 12$$

Sum of rank for Group A

$$\sum R_A = 8.5 + 8.5 + 12 + 14.5 + 16.5 + 18.5 + 18.5 + 20.5 + 20.5 + 22 + 23 + 24 = 207$$

The computation for Maa-Whitey U Statistics is given as;

$$U_A = n_A n_B + \frac{n_A(n_A + 1)}{2} - \sum R_A$$

$$U_A = (12 \times 12) + \frac{12(12+1)}{2} - 207 = 15$$

$$U_B = n_A n_B - \sum R_A = (12 \times 12) - 15 = 129$$

For the test statistics (U), we chose the smaller value = $U_A = 15$

Table 4. Scores Ranking

Score	Group	Ranks
37	B	1
62	B	2
65	B	3
66	B	4
68	B	5
70	B	6
72	B	7
73	A	8.5
73	A	8.5
74	B	10
75	B	11
76	A	12
77	B	13
78	A	14.5
78	B	14.5
82	A	16.5
82	B	16.5
84	A	18.5
84	A	18.5
85	A	20.5
85	A	20.5
88	A	22
90	A	23
95	A	24

d. Interpretation at $\alpha = 0.05$

At the 0.05 level of significance (two-tailed), and with both groups having 12 participants, the critical value is approximately 37. This means that we reject the null hypothesis if the calculated Mann-Whitney U value is 37 or less. Since the obtained U value meets this condition, the null hypothesis is rejected. This means that there is a statistically significant difference in mathematics performance between the

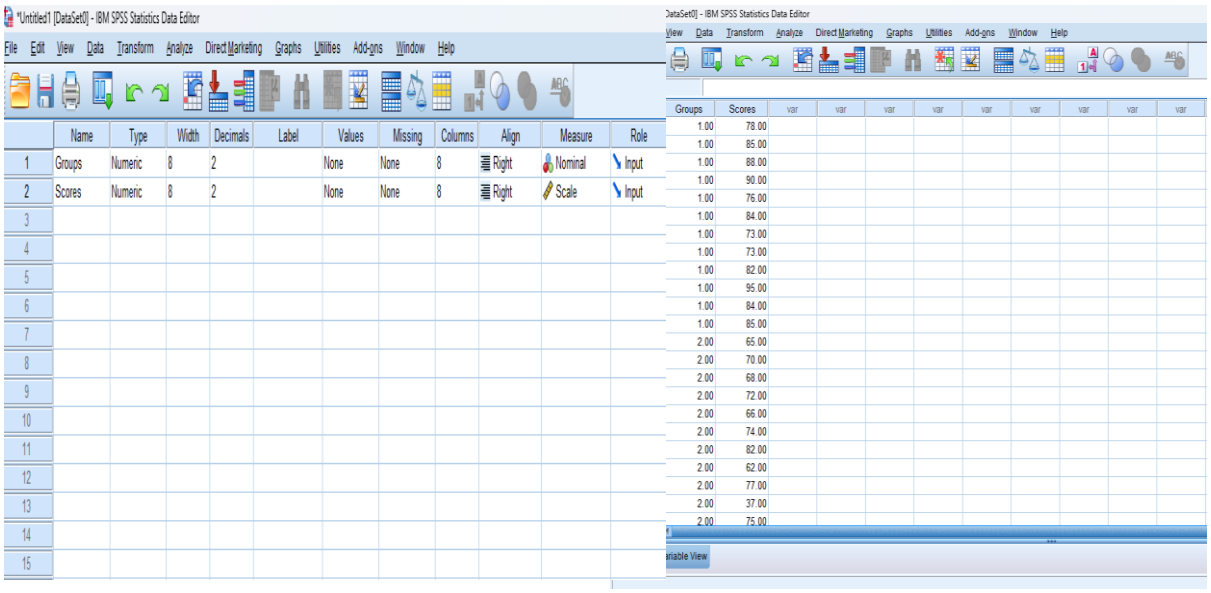
two groups at the 0.05 level. The results further explain that students taught circle theorem using GeoGebra (Group A) generally have high ranks, indicating better performance compared to those taught using traditional methods.

Using SPSS

Step 1: Enter the Data

First, open IBM SPSS Statistics and go to Variable View, where you create two variables: one called *Group* (set as numeric, with 1 representing Group A: students taught the circle theorem using GeoGebra and 2 representing Group B: students taught using traditional methods) and another called *Score* (also set as numeric). After that, move to Data View and enter the data in the appropriate columns, as shown in Figure 1.

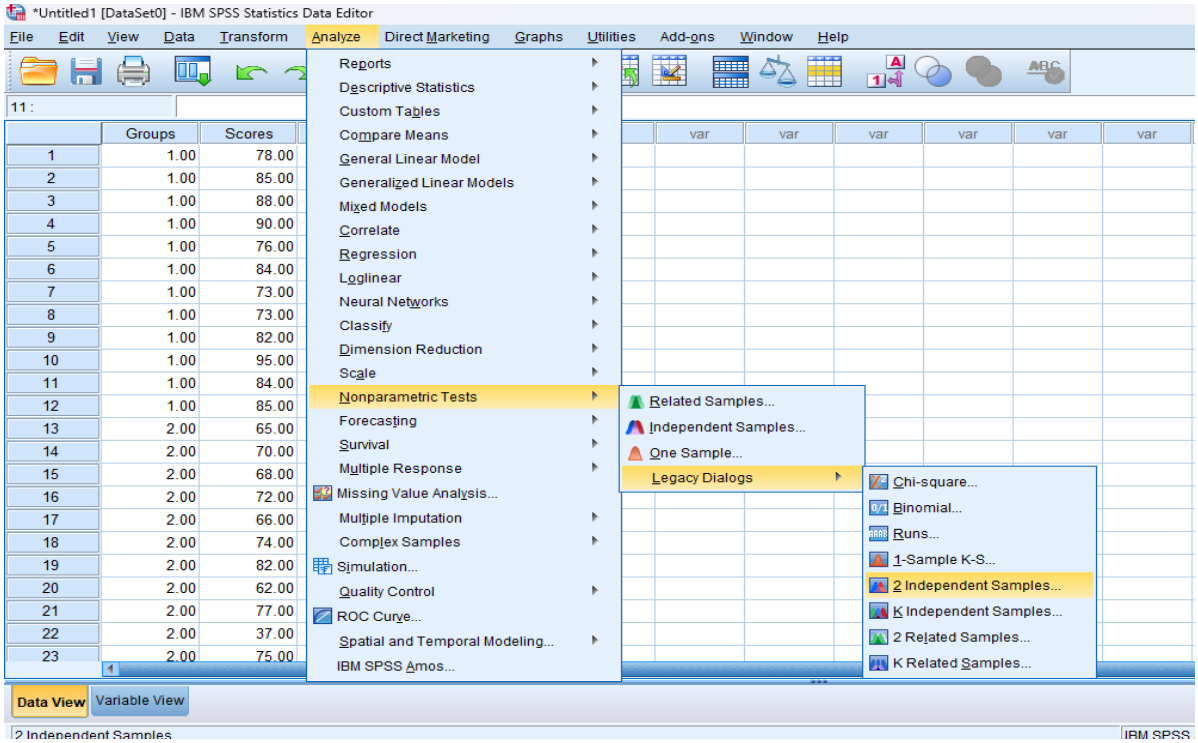
Figure 2. Data Entry in SPSS (ver. 23) (Source: Authors, 2026)



Step 2: Run the Mann-Whitney U Test

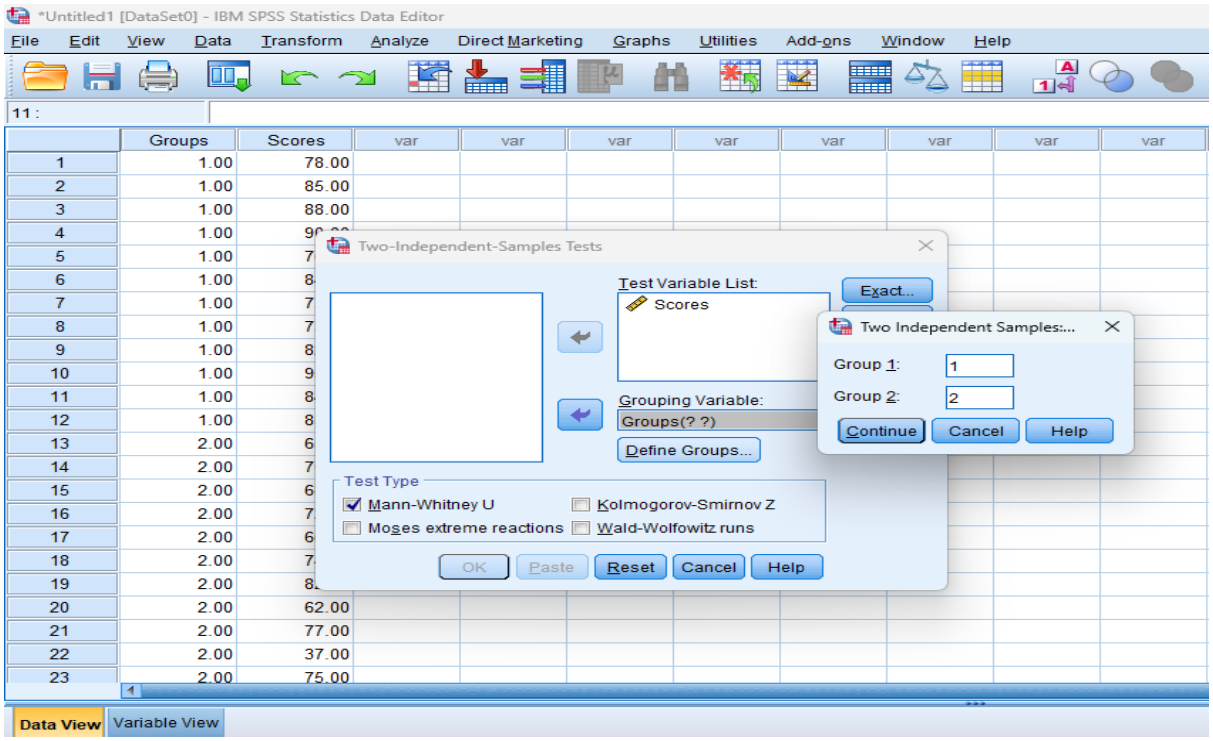
To perform the Mann-Whitney U test in SPSS, go to Analyze → Nonparametric Tests → Legacy Dialogs → 2 Independent Samples, as shown in Figure 3.

Figure 3. Steps for selecting 2-Independent Samples (Source: Authors' Creation, 2026)



Move *Scores* to the Test Variable List and *Groups* to the Grouping Variable, define the groups as 1 and 2, as shown in Figure 4. Click on "Continue."

Figure 4. Moving of Variables and Defining Groups (Source: Authors' Creation, 2026)



Click on “Option” and select the " Exclude cases listwise under Options, as shown in Figure 5, and finally click OK to run the analysis and display the results in Table 5.

Figure 5. Selecting Excluding Case Listwise (Source: Authors’ Creation, 2026)

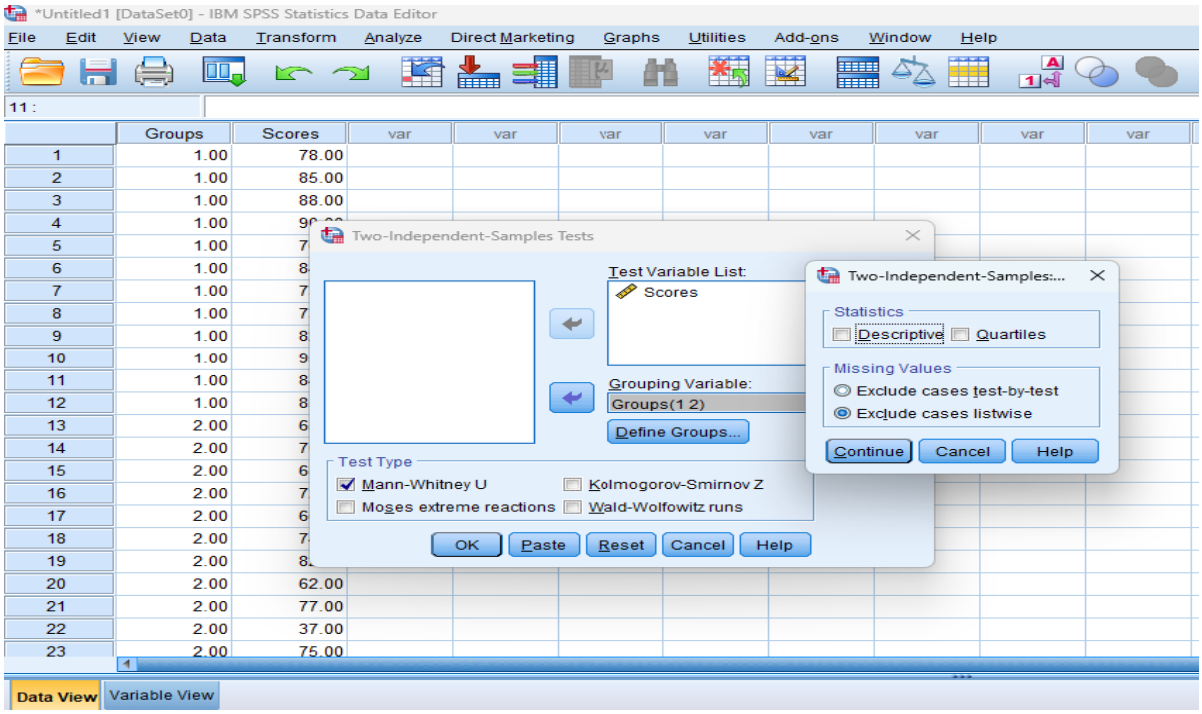


Table 5 shows the results from SPSS (version 23) for comparing students taught the circle theorem using GeoGebra (Group A) and those taught using the traditional method (Group B) with a Mann-Whitney U test. The Mann-Whitney U value is 15, which measures the difference in ranks between the two groups, and the Wilcoxon W value is 93, representing the sum of ranks for one group. The Z value of -3.294 shows how far the observed difference is from what we would expect if there were no real difference between the groups.

Table 5. Statistical Results from SPSS (ver. 23) for comparing students taught circle theorem using GeoGebra (Group A) and the Traditional Method (Group B)

Test Statistics	
	Scores
Mann-Whitney U	15.000
Wilcoxon W	93.000
Z	-3.294
Asymp. Sig. (2-tailed)	.001
Exact Sig. [2*(1-tailed Sig.)]	.000[b]

a. Grouping Variable: Groups
b. Not corrected for ties

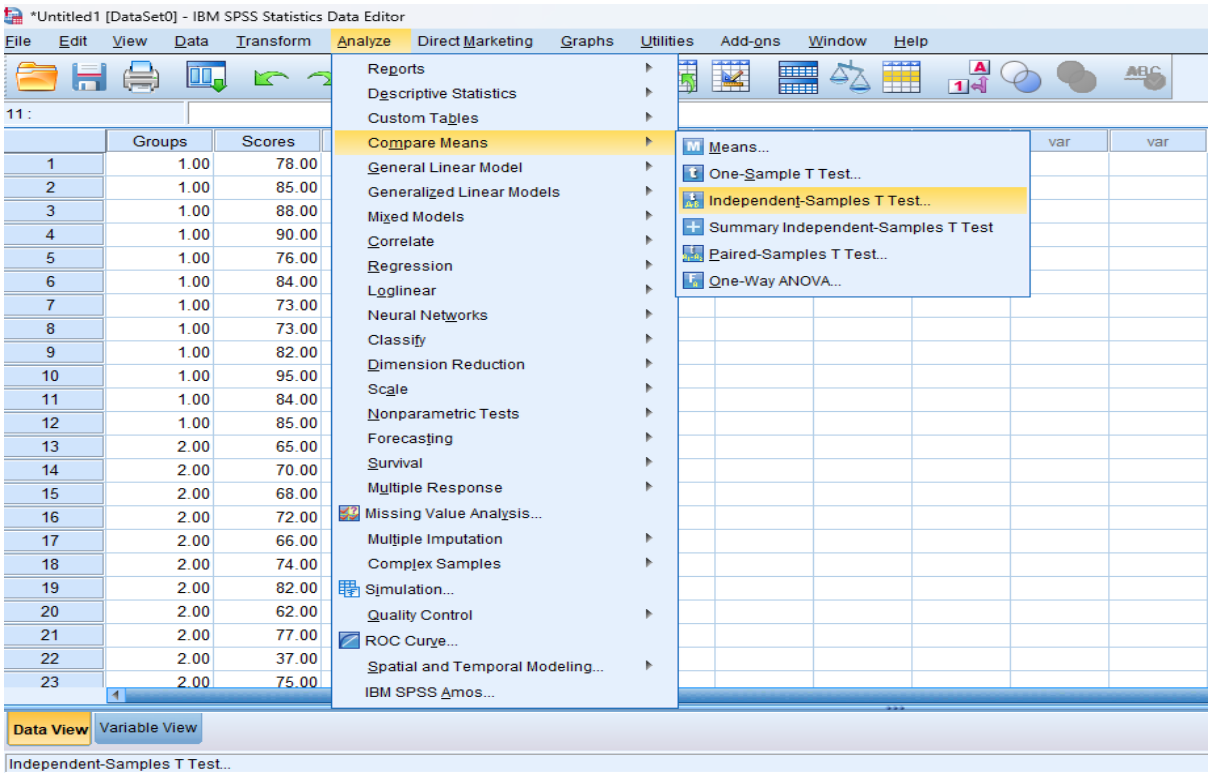
Moreover, the two-tailed p-value (Asymp. Sig.) is 0.001, which is less than 0.05, meaning the difference is statistically significant, and the exact significance value of 0.000 also supports this result.

In simple terms, this means that students taught using GeoGebra performed significantly better than those taught using the traditional method.

How to Perform an Independent Samples t-Test Via SPSS (ver. 23)

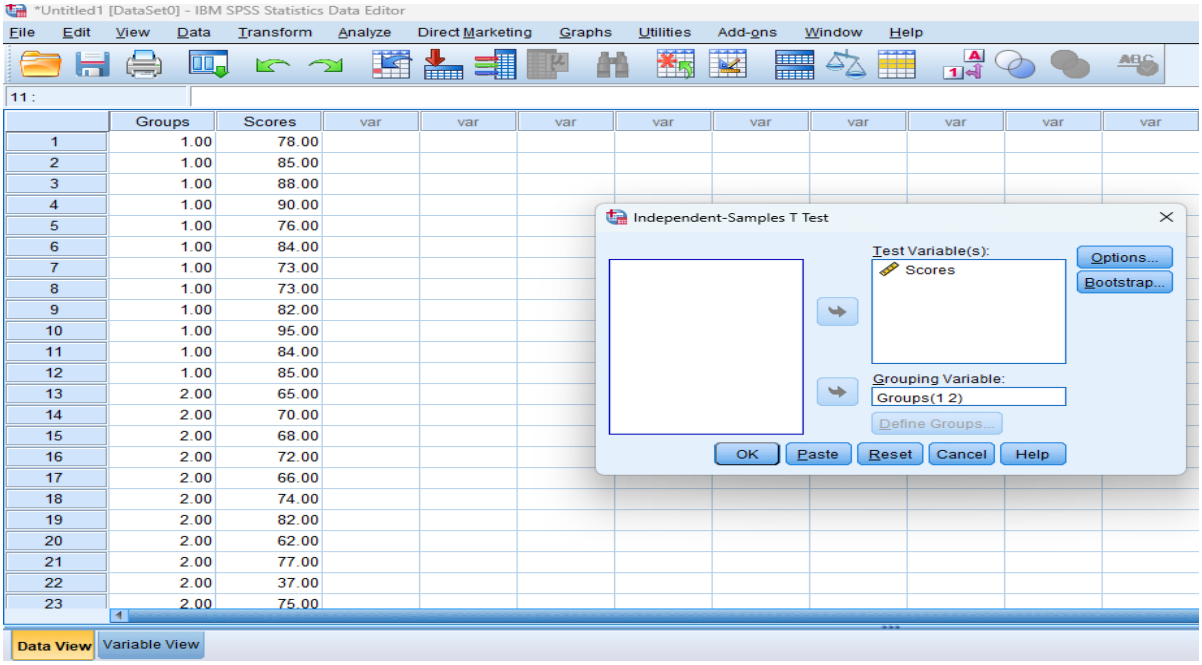
To conduct an independent samples t-test in IBM SPSS Statistics, click on Analyze, select Compare Means, and then choose Independent-Samples T Test, as shown in Figure 6.

Figure 6. Independent-Samples T Test Analysis Procedure (Source: Authors' Creation, 2026)



Next, move the dependent variable (*Scores*) into the Test Variable(s) box and the grouping variable (Groups) into the Grouping Variable box, as shown in Figure 7.

Figure 7. Moving Variables into the Test Variables and the Grouping Variables (Source: Authors' Creation,



2026)

After that, click on Define Groups, enter the group codes (1 for GeoGebra and 2 for the traditional method), and click Continue. Finally, click OK to run the test. The analysis results of the independent samples t-test are presented in Table 6.

The statistics result (see Table 6) revealed that there was a statistically significant difference between the two groups, t (22) = 3.588, p = 0.002. Levene's test for equality of variances was not significant (F = 0.811, p = 0.378), indicating that the assumption of equal variances was met. The mean difference between the groups was 13.92, with a 95% confidence interval ranging from 5.87 to 21.96, suggesting a meaningful difference in scores between the groups.

Comparison of the Independent Samples t-Test and the Mann-Whitney U test

To enhance the robustness and originality of the analysis, both parametric and non-parametric tests were conducted using IBM SPSS Statistics to compare differences between two independent groups. Table 7 compares the results of the independent samples t-test and the Mann-Whitney U test. The t-test, which focuses on comparing group means, showed a value of t = 3.588 with a p-value of 0.002, indicating a significant difference between the groups since the p-value is less than 0.05.

Table 6. Independent Samples t-test Result

	Levene's Test for Equality of Variances								
	F	Sig.	t	df	p	MD	Std. Error	95% CI Lower	Upper
Equal variances assumed	0.811	0.378	3.588	22	0.002	13.92	3.88	5.87	21.96
Equal variances not assumed			3.59	17.68	0.002	13.92	3.88	5.76	22.077

Table 7. Comparative Results of t-Test and Mann-Whitney U Test

Test Method	*Focus of Test*	*Test Statistic*	*Value*	*p-value*	*Decision (α = 0.05)*	*Conclusion*
Independent Samples t-Test	Difference in means	t	3.588	.002	Significant	There is a statistically significant difference
Mann-Whitney U Test	Difference in medians/ranks	U	15.000	.001	Significant	There is a statistically significant difference

In the same way, the Mann-Whitney U test, which compares median ranks and does not depend on normality assumptions, also showed a significant result (U = 15.000, p = 0.001). Because both tests produced significant results, it means that the difference between the groups is real and not due to chance. The fact that both the parametric and non-parametric tests agree makes the findings more reliable and strengthens confidence in the results.

DISCUSSION

The Mann-Whitney U test is a useful tool for comparing two independent groups, especially when real-world data don't meet the strict requirements of parametric tests like the t-test. Unlike parametric tests, it doesn't assume that the data are normally distributed or measured on an interval scale; it only requires ordinal or continuous data with independent observations (Emerson, 2023). This makes it particularly helpful in education, psychology, and health research, where small sample sizes, skewed data, or outliers are common. By focusing on ranks rather than raw scores, researchers can still detect meaningful differences between groups without the risk of errors that come from using inappropriate parametric methods. However, the test has some limitations. It assumes that the distributions of the two groups are similarly shaped, so if the groups differ a lot in spread or shape, a significant result may reflect differences in overall distribution rather than a true shift in the median (Karch, 2021). Also, when the data actually meet parametric assumptions, a t-test can be more powerful, which is why it's important to consider both the data and research goals before choosing a method. For researchers, the

Mann-Whitney U test offers a reliable way to analyze data that don't meet parametric assumptions, ensuring valid conclusions even with small or skewed samples. For educators, it provides a practical method to compare student groups, survey responses, or the effects of interventions without needing perfect normal data. By understanding when and how to use this test, both researchers and educators can make stronger, evidence-based decisions and better interpret the results of their analyses.

CONCLUSION

The Mann-Whitney U test is a reliable and practical method for comparing two independent groups, especially when the data are not normally distributed or are measured on an ordinal scale. Ranking the data instead of using raw values, it reduces the impact of outliers and skewed distributions, making it useful in fields like education, health, psychology, and social sciences. Its main strength is flexibility, but it also has limitations: it measures differences in ranks rather than exact medians, which can be misleading if the groups' distributions have very different shapes, and it may be less powerful than parametric tests when normality holds. Compared with t-tests, it works best for small samples or ordinal data, while t-tests are more appropriate for normally distributed interval data. In all, the Mann-Whitney U test provides a simple, effective way to examine group differences, as long as researchers interpret the results carefully, check assumptions, and report effect sizes for practical significance. Future studies should explore robust alternatives such as the Brunner-Munzel test in cases of unequal distributions.

Author contribution: All the authors were involved in concept, interpretation, writing, and critically revising the article. Authors approve the final version of the article.

Declaration of interest: The authors have no conflict of interest to declare regarding the content of this article.

Ethics declaration: This paper is a methodological study aimed at illustrating. The analysis is based on illustrative (or secondary) data and does not involve direct interaction with human participants. Consequently, ethical approval was not required for this study.

Funding: No funding source is reported for this study.

REFERENCES

Abreh, M. K., Okyere, M., Osiakwan, J., & Adom, G. (2025). Navigating the Use and Misuse of Likert Scales in Counseling and Development Research. *Measurement and Evaluation in Counseling and Development*, *58*(4), 347-353. https://doi.org/10.1080/07481756.2025.2499964

Bakker, M., & Wicherts, J. M. (2014). Outlier removal, sum scores, and the inflation of the type I error rate in independent samples t tests: The power of alternatives and recommendations. *Psychological Methods*, *19*(3), 409-427. https://doi.org/10.1037/met0000014

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. In *BioData Mining* (Vol. 18, Issue 1). BioMed Central. https://doi.org/10.1186/s13040-025-00465-6

Cohen, L., Manion, L., & Morrison, K. (2007). *Research methods in education* (6th ed.). London & New York: Routledge.

Dehaene, H., De Neve, J., & Rosseel, Y. (2021). A Wilcoxon–Mann–Whitney Test for Latent Variables. *Frontiers in Psychology*, *12*(November). https://doi.org/10.3389/fpsyg.2021.754898

Dennis, J. L., & Somerville, M. P. (2023). Supporting thinking about thinking: examining the metacognition theory-practice gap in higher education. *Higher Education*, *86*(1), 99–117. https://doi.org/10.1007/s10734-022-00904-x

Gibbons, J. D., Chakraborti, S., Taylor, P., & Chakraborti, S. (1991). Comparisons of the Mann-Whitney, Student's t, and Altemate t Tests for Means of Normal Distribution. *The Journal of Experimental Education*, *59*(3), 258-267.

Khan, A., Ahmad, F. H., & Malik, M. M. (2017). Use of digital game-based learning and gamification in secondary school science: The effect on student engagement, learning, and gender difference. *Education and Information Technologies*, *22*(6), 2767-2804. https://doi.org/10.1007/s10639-017-9622-1

Lin, T., Chen, T., Liu, J., & Tu, X. M. (2021). Extending the Mann-Whitney-Wilcoxon rank sum test to survey data for comparing mean ranks. *Statistics in Medicine*, *40*(7), 1705-1717. https://doi.org/10.1002/sim.8865

Liu, Y., & Xie, J. (2020). Cauchy Combination Test: A Powerful Test With Analytic p-Value Calculation Under Arbitrary Dependency Structures. *Journal of the American Statistical Association*, *115*(529), 393-402. https://doi.org/10.1080/01621459.2018.1554485

Lotsch, J., & Ultsch, A. (2020). A non-parametric effect-size measure capturing changes in central tendency and data distribution shape. *PLoS ONE*, *15*(9 September), 1-19. https://doi.org/10.1371/journal.pone.0239623

Mann, H. B., & Whitney, D. R. (1947). On a Test of Whether One of Two Random Variables is Stochastically Larger than the Other. *Institute of Mathematical Statistics Stable*, *18*(1), 50-60. https://about.jstor.org/terms

Nachar, N. (2008). The Mann-Whitney U: A Test for Assessing Whether Two Independent Samples Come from the Same Distribution. *Tutorials in Quantitative Methods for Psychology*, *4*(1), 13-20. https://doi.org/10.20982/tqmp.04.1.p013

Norman, G. (2010). Likert scales, levels of measurement, and the "laws" of statistics. *Advances in Health Sciences Education*, *15*(5), 625–632. https://doi.org/10.1007/s10459-010-9222-y

Okeh, U. M. (2009). Statistical analysis of the application of the Wilcoxon and Mann-Whitney U test in medical research studies. *Biotechnology and Molecular Biology Reviews*, *4*(6), 128-131. http://www.academicjournals.org/bmbr

Olvera, O. L. (2020). Issues, problems, and potential solutions when simulating continuous, non-normal data in the social sciences. *Meta-Psychology*, *4*(Mcmc). https://doi.org/10.15626/mp.2019.2117

Pérez, N. P., Guevara López, M. A., Silva, A., & Ramos, I. (2015). Improving the Mann-Whitney statistical test for feature selection: An approach in breast cancer diagnosis on mammography. *Artificial Intelligence in Medicine*, *63*(1), 19-31. https://doi.org/10.1016/j.artmed.2014.12.004

Rutkowski, I. P. (2025). Stevens' measurement scales in marketing research – A continuation of the discussion on whether researchers can ignore the Likert scale's limitations as an ordinal scale. *Marketing of Scientific and Research Organizations*, *55*(1), 39-55. https://doi.org/10.2478/minib-2025-0003

Sedgwick, P. (2015). A comparison of parametric and non-parametric statistical tests. *BMJ (Online)*, *350*(April), 1-2. https://doi.org/10.1136/bmj.h2053

Sihombing, P. R. (2025). Methodological Comparison of Frequentist and Bayesian Two-Sample Tests on Regional Socioeconomic Disparities in Indonesia. *Forum Analisis Statistik*, *5*(2), 60-70.

Vickers, A. J. (2005). Parametric versus non-parametric statistics in the analysis of randomized trials with non-normally distributed data. *BMC Medical Research Methodology*, *5*, 1-12. https://doi.org/10.1186/1471-2288-5-35

Wall Emerson, R. (2023). Mann-Whitney U test and t-test. *Journal of Visual Impairment and Blindness*, *117*(1), 99-100. https://doi.org/10.1177/0145482X221150592

Wang, Z., Liu, H., & Fan, X. (2025). Integrating machine learning and statistical testing for studying emotional influence in online reviews. In *Business Process Management Journal*. https://doi.org/10.1108/BPMJ-12-2024-1245

White, N. M., Balasubramaniam, T., Nayak, R., & Barnett, A. G. (2022). An observational analysis of the trope "A p-value of < 0.05 was considered statistically significant" and other cut-and-paste statistical methods. *PLoS ONE*, *17*(3 March), 1-15. https://doi.org/10.1371/journal.pone.0264360

Wilcox, R. R. (2005). Comparing medians: An overview plus new results on dealing with heavy-tailed distributions. *Journal of Experimental Education*, *73*(3), 249-263. https://doi.org/10.3200/JEXE.73.3.249-263

Wilcoxon, F. (1945). Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, *1*(6), 80. https://doi.org/10.2307/3001968

Wonu, N., & Anaekwe, E. (2023). Application of the Mann-Whitney U-test in the Analysis of Gender-based Leadership and Collaborative Scholarly Writing Performance. *FNAS Journal of Mathematical and Statistical Computing*, *1*(1), 29-39. www.fnasjournals.com

Zarrin, M., Abyani, M., & Asgarian, B. (2020). A statistical study on lognormal central tendency estimation in probabilistic seismic assessments. *Structure and Infrastructure Engineering*, *16*(5), 803-819. https://doi.org/10.1080/15732479.2019.1668813

Please Cite: Asare, B., Arthur, Y. D., Obeng, B. A., Alija, S., Orhani, S. (2026). Mann-Whitney U Test for Two Independent Groups. *Journal of Research in Didactical Sciences,* 5(1), 13-31. https//doi.org/10.51853/jorids/512

Received: 07.12.2025 ▪ Accepted: 05.04.2026