Studies in Second Language Acquisition (2026), **48**, 535–570
doi:10.1017/S0272263126101600

STATE OF THE SCHOLARSHIP

Univariate normality checking practices in L2 research: An AI-assisted systematic review

Vahid Aryadoust and Yichen Jia

National Institute of Education, Nanyang Technological University, Singapore
Corresponding author: Vahid Aryadoust; Email: vahid.aryadoust@nie.edu.sg

(Received 14 July 2025; Revised 09 January 2026; Accepted 19 January 2026)

Abstract

Regardless of whether one is analyzing quantitative data from research involving generative artificial intelligence (GenAI) or more classical methods, testing for normality remains a necessary step in statistical analysis. Although over 60 methods have been proposed for assessing univariate normality, previous systematic reviews show that normality testing remains underreported in L2 research. This paper addresses this gap by first reviewing the concept of normality and its role in parametric statistical inference. We then examine 12 normality assessment methods including five graphical and seven analytical methods selected based on their prominence in statistical literature and availability in commonly used software. Each method is explained in terms of its underlying mechanism and sensitivity to specific forms of nonnormality, such as skewness, tail heaviness, and multimodality. In the second part of the study, we review 237 empirical articles published between 2020 and 2025 in ten selected L2-focused Q1 journals, using AI-assisted annotation. Our findings reveal inconsistencies in how graphical tests are reported, a tendency to rely on tests such as the Kolmogorov-Smirnov without explicit attention to sample size constraints, and limited justification provided for critical values of skewness and kurtosis. These results indicate some divergence between recommended statistical practices and the procedures for normality testing reported in the L2 publications examined. The paper concludes with actionable recommendations for selecting and interpreting normality tests in L2 research contexts.

Keywords: analytical tests; graphical methods; normality assessment; second language research; univariate analysis

Introduction

Normality in statistical analysis refers to the (probability) distribution of a dataset characterized by a symmetric bell-shaped curve, where observations cluster around the mean and taper symmetrically toward both tails (Hair et al., 2019, p.48). Normality of

© The Author(s), 2026. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/S0272263126101600 Published online by Cambridge University Press

the error[1] (rather than the raw data; Pek et al., 2018) is a prerequisite of many statistical analyses, such as the *t*-test, ANOVA, correlation analyses, and linear regression, which determines the application of parametric methods or nonparametric counterparts (Razali & Wah, 2011; Uhm & Yi, 2021). Given the significant role of normality, it has attracted great attention from statistical researchers, with over 60 normality checking methods being proposed (Hernandez, 2021) and new approaches continuing to be developed (e.g., Bayoud, 2021).

While the diversity of methods provides researchers with various options, it also poses challenges. Normality tests vary significantly in their implementation complexity and reliability. Some are purely descriptive and require a researcher to visually assess whether a distribution follows a bell-shaped curve (e.g., Histogram), but they lack universal criteria for justifying normality (Wijekularathna et al., 2019). In contrast, analytical methods such as Shapiro-Wilk and Kolmogorov-Smirnov tests provide quantified results but can be complex and less intuitive or might not be readily available in some data analysis software (see Yap & Sim, 2011, p. 2142). Furthermore, these methods focus on different aspects of normality, with their sensitivity influenced by the sample size (Kim, 2013). Without understanding the underlying logic of normality tests, their differences, and the properties of nonnormality, researchers may struggle to identify which method best aligns with their specific needs.

Research shows that despite the significance and diversity of normality checking methods, researchers in the field of second language (L2) research have paid insufficient attention to this critical assumption (Hu & Plonsky, 2021; Lindstromberg, 2016; Plonsky & Ghanbar, 2018). Lindstromberg (2016) reported that only 21 out of 96 quantitative studies reported checking underlying statistical assumptions. Among these, normality was the most frequently tested, yet details about how it was assessed were not always provided. In another synthesis study, Plonsky and Ghanbar (2018) found that only 15.53% L2 studies using multiple regression assessed normality. Hu and Plonsky (2021) further reviewed 107 studies from 2012 to 2017 in *Language Learning* and *Second Language Research* that applied analyses requiring normal distribution assumptions (i.e., independent sample *t*-test (K = 38), one-way ANOVA (K = 42), Pearson correlation (K = 37), multiple regression (K = 17)). Although normality was one of the two most frequently assessed assumptions, it accounted for only 6.7% of all cases. Such a low checking rate is not unique to linguistics but is also true for psychology (Ernst & Albers, 2017), education (Shatz, 2023), and medical research (Nielsen et al., 2019). The limited reporting of assumption checking across multiple fields threatens research transparency, which is fundamental to study quality evaluation and reproducibility (Plonsky, 2024).

While previous systematic reviews have offered valuable insights into how frequently the normality assumption is examined in L2 research, to our knowledge, no L2 study has reviewed the available normality assessment methods or evaluated how accurately these methods are being applied. The present study, rather than reiterating the adequacy of normality checking practices documented in previous research, focuses on the nature of univariate normality. Specifically, we examine common univariate methods for checking normality, including five graphical and seven analytical methods

[1]According to Pek et al. (2018, p. 2), in the general linear model including ANOVA and linear regression, "[i]t would be inappropriate to examine the distribution of $y = (y_1,\ldots, y_N)'$ independent of the *K* predictors because the distributional assumption is about $\boldsymbol{\varepsilon}$ [error]. It is only under the intercept-only model, where the distributional properties of y are identical to $\boldsymbol{\varepsilon}$, that in this vein, the distribution of y is analogous to *e*." Also see Pek et al. (2017, pp. 3–4) for further information.

https://doi.org/10.1017/S0272263126101600 Published online by Cambridge University Press

frequently reported in statistical research and available in popular GUI-based (graphical user interface) statistical analysis software or R. We examine how various normality tests function and evaluate their capacity to detect specific forms of deviation from normality, such as skewness, bimodality, and variations in tail heaviness. Moreover, building on Hu and Plonsky's (2021) work, we broaden the scope from two L2-focused journals to ten Q1 L2-focused journals, and review 237 empirical studies published between 2020 and 2025 that documented normality testing procedures. Rather than calculating the proportion of studies that conducted normality checks, we focus on analyzing the particular tests used, the criteria adopted, and the degree to which researchers followed guidelines recommended by statisticians. We conclude by summarizing common practices currently used in L2 research and suggesting considerations for selecting normality checking methods appropriate to different research contexts.

Literature review

Meaning of univariate normality

Univariate normality refers to the extent to which the distribution of sample data corresponds to a normal distribution, which is typically represented as a bell-shaped curve with data values on the x-axis and their frequencies on the y-axis (Hair et al., 2019, p. 48; Thode, 2002). Three key features characterize data distribution: asymmetry, tail weight, and peak characteristics (including number and shape), where sample size potentially moderates these properties (Cain et al., 2017). Asymmetry is quantified through skewness, which measures the distribution's deviation from symmetry around its mean value (Field, 2018). A distribution with a longer or fatter tail on the left is considered negatively skewed, while one with a longer or fatter tail on the right is positively skewed. Tail weight, referring to the relationship between the central peak and the distribution's tails, is quantified by kurtosis. Positive kurtosis indicates heavier-than-normal tails on both sides, while negative kurtosis suggests lighter-than-normal tails (Field, 2018, pp. 344–345). Moreover, the presence of multiple peaks, even when individual peaks exhibit bell-shaped characteristics, can still make the overall distribution deviate from normality and potentially affect statistical analyses (Ventura-León et al., 2023).

In addition to distribution characteristics, sample size plays a critical role in normality assessment (Hanusz et al., 2017; Khatun, 2021). Small samples may pass a normality test with little power, while large samples may flag minor deviation from normality as significantly nonnormal (Öztuna et al., 2006). The Central Limit Theorem (CLT) addresses this challenge by establishing that sampling distributions of means approximate normality as sample size increases, regardless of the underlying population distribution (Altman & Bland, 1995; Kwak & Kim, 2017). However, debate exists regarding the minimum sample size required for CLT application. Some researchers suggest that the sampling distribution can be assumed normal at $K = 30$ (Field, 2018, p. 111; Weinberg & Abramowitz, 2002, p. 276), while others recommend continued normality testing and correction for highly skewed data even with samples exceeding 100 (Wilcox, 2010, pp. 63–85). A more conservative approach suggests that CLT assumptions are appropriate only when samples exceed 200 and the subsequent analyses can accommodate potential normality violations (Demir, 2022; Hair et al., 2019, p. 219). These considerations highlight the importance of interpreting normality test results in light of sample size.

https://doi.org/10.1017/S0272263126101600 Published online by Cambridge University Press

Significance of normality

Normality is one of the fundamental assumptions underlying parametric inferential statistics (Aryadoust & Raquel, 2019; Field, 2018, p. 331; Thode, 2002, p. 2). In a perfectly normal distribution, the mean, median, and mode coincide (are equal), and both skewness and "excess kurtosis" are equal to zero (Kim, 2013; West et al., 1995). Many GUI statistical packages used in L2 research, such as SPSS, Jamovi, and JASP, report excess kurtosis, which is computed by subtracting 3 from the standard kurtosis (kurtosis proper) index (Kim, 2013). According to West et al. (1995), the absolute value of kurtosis must be below 7 to avoid substantial deviations from normality, although this threshold refers to standard kurtosis, not excess kurtosis. Excess kurtosis specifically measures the heaviness of the tails of a distribution relative to the normal distribution. A positive excess kurtosis indicates a distribution with heavier tails and a sharper peak, while a negative value suggests lighter tails and a flatter peak. As Kim (2013) notes, an excess kurtosis value substantially above or below zero may indicate that the data are not normally distributed and could affect the robustness of parametric statistical analyses that assume normality.

Although nonparametric alternatives are available for situations where the assumption of normality is violated, these methods generally have lower statistical power than their parametric counterparts when applied to data that are normally distributed (Keren & Lewis, 2014). Therefore, it may not always be appropriate to default to nonparametric tests to avoid normality testing (Le Cessie et al., 2020; Parsons et al., 2012). Instead, for moderately nonnormal datasets, correction or transformation methods should be employed to fit parametric analyses (Pek et al., 2018). For example, in their review of 61 statistics textbooks, Pek et al. (2018) identified a range of recommended strategies for addressing nonnormality, including data transformations, heteroscedasticity-consistent standard errors, bootstrap-based methods, and likelihood-based adjustments. Among these, power transformations—such as the Box-Cox and Yeo-Johnson transformations—are commonly applied to stabilize variance and improve the approximation to normality, particularly when moderate skewness or kurtosis is present (Raymaekers & Rousseeuw, 2024). More recent contributions, such as those by Raymaekers and Rousseeuw (2024), have proposed robust methods for estimating transformation parameters in the presence of outliers, which help preserve the central tendency of the data without unduly distorting its tails.

The importance of normality extends beyond the choice between parametric and nonparametric analyses (Vrbin, 2022). It influences parameter estimation, confidence intervals, and hypothesis testing (Field, 2018, p. 331). A recent study by Ventura-León et al. (2023) demonstrated how various nonnormal distributions affect correlation coefficient estimation. They investigated bimodal nonnormality (with outliers at means of 3 or 6), distributions with varying degrees of skewness (0 to 2.5), and kurtosis (–1.39 to 25), analyzing their impact on Pearson, Spearman, and Pearson Winsorized correlation coefficients (a robust version of the Pearson correlation designed to reduce the influence of outliers) across sample sizes (between 50 and 1000). Their findings revealed large discrepancies in bimodal distributions with outliers (mean = 6) for Spearman coefficients across all sample sizes (Ventura-León et al., 2023, p. 414). Although this study did not examine highly skewed distributions, which limits its representation of nonnormality, it illustrates two critical aspects of normality. On the one hand, it demonstrates how certain types of departure from normality can significantly affect statistical outcomes, which supports the premise that the validity of

https://doi.org/10.1017/S0272263126101600 Published online by Cambridge University Press

inferential results depends on how well data approximate a normal distribution (Uhm & Yi, 2023). On the other hand, it shows that not all departures from normality can substantially impact statistical analyses, as some non-normal distributions produce no noticeable differences in coefficient estimates (Ventura-León et al., 2023). It aligns with previous findings that suggest some statistical approaches can effectively handle moderate normality violations, particularly with adequate sample sizes (Cain et al., 2017; Knief & Forstmeier, 2021). For example, in very large samples, it is often unnecessary for the errors of the general linear model (e.g., ANOVA and linear regression) to follow a strictly normal distribution due to the CLT. As Pek et al. (2018, p. 4) explained:

> [r]egardless of the distribution of ε [errors in the general linear model], the CLT assures that the sampling distribution of the estimates will converge toward a normal distribution as N increases to infinity, when ε are independent and identically distributed, and when σ^2 is finite.

Thus, for sufficiently large samples, the sampling distribution of parameter estimates tends to be normal even if the error terms themselves are not, thereby preserving the validity of inferential tests under standard linear modeling frameworks.

These findings suggest that while normality is a crucial statistical assumption that could influence statistical validity, it should not be treated as a binary criterion for choosing between parametric and nonparametric analyses, nor should deviation from perfect normality be regarded as an "evil" that necessarily leads to bias in parameter estimations (Knief & Forstmeier, 2021). Moreover, it is important to understand the nature of the data and the underlying nature of the test. For instance, the correlations mentioned above can only detect monotone tendencies, meaning that they may fail to reveal the true relationship if the data are entrenched with nonmonotone tendencies; for such cases, more recent correlation alternatives can be used (Chatterjee, 2021) to reveal the true relationship regardless of the distribution of the data. Although a comprehensive discussion of these considerations extends beyond our study's scope, the severity of normality violations and the robustness of analytical methods to such violations are factors that may inform how results are interpreted and whether alternative analyses are appropriate. This context underscores the importance of our study, which investigates the functioning and sensitivity of various normality tests in detecting different forms of nonnormality. The analysis provides insight into how these tests operate and the implications of their outcomes for subsequent statistical procedures.

Univariate normality tests

Given the significant role of normality in statistics, researchers have proposed over 60 normality assessment approaches (Hernandez, 2021, pp. 3–11), which can be categorized into graphical methods, tests based on the empirical cumulative distribution function (ECDF), tests based on regression and correlation (especially using order statistics), tests based on moments (such as skewness and kurtosis), and a range of other analytical methods (see Table 1). These five categories can be broadly grouped into two overarching types: graphical methods and analytical tests (Field, 2018; Hair et al., 2019).

Extensive research has been conducted to identify the "best" analytical normality tests across diverse distributions and sample sizes (e.g., Arnastauskaitė et al., 2021; Patrício et al., 2017; Thode, 2002). In the comprehensive review summarized in Table 1,

https://doi.org/10.1017/S0272263126101600 Published online by Cambridge University Press

Table 1. Overview of 61 normality tests categorized by methodological type and primary references presented by Hernandez (2021, pp. 3–11)

Test abbreviation	Test name and category
	Graphical method
HIST	Histogram
STEM	Stem-and-leaf Plot
BOX	Box-and-whisker Plot
CDF	Cumulative Distribution Function Plot
PPP	P-P Plot (percent-percent plot)
QQP	Q-Q Plot (quantile-quantile plot)
	Empirical cumulative distribution function
AD	Anderson-Darling
KS	Kolmogorov-Smirnov
LF	Lilliefors
CVM	Cramer-von Mises
WA	Watson
AJN	Ajne
ZA	Zhang-Wu ZA
KU	Kuiper
GLB	Glen-Leemis-Barr
ZC	Zhang-Wu ZC
FRO	Frosini
GT	G Test
ND	N-distanceleft
	Moments
HLM	Hosking L-moments
BHSBS	Brys-Hubert-Struyf-Bonett-Seier
AJB	Adjusted Jarque-Bera
JB	Jarque-Bera
PDAB	Pearson-D'Agostino-Bowman
DAK	D'Agostino-Pearson Kurtosis
DAP	D'Agostino-Pearson K2
BS	Bonett-Seier
DAS	D'Agostino-Pearson Skewness
DL	Desgagné-Lafaye
DH	Doornik-Hansen
RJB	Robust Jarque-Bera
BM	Bontemps-Meddahi
DDL	Directional Desgagné-Lafaye
BHS	Brys–Hubert–Struyf
CCK	Cabaña-Cabaña Kurtosis
TLM	Trimmed L-moments
CCS	Cabaña-Cabaña Skewness
	Regression and correlation
SW	Shapiro-Wilk
SF	Shapiro-Francia
CS	Chen-Shapiro
CO	Coin
WB	Weisberg-Bingham
RG	Rahman-Govindarajulu
FB	Filliben
HG2	Hegazy-Green 2
RJ	Ryan-Joiner
ROY	Royston
DAD	D'Agostino D
BCMR	del Barrio-Cuesta-MatrÃ¡n-RodrÃguez

(*Continued*)

https://doi.org/10.1017/S0272263126101600 Published online by Cambridge University Press

Table 1. (*Continued*)

Test abbreviation	Test name and category
ZH	Zhang
GK0	Gan-Koehler k0
GK	Gan-Koehler k2
DWV	De Wet-Venter
HG1	Hegazy-Green 1
	Others
VA	Vasicek
GMG	Gel-Miao-Gastwirth
CHI^2 (χ^2)*	Chi-squared
GE	Geary
EP	Epps-Pulley
SH	Spiegelhalter
MI	Martinez–Iglewicz

*Although the Chi-squared (Chi^2/χ^2) test can be configured to assess normality, it is a general goodness-of-fit procedure rather than a test specifically designed for this purpose. Its performance depends strongly on how class intervals are defined and on sample size, which limits its usefulness for evaluating normality in some applications. Thus, despite Hernandez (2021, pp. 3–11) mentioning it among normality tests, we consider it tangentially relevant to normality assessment.

Hernandez (2021) analyzed the statistical power of 55 analytical normality tests and identified the Shapiro-Wilk (SW) test as the most frequently used and effective approach. However, this conclusion needs careful interpretation. First, as Hernandez (2021) noted, the high ranking of the SW test may be attributed to its frequent inclusion in these studies. Second, test performance varies with dataset characteristics, which can be influenced by sample size and distribution type (Nosakhare & Bright, 2017; Wijekularathna et al., 2019). Thus, combining or averaging results from studies that differ in distributional characteristics and sample sizes may lead to potentially misleading conclusions.

Current comparative studies typically evaluate numerous normality tests (e.g., Arnastauskaitė et al., 2021, evaluating 31 tests) across multiple non-normal distributions (e.g., Jiménez-Gamero, 2023, evaluating 12 distributions). While existing studies provide valuable insights into the performance of normality tests under various conditions, opportunities remain to extend this work. Most studies focus on reporting test statistics and distribution parameters. This could be complemented by a deeper examination of the underlying principles and how specific distributional characteristics influence test performance. Some researchers have begun categorizing non-normal distributions based on features such as tail weights (Jelito & Pitera, 2021) or symmetry (Uhm & Yi, 2021), which opens promising avenues for more systematic investigation. Building on these foundations, a clearer understanding of distributional characteristics and test sensitivities would help L2 researchers and practitioners make more informed choices when selecting suitable normality tests. Additionally, although graphical normality tests present interpretive challenges (Wijekularathna et al., 2019), they remain an underexplored aspect of normality assessment and could be used alongside quantitative methods to provide a more complete understanding of distributional characteristics.

To address these limitations, we examine both graphical and frequently used analytical normality tests, where we explain their underlying mechanisms. We selected analytical tests based on Hernandez's (2021) and Thode's (2002) studies, focusing on

https://doi.org/10.1017/S0272263126101600 Published online by Cambridge University Press

those appearing in more than 10 out of 20 studies, including Shapiro-Wilk (SW), Shapiro-Francia (SF), Kolmogorov-Smirnov (KS), Lilliefors, Cramer-von Mises (CVM), Anderson-Darling (AD), and Jarque-Bera (JB), alongside skewness and kurtosis measures.

Graphical measures

Graphical measures for assessing normality can be categorized into two main approaches: those based on raw data or frequency visualization such as Histograms, and those employing goodness-of-fit (GoF) principles to compare sample distributions against theoretical normal distributions such as quantile-quantile (Q-Q) plots (Thode, 2002).

Raw-data-based graphical approaches. Raw-data-based methods comprise three commonly used approaches: Histograms, Stem-and-leaf plots (Tukey, 1977), and boxplots (box-and-whisker plots) (Tukey, 1977).

Histograms. Histograms display frequency distributions by grouping values into contiguous intervals, or bins, along the x-axis, with corresponding bars representing the frequency of observations on the y-axis. This visual method enables a quick assessment of the distribution's shape, including its symmetry, skewness, modality, and potential outliers. A bell-shaped histogram suggests normality, whereas noticeable skew, flatness, or multiple peaks may indicate deviations from a normal distribution. Although histograms can be applied to datasets of various sizes, their interpretability is highly sensitive to bin width, which can either obscure or exaggerate distributional features if not appropriately chosen (Correll et al., 2019; Indira et al., 2011). As such, histograms tend to produce more stable and informative visualizations with larger samples, typically when the sample size reaches approximately 200 or more observations (Neter et al., 2005; Oppong & Agbedra, 2016).

Stem-and-leaf plots. Stem-and-leaf plots offer a more detailed visualization by organizing individual data points as "leaves" attached to categorical "stems." Unlike histograms that represent frequency through bar height, this method preserves and displays individual values, which enables more granular data examination (Oppong & Agbedra, 2016). The distribution's shape can be interpreted by observing the symmetry of the leaves around the center; a roughly symmetric spread suggests normality, while a concentration of leaves on one side indicates skewness. Gaps or clusters may reveal data irregularities, and repeated values are easily spotted. This makes stem-and-leaf plots particularly useful for identifying patterns, outliers, and the central tendency in small to moderately sized datasets. However, its readability decreases with larger datasets and may remove data points to maintain clarity (Ajao et al., 2018).

Boxplots. Boxplots (box-and-whisker plots), which also utilize raw data, summarize data dispersion through a combination of box and whiskers. They display the mean, quartiles, range, and potential outliers, which is effective for dataset comparison (Mishra et al., 2019). However, because boxplots visualize summarized data using five key statistics (minimum, quartile 1 (Q1), median, quartile 3 (Q3), maximum) rather than individual data points, they can hardly show specific distribution or frequency information of the dataset (see Figure 1)—a quartile divides an ordered dataset into four equal parts, each containing 25% of the data. Moreover, their utility may be limited with small datasets, where single outliers can disproportionately influence quartiles calculation and whiskers position, and potentially distort the perceived distribution (Ahmadiyeh et al., 2023).

https://doi.org/10.1017/S0272263126101600 Published online by Cambridge University Press

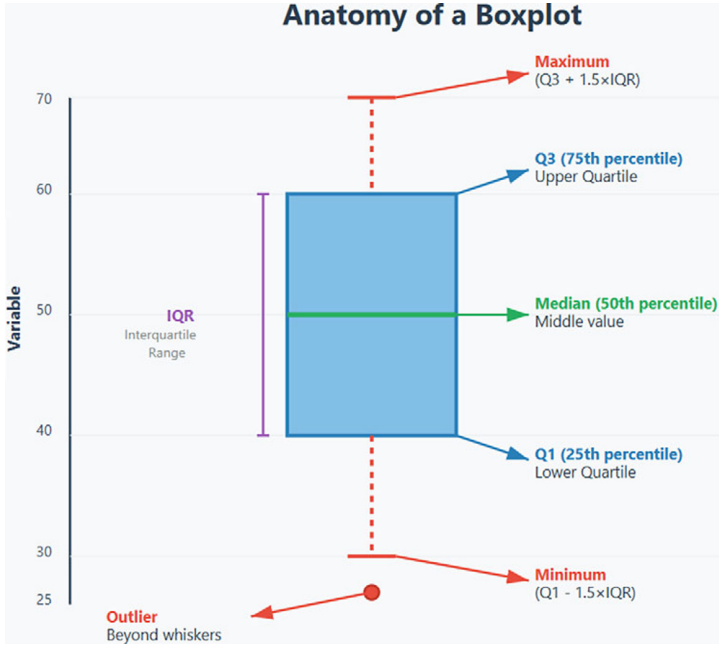


Figure 1. A representation of a Boxplot.

Figure 1 presents a labeled boxplot, which is a graphical method for visualizing the distribution, central tendency, and spread of a continuous variable. The box in the center represents the interquartile range (IQR), which spans from the first quartile (Q1, 25th percentile) to the third quartile (Q3, 75th percentile). This range captures the middle 50% of the data. The thick horizontal line inside the box marks the median (50th percentile), indicating the central value of the distribution. Whiskers extend from either end of the box to the smallest and largest values within 1.5 times the IQR below Q1 and above Q3, respectively. Any data points that fall outside this range are plotted individually and considered outliers. One can infer the skewness of the distribution by examining the symmetry of the box and the length of the whiskers. A perfectly symmetric box with equal-length whiskers suggests normality, whereas noticeable asymmetry indicates skew. The presence and number of outliers provide additional insight into the variability and potential anomalies in the data.

However, boxplots are less capable of dealing with bimodal distributions. Choonpradub and McNeil (2005) discussed in detail how a boxplot could look the same as that of a normal distribution when the raw data actually follows a perfectly symmetric bimodal distribution (a set of data that has two main peaks or high points), meaning that it could mask the real distribution shown in a histogram.

Goodness-of-fit (GoF)-based graphical approaches. Unlike raw-data approaches that visualize data distribution directly, GoF-based measures, like P-P plots and Q-Q plots, compare sample distributions against their theoretical normal counterparts.

Probability-Probability (P-P) plots. P-P plots compare the theoretical cumulative probability function (CDF) of a variable with the ECDF (Rani Das, 2016). These plots display theoretical cumulative probabilities on the x-axis against empirical cumulative probabilities on the y-axis, with a diagonal line indicating perfect alignment with the

https://doi.org/10.1017/S0272263126101600 Published online by Cambridge University Press

normal distribution. The cumulative nature of P-P plots makes them particularly sensitive to central distribution deviations (Delignette-Muller & Dutang, 2015; Wilk & Gnanadesikan, 1968). Minor deviations in the central region produce notable divergences from the diagonal reference line, while tail deviations may be less apparent due to the cumulative calculation diluting outlier effects (see Figure 2A).

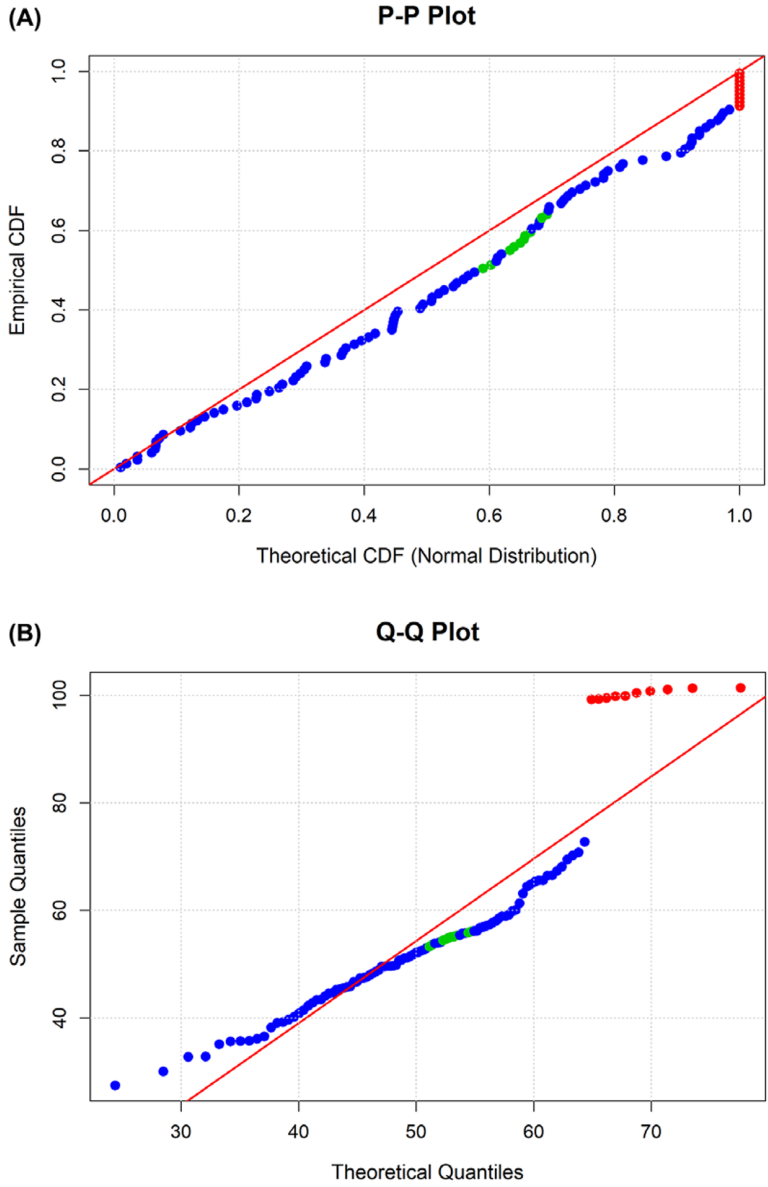


Figure 2. P-P plot (A) and Q-Q plot (B) with 20 unexpected values.

https://doi.org/10.1017/S0272263126101600 Published online by Cambridge University Press

Quantile-Quantile (Q-Q) plots. Q-Q plots also employ a diagonal line to indicate perfect normality but utilize quantile comparisons rather than cumulative probabilities. A quantile is a point that divides a dataset (or a theoretical distribution) into equal-sized portions. Theoretical quantiles (calculated using an inverse normal function) are plotted on the x-axis against empirical quantiles on the y-axis (Wilk & Gnanadesikan, 1968). This approach excels at revealing tail behaviors in data distributions but may be less informative about central tendencies, which often appear as a straight line (see Figure 2B), particularly in large samples (Ajao et al., 2018; Rodu & Kafadar, 2022).

Figure 2 demonstrates the complementary strengths of these GoF-based graphical approaches using a dataset with both tail anomalous values (red dots, mean = 100, SD = 2) and upper middle anomalous values (green dots, mean = 55, SD = 2). The P-P plot effectively highlights central distribution anomalies (green dots), while the Q-Q plot clearly reveals tail deviations (red dots) through marked departures from the diagonal line. This comparison illustrates how P-P plots excel at detecting central distribution deviations while Q-Q plots are more effective in identifying tail anomalies (Gan & Koehler, 1990).

Analytical measures

While graphical methods offer intuitive visualization and are widely available in commonly used statistical software in L2 research including SPSS, Jamovi, JASP, and R, as demonstrated in Table 2, their subjective interpretation may not be sufficient for normality assessment (Henderson, 2006). Therefore, analytical measures are indispensable and should be interpreted alongside graphical tests to validate statistical analysis choices (Keselman et al., 2013; Öztuna et al., 2006). As demonstrated in Figure 3, analytical tests can be categorized into moment-based (e.g., Jarque-Bera test) and GoF-based tests, with the latter further divided into empirical distribution function-based measures (e.g., Kolmogorov-Smirnov test) and regression/correlation-based measures (e.g., Shapiro-Wilk test) (Hernandez, 2021; Marange & Qin, 2021). These analytical methods will be discussed next.

In reviewing analytical normality test formulas, it is important to note that variations exist between sample-based and population-based calculations (Doane & Seward, 2011). This study focuses on sample-based formulas for two reasons. First, sample-based formulas explicitly incorporate the sample size of the study, which is useful in understanding why normality tests are sensitive to sample sizes. Second, there is a practical consideration that most, if not all, statistical software packages use sample-based rather than population-based normality tests. Understanding sample-based formulas therefore helps researchers better interpret the results their software produces.

Moment-based methods of assessing normality. Moment-based methods of assessing normality evaluate the shape of a distribution using its statistical moments—particularly the third and fourth central moments—to assess deviations from normality (Thode, 2002). As previously discussed, skewness, derived from the third moment, measures asymmetry, indicating whether values lean to the left or right of the mean. Kurtosis, based on the fourth moment, quantifies the heaviness of the distribution's tails and the sharpness of its peak. The Jarque-Bera (JB) test integrates these two metrics into a single goodness-of-fit measure by computing a test statistic using sample skewness and kurtosis to assess normality (Jarque & Bera, 1987; Thadewald & Büning, 2007). These methods are discussed next.

Skewness. Skewness is a descriptive statistic measure that quantifies the degree of asymmetry in a probability distribution around its mean (Doane & Seward, 2011; Ho &

https://doi.org/10.1017/S0272263126101600 Published online by Cambridge University Press

https://doi.org/10.1017/S0272263126101600 Published online by Cambridge University Press

Table 2. An overview of graphical and analytical methods for assessing normality and their availability in common statistical packages (as of November 2025)

Type	Analysis	Best uses	Jamovi	SPSS*	JASP	R
Graphical						
Raw data plotting	Histogram	Visualizes the distribution of data, identifies modes, skewness, and outliers. Preferred for large samples (n ≥ 200)	✓	✓	✓	✓
	Stem-and-leaf plots/ tables	Similar to histograms but provides individual values, ideal for small to medium datasets.		✓	✓	✓
	Boxplots	Effective for comparing distributions across multiple univariate variables, highlighting medians, quartiles, and outliers. Less effective for very small samples.	✓	✓	✓	✓
Goodness-of-fit (GoF) plotting	P-P plot	Assesses how closely data conforms to a specified theoretical distribution, especially effective in the distribution's body but less sensitive to outliers.	**	✓	✓	✓
	quantile-quantile (Q-Q) plot	Compares quantiles against a normal distribution. Sensitive to deviations in the tails.	✓	✓	✓	✓
Analytical						
Moment-based	Skewness	Detects asymmetry in a distribution. Use z-score for small to moderate samples, absolute values for large samples.	✓	✓	✓	✓
	Kurtosis	Detects peakedness and heavy tails in a distribution. Use z-score for small to moderate samples, absolute values for large samples.	✓	✓	✓	✓
	Jarque-Bera (JB) test	Measures skewness and kurtosis concurrently. Sensitive to sample size.		✓		✓
GoF-based measure: Empirical distribution function (EDF)-based measure	Kolmogorov-Smirnov (KS)	Versatile, no predetermined parameters. Works on all continuous datasets to identify the point of largest discrepancy between observed and theoretical data. Less powerful, more reliable as sample size grows but can be oversensitive in very large datasets.	✓	✓	✓	✓
	Lilliefors test	A modification of the KS test; uses the Monte-Carlo method to calculate theoretical normality, thus being more powerful than KS test.		✓		✓
	Cramer-von Mises (CVM) test	Versatile, no predetermined parameters. Focuses on the discrepancy across the entire dataset rather than identifying a single point.		✓	✓	✓
	Anderson-Darling (AD) test	A modification of CVM test, better at handling extreme values and detecting outliers in the tails.	✓	✓	✓	✓
GoF-based measure: Regression/ correlation-based measure	Shapiro-Wilk (SW) test	Compares ordered sample values with expected normal order statistics relying on predetermined coefficients. Specifically designed for normality tests; robust and better suited for small samples (3–50 coefficients available).	✓	✓	✓	✓
	Shapiro-Francia (SF) test	A modification of SW test without the need for predetermined coefficients. Less constrained by sample size but still performs better in small samples.		✓		✓

*The SPSS functions described are based on version 30+. Some features may not be available in earlier versions.

**As of November 2025, Jamovi does not natively generate P-P plots. They can be created via the *Rj editor* using R packages such as *ggplot2, car*, or *ggpubr.* In addition, some community modules in the Jamovi library (for example, *GAMLj* or *Rj*) allow users to run R code that includes P-P plot generation.

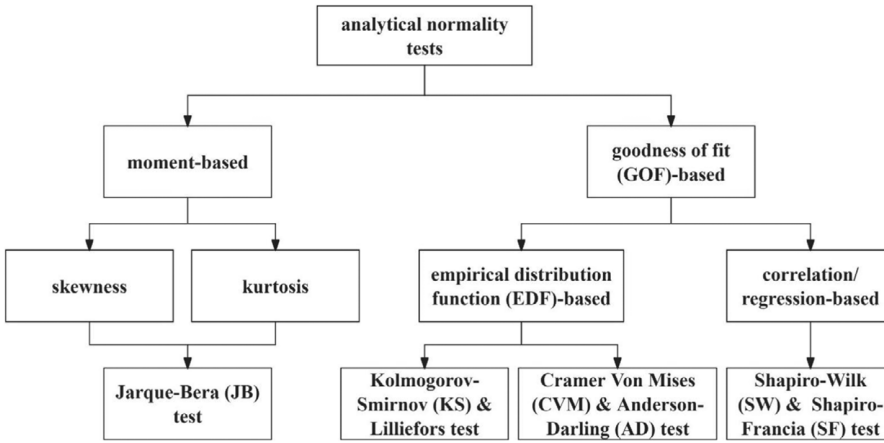


Figure 3. Classification of analytical normality tests by methodological approach.

Yu, 2015). It is calculated using the third moment of the mean and standardized by the cube of the SD:

$$Skewness = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{SD} \right)^3,$$

where n is the sample size, X_i is each data point, $\bar{X}$ the mean of the data and SD the standard deviation (DeCarlo, 1997). In a perfectly normal distribution, skewness equals zero, while the mean equals the sample's median and mode (Hatem et al., 2022). The cubic power in the formula intensifies the influence of values far from the mean, making it sensitive to outliers at the tails and amplifying the direction of the tails (Doane & Seward, 2011). A positively skewed distribution features a long tail on the right with fewer large values, whereas a negatively skewed distribution shows a concentration of data on the right with a tail extending to the left (Hair et al., 2019).

The interpretation of skewness values is influenced by sample size (Kim, 2013). To standardize skewness across different sample sizes and make it comparable, Z-scores, which account for the standard error of skewness, are recommended when dealing with small samples. The Z-score is computed by dividing the skewness (or kurtosis) values by their standard errors (Tabachnick & Fidell, 2013). For small samples ($n < 50$), the significant deviations are typically assessed against a critical Z-value of ±1.96 at an alpha level of 0.05. For medium-sized samples ($50 < n < 300$), significant deviations are usually assessed against a critical Z-value of ±3.29. For large samples ($n > 300$), an absolute skewness value greater than 2 might suggest nonnormality (Kim, 2013). On the other hand, according to Mayers (2013), for sample sizes between 50 and 175, a more stringent critical value of ±2.58, which corresponds to a 1% significance level ($p < 0.01$), is appropriate for z-scores of skewness and kurtosis, rather than the conventional ±1.96 (corresponding to $p = .05$). Proposed absolute skewness thresholds for normality assessment vary considerably in the literature, ranging from conservative values of 0.5 (Hatem et al., 2022) to more lenient thresholds of 1.5 (George & Mallery, 2001) and 3 (Kline, 2016).

https://doi.org/10.1017/S0272263126101600 Published online by Cambridge University Press

Kurtosis. Kurtosis is another moment-based measure that is calculated as the standardized fourth moment about the mean:

$$Kurtosis = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{SD} \right)^4,$$

where n is the sample size, x_i denotes each observation value, $\bar{x}$ is the sample mean, and SD is the sample mean, and SD is the sample standard deviation (Cain et al., 2017; DeCarlo, 1997). While commonly misinterpreted as merely reflecting distribution peak height, with positive kurtosis indicating a higher peak and negative kurtosis a lower peak, kurtosis characterizes both tail weight and peak characteristics of the data distribution (Balanda & MacGillivray, 1988). As discussed earlier, a normal distribution has an excess kurtosis of 0 (excess Kurtosis = Kurtosis proper – 3), representing balanced tail weights (Hair et al., 2019). Distributions with positive excess kurtosis display both increased peak height and heavier tails, indicating a higher probability of extreme values compared to a normal distribution. Conversely, negative excess kurtosis manifests as both a flatter peak and lighter tails, reflecting fewer extreme values.

Like skewness, the interpretation of kurtosis depends on sample size. While Kline (2016) proposed a general threshold of ±10, other researchers advocate for sample size-dependent criteria. Large samples usually provide more reliable excess kurtosis estimates. For small samples, where outliers can disproportionately affect the measure, Kim (2013) recommended using Z-score: ±1.96 for samples smaller than 50, and ±3.29 for sample sizes ranging from 50 to 300 (Kim, 2013). Mayers (2013) further refined these guidelines to suggest a Z-score range of ±2.58 for sample sizes between 50 and 175.

Jarque-Bera (JB) test. Although skewness and kurtosis are critical measures for distributional properties, neither measure alone comprehensively captures all features of normality (Kim, 2021). To synthesize both skewness and kurtosis, Jarque and Bera (1987) proposed the JB test as follows:

$$JB = \frac{n}{6} \left(Skewness^2 + \frac{1}{4} (Kurtosis - 3)^2 \right),$$

where *n* is the sample size. Under the null hypothesis of normality, the JB value, which combines both skewness and excess kurtosis, should theoretically converge to zero. This statistic follows a chi-squared distribution with two degrees of freedom, where a larger JB value indicates a stronger deviation from normality. The null hypothesis is rejected when the JB statistic exceeds the critical value derived from the chi-squared distribution (Kim, 2021; Thadewald & Büning, 2007). For example, at a significance level of 0.05, the critical value from the chi-squared distribution with two degrees of freedom is approximately 5.99. If the computed JB statistic exceeds this threshold, the data is considered nonnormally distributed. Conversely, a JB statistic below this critical value suggests insufficient evidence to reject the assumption of normality. As an integration of skewness and kurtosis, this test is also sensitive to sample size (Demir, 2022) and has multiple criteria to interpret the results (see Wijekularathna et al., 2019), but it is particularly useful in large samples, where its asymptotic properties yield reliable results.

Goodness-of-fit (GoF)-based measures. GoF-based measures complement moment-based approaches in normality testing by comparing observed data distributions with

https://doi.org/10.1017/S0272263126101600 Published online by Cambridge University Press

theoretical normal distributions through various computational methods, such as empirical distribution function (EDF)-based measures and correlations/regressions (Marange & Qin, 2021). GoF methods for assessing normality include the Kolmogorov-Smirnov test, Lilliefors test, Cramér-von Mises test, Anderson-Darling test, Shapiro-Wilk test, and Shapiro-Francia test, all of which are discussed in the following sections.

Kolmogorov-Smirnov (KS) test. The KS test quantifies the maximum discrepancy between the EDF of the data and the CDF of a normal distribution, expressed as:

$$\mathrm{KS} = \sup_x |F_n(x) - F(x)|,$$

where sup denotes the maximum difference over all possible values, $F_n(x)$ represents the empirical distribution function of the data and $F(x)$ is the theoretical CDF (Armitage & Colton, 1998).

This test does not assume the data must follow any specific parametric distribution, which allows for its application across any continuous distribution (Kim, 2013). However, this flexibility comes at the cost of reduced efficiency and power compared to other normality tests (Arnastauskaitė et al., 2021; Emmanuel et al., 2020; Steinskog et al., 2007). Although it demonstrates improved performance with larger datasets (Keselman et al., 2013), the test becomes overly sensitive as sample size increases, partly due to its critical value formula for n > 30: $0.886/\sqrt{n}$ at α = .05 (Lilliefors, 1967). This inverse relationship between critical value and sample size can lead to the rejection of the null hypothesis for minor deviations in large samples (Okeniyi et al., 2020). Moreover, debate exists regarding its sensitivity patterns: Lanzante (2021) argues that its cumulative nature intensifies its sensitivity to central distribution discrepancies, while others suggest it is more sensitive to tail deviations where maximum differences typically occur (Baghban et al., 2013; Darling, 1957; Lilliefors, 1967). These limitations have led some researchers to advise against using the KS test for normality assessment (D'Agostino et al., 1990; Schoder et al., 2006).

Lilliefors test. Given these limitations of the KS test, the Lilliefors test, a modification of the KS test, was proposed (Lilliefors, 1967). The original KS test assumes known population parameters (mean and standard deviation), but in practice, researchers must estimate these parameters from sample data. The Lilliefors test addresses this limitation by using estimated sample parameters and applying modified critical values that account for the additional uncertainty introduced by parameter estimation. These critical values were derived through Monte Carlo simulation methods to provide more accurate probability assessments when population parameters are unknown. This modification results in a more conservative and reliable normality test, particularly effective for small to moderate sample sizes (Abdi & Molin, 2007; Razali & Wah, 2011).

Cramer Von Mises (CVM) test. Like the KS and Lilliefors tests, the CVM test also employs the EDF to compare sample and theoretical distributions. It is computed as:

$$CVM = n\int_{-\infty}^{\infty} [F_n(x) - F(x)]^2 \Psi(x) dF(x),$$

where *n* represents the sample size, $F_n(x)$ is the cumulative distribution function (ECDF) of the sample, and $F(x)$ the theoretical CDF of the normal distribution, and $\Psi(x)$ serves as a nondecreasing weight function (Ahad et al., 2011; Cramér, 1928; Darling, 1957;

https://doi.org/10.1017/S0272263126101600 Published online by Cambridge University Press

Durbin & Knott, 1972). This method is nonparametric and is increasingly sensitive to larger sample sizes, which could potentially flag trivial deviations from normality that lack practical significance (Bayoud, 2021; Yap & Sim, 2011). However, instead of identifying maximum point-wise deviations like the KS test, the CVM test quantifies the sum of squared differences between the empirical and theoretical distributions across the entire range, with a uniform weighting ($\Psi(x) = 1$) to deviations throughout the distribution (Mala et al., 2021).

Anderson-Darling (AD) test. As a modification of the CVM test, the AD test employs a similar mechanism, which computes the statistic as:

$$AD = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)[\,ln\,F(x_i) + ln(1 - F(x_{n+1-i})],$$

where n is the sample size, x_i is the ordered sample values, and $F(x_i)$ as the CDF of the normal distribution (Anderson & Darling, 1954; Pettitt, 1977). The key distinction lies in its application of logarithmic transformation within the weighting function. As it approaches 0 or 1 at distribution tails, it is more sensitive to deviations in tails (Farrel & Rogers-Stewart, 2006; Razali & Wah, 2011).

Shapiro-Wilk (SW) test. Within the category of GOF test, in addition to the EDF-based tests discussed, there are regression/correlation-based measures represented by the SW test. The SW test calculates the square of the Pearson correlation coefficient between the ordered sample values and the expected normal order statistics, expressed as:

$$SW = \frac{\left(\sum_{i=1}^{n} a_i X_{(i)}\right)^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2},$$

where $X_{(i)}$ are the ordered sample values, $\bar{X}$ the sample mean, and a_i the coefficients that are derived from the expected values and the covariance matrix of the order statistics of a normal distribution (Shapiro & Wilk, 1965).

The distinctive feature of the SW test lies in its usage of predetermined coefficients, a_i, which are fixed for each sample size, and produce a test statistic ranging from 0 to 1. Values approaching 1 suggest a probable normal distribution origin, while values near 0 indicate normality departures. Hypothesis testing decisions rely on comparing the test statistic against established critical values (González-Estrada et al., 2022). While the SW test demonstrates robust performance in normality assessment, it is recommended to be applied in small samples (3–50) (Khan & Ahmad, 2017), where the coefficients were calculated by Shapiro and Wilk (1965, pp. 603–605). Royston (1982, 1992) provided an approximation to transform the Shapiro-Wilk W statistic to a normal variate z, which extended the test computation to larger samples (n ≤ 2000), with errors less than 0.001 for samples up to n = 1000 (Royston, 1992, p. 118). However, uncertainty remains regarding which computational methods statistical software packages actually implement. While Royston's extensions theoretically enable reliable SW testing for large samples, some software may rely on alternative approximation methods that can become unreliable when sample sizes exceed 50 (Royston, 1992, p. 118). This could explain why researchers cautioned against the use of the SW test beyond n = 50, as its validity for larger samples remains uncertain (Park, 2008, p. 8).

https://doi.org/10.1017/S0272263126101600 Published online by Cambridge University Press

Shapiro-Francia (SF) test. The SF test offers a modification of the SW test. It computes the test statistic as:

$$SF = \frac{\left(\sum_{i=1}^{n} m_i X_{(i)}\right)^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2},$$

where m_i is a vector of standard normal ordered statistics (Shapiro & Francia, 1972). Unlike the SW test that requires the calculated a_i, the m_i in the SF test can be approximated using the quantile function of the standard normal distribution, thus being more flexible in application (Mbah & Paothong, 2015). This test is not available in earlier versions of SPSS, which is one of the commonly used packages in L2 research, but it would be useful for evaluating samples larger than 50—beyond the range for which the Shapiro-Wilk test is recommended—up to approximately 2,000 observations (Shapiro & Francia, 1972). However, software like SAS, which is used in around 8% of L2 research as reported in Loewen et al. (2014) and mentioned as a common software tool (along with SPSS) in language research by Brown and Rodgers (2002), does not support this test (Park, 2008, p. 8).

Method

Paper inclusion

To better understand normality checking practices in L2 research, we followed Hu and Plonsky's (2021) research approach and examined empirical studies used the search function to identify keywords such as "normality," "normally distributed," "right/left/positively/negatively skewed," "skewness," "kurtosis," "distribution," and "normal" in full manuscripts. The selected journals include the *Language Learning* and *Second Language Research*, used in Hu and Plonsky (2021), which are known to publish largely quantitative research in L2 research (Gass, 2009). To broaden our coverage of L2-focused research, we expanded this journal pool by identifying additional high-quality journals. Using L2-related search terms ("second language," "foreign language," "TESOL"), we searched the Q1 journal list (2024) under the linguistics and language subcategory in the SCImago Journal Rank (SJR). From 362 Q1 journals, we identified eight additional journals containing these keywords: *Asian-Pacific Journal of Second and Foreign Language Education*, *Journal of Second Language Studies*, *Journal of Second Language Writing*, *Reading in a Foreign Language*, *Studies in Second Language Acquisition*, *Studies in Second Language Learning and Teaching*, *TESOL Journal*, *TESOL Quarterly*.

Figure 4 presents the PRISMA flow diagram for study selection. We conducted an initial search using ISSN numbers for all ten journals in Scopus, covering publications from 2020 to 2025 (June). We excluded editorials, notes, and reviews and focused exclusively on research articles. Given that all target journals publish in English, no language restrictions were applied. The specific search query was:

> ISSN (25423835 OR 23635169 OR 15390578 OR 00398322 OR 19493533 OR 10603743 OR 20835205 OR 02722631 OR 14770326 OR 14679922) AND PUBYEAR > 2019 AND PUBYEAR < 2026 AND (LIMIT-TO (DOCTYPE, "ar")).

This initial search yielded 2,002 papers. Despite limiting document types to "articles," we identified additional review papers and did manual exclusion. Using title-based

https://doi.org/10.1017/S0272263126101600 Published online by Cambridge University Press

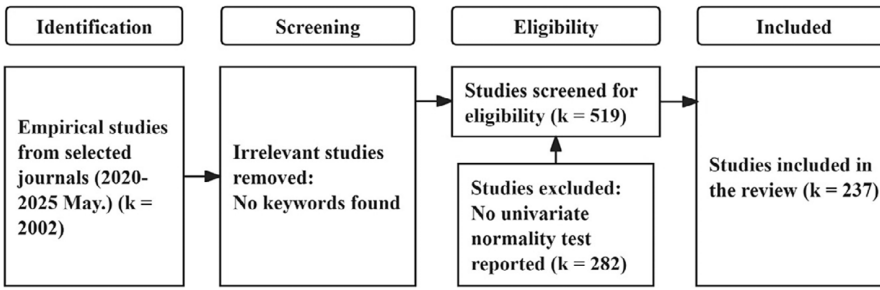


Figure 4. PRISMA flow diagram for the review.

keyword searches for terms like "review," "meta-analysis," and "bibliometric," we excluded 45 review papers, which yielded a refined dataset of 1,957 papers.

Given the large amount of the refined dataset, it is impractical to manually review nearly 2,000 papers in full texts, thus we employed an AI-powered screening. Since normality testing applies specifically to quantitative research, we used ChatGPT-4o to classify research methodologies based on titles and abstracts of selected papers. We used the following prompt for classification:

> We are conducting a study on normality testing, which is only relevant for quantitative research. Due to the large number of papers involved (thousands), manual review is not feasible for all entries. Therefore, I will upload batches of 20 studies, each containing only the title and abstract. Your task is to examine each and classify the research method as one of the following: Quantitative, Qualitative, Mixed Methods, and NA (Not Available/Not Clear—use this when the method cannot be reliably determined from the title and abstract alone). Please make your best judgment based on the content provided. Do not guess. If the method is not clear, simply assign "NA."

The AI classification yielded 574 qualitative papers, 1,023 quantitative papers, 230 mixed methods papers, and 130 papers with unclear methodology. To validate this classification approach, we manually re-assessed 60 randomly selected papers (approximately 10%) from those labeled as qualitative. This validation revealed only one misclassified quantitative study, indicating a 98.3% accuracy rate. Based on this high accuracy, we excluded the 573 correctly identified qualitative papers from further analysis, resulting in a final dataset of 1,384 papers.

Among our target journals, the *Journal of Second Language Studies* was not accessible through our institutional subscriptions. Consequently, we included only 20 of 68 papers from this journal, which we obtained through author correspondence or open-access availability. We successfully obtained full-text access to 1,336 empirical papers and conducted comprehensive keyword searches for terms related to distributional properties: "normality," "normally distributed," "right/left/positively/negatively skewed," "skewness," "kurtosis," "distribution," and "normal." This systematic search identified 570 empirical studies that explicitly addressed distributional properties, forming the basis for our subsequent analysis of normality checking practices in L2 research.

Among these 570 studies, 45 were further excluded as they addressed other types of distributions (e.g., multivariate, bimodal) rather than univariate normal distributions,

https://doi.org/10.1017/S0272263126101600 Published online by Cambridge University Press

and 6 mentioned normality specifically to emphasize that their analyses did not require normality assumptions (e.g., Bayesian mixed-effects models). Of the remaining 519 studies, 282 were additionally excluded because, although they reported whether their data violated normality assumptions, they did not explicitly state which specific test(s) they performed to reach these conclusions. It is worth noting that many studies directly applied logarithmic transformations to theoretically nonnormal data, such as response times, which are typically right-skewed (e.g., Ahn & Jiang, 2023; Berghoff, 2022) and word frequency data following Zipfian distributions (e.g., Saito, 2020), or simply assumed data normality when sample sizes reached 30 (e.g., Akbarian & Elyasi, 2023). Almost no studies reported whether the transformed data met normality assumptions, with only a few exceptions that performed graphical (histogram) or statistical (skewness and kurtosis) checks on the transformed data (e.g., Ahn, 2021; Nakata et al., 2023).

Ultimately, 237 studies were included in the final review: 41 from *Language Learning* and 10 from *Second Language Research*, 4 from *Journal of Second Language Studies*, 33 from *Asian-Pacific Journal of Second and Foreign Language Education*, 17 from *Reading in a Foreign Language*, 15 from *TESOL Quarterly*, 11 from *TESOL Journal*, 27 from *Studies in Second Language Learning and Teaching*, 65 from *Studies in Second Language Acquisition*, and 14 from *Journal of Second Language Writing*.

Coding scheme

Since normality checking approaches and their interpretation can be sensitive to sample size, the coding scheme for this study includes four dimensions: (1) sample size, (2) normality approach type, (3) specific normality checking approach, and (4) criteria used to understand what is commonly practiced.

Following Hu and Plonsky (2021), we searched for keywords such as "normality," "normal distribution," "normally distributed," and "distribution is normal," then closely examined the methods, data analysis, results, notes, and appendices to identify the specific normality checks and criteria used. One of the authors conducted the first-round annotation of the 237 studies and used ChatGPT-4o as a second coder in October 2025 to review 48 studies (20%) and retrieve information using the prompt shown in Figure 5. The entire manuscript in PDF format was uploaded to the chatbot one article at a time, with one article per prompt and data controls for model improvement disabled to protect the intellectual property of the authors. According to OpenAI's privacy policy, content uploaded with these settings is not used for model training. No preprocessing of the manuscripts, such as conversion to plain text, was performed, as this could distort figures and tables and potentially result in the loss of information related to normality checking. Given that the AI model used, ChatGPT-4o, is designed to process multimodal information (OpenAI, 2024) and has been successfully integrated into reference management software plugins that directly process manuscripts in PDF format (e.g., GPT Pro in Zotero), we chose to upload the complete manuscripts in their original PDF format to preserve all information.

No interactive follow-up prompts were used to maintain independence of the AI review. When token limits were reached, the authors waited for the limit to reset and re-uploaded both prompt and manuscript in a new conversation window. No other output issues were encountered. When discrepancies arose between human coding and AI annotations, the authors referred back to the original manuscript to determine the source of error. For studies reporting both univariate and multivariate

https://doi.org/10.1017/S0272263126101600 Published online by Cambridge University Press

You are an expert systematic reviewer. Your task is to extract information about normality testing practices from uploaded research manuscripts.
For each paper, identify and record the following four categories:

- Sample Size: the number of participants or observations included in the normality checking
- Normality Approach Type
  - Graphical (e.g., histogram, Q-Q plot, boxplot, P-P plot, etc.)
  - Analytical (Statistical) (e.g., Shapiro–Wilk, Kolmogorov–Smirnov, Jarque–Bera, etc.)
  - Both (if the paper uses both graphical and analytical checks).
- Specific Normality Checking Approach
  - List the exact method(s) used.
    Examples: "Shapiro–Wilk test," "Kolmogorov–Smirnov test," "visual inspection of Q–Q plots," etc.
  - If the paper reports multiple approaches, include all of them.
- Criteria Used to Judge Normality
  - Describe the criterion or threshold used to determine normality, e.g., skewness $< |1|$, kurtosis $< |3|$).
  - If the study did not specify the criterion but mentioned cites another source (e.g., "following Kline, 2011"), record that citation.

Notes:

- If multiple studies are described in one paper, create a separate row for each.
- If unclear or missing, write "NA"

Figure 5. Prompt for ChatGPT-4o to code manuscripts.

normality tests, only univariate annotations were included in the reliability assessment, though the AI review annotated both. Example AI reviewer outputs are provided in Supplementary Material 1.

The comparison revealed 11 mismatches out of 192 (48 * 4 categories) annotation points, yielding 94.27% absolute agreement between the human coder and AI model. Analysis of these 11 mismatches showed they fell into two categories: sample size discrepancies and normality testing discrepancies. Seven mismatches involved sample size identification (85.42% agreement). In one instance, the human coder used the initial participant count without noting that several participants had been excluded from the final analysis. In the remaining six cases, the AI model incorrectly identified the analytical sample. For example, some studies examined rater behavior using rating materials from *n* speakers. The AI model appeared to identify the *n* speakers as the sample size, when the raters themselves were the actual study participants whose data were analyzed. After resolving discrepancies by consulting original manuscripts, the human coder achieved 97.92% accuracy (47/48) for sample size identification, while the AI model achieved 87.5% accuracy (42/48). The remaining four mismatches concerned normality checks. The human coder overlooked two normality tests that had been reported. In two other studies, the AI model annotated normality checks as "NA" because the normality testing results appeared in supplementary materials that were not provided to the model. The human coder subsequently accessed these online

https://doi.org/10.1017/S0272263126101600 Published online by Cambridge University Press

supplementary materials and confirmed that the original human annotations had correctly captured the normality testing information. These two cases affected both the "normality checking approach type" category and the "specific normality checking approach" category. However, to avoid duplicate counting of the same discrepancy, they were recorded as two mismatches under the "normality checking approach type" category only, rather than four mismatches across both categories.

Overall, in the 48 studies (20% of the total papers), we identified 72 instances of normality assessment. The human annotator correctly coded 70 of these instances (97.22% accuracy) in the first round, having overlooked two reported normality tests. The AI model also correctly identified 70 instances from the materials provided to it. The two instances it missed appeared only in supplementary materials that were not uploaded, which was a data access limitation rather than a coding error. When considering only the information available to the AI model, its accuracy was 100%. Inter-coder reliability for the categorical variable "normality checking approach type" yielded Cohen's $\kappa = 0.895$, which indicates almost perfect agreement (Landis & Koch, 1977). This high reliability is expected, given that the annotation in this study was quite objective and straightforward.

Results

A total of 382 normality checking instances were performed across the 237 studies, as 113 studies reported using more than one approach for normality checking. Details about the test methods used are summarized in Table 3. Overall, analytical approaches were substantially more common (75.92%) than graphical ones (24.08%). Within the analytical category, moment-based methods—especially skewness and kurtosis—were most frequently applied, while the Shapiro-Wilk and Kolmogorov-Smirnov tests dominated among GoF-based tests. In contrast, Q-Q plots and histograms were the preferred graphical techniques. Note that the "Combined" category refers to studies that employed both analytical and graphical methods for checking normality within the

Table 3. Normality checking reported for each study

Normality checking type	Normality checking approach	K	Percentage Category (overall %)
Graphical (k = 66)		*92*	*100 (24.08)*
	Histogram	30	32.61 (7.85)
	P-P plot	5	5.43 (1.31)
	Q-Q plot	42	45.65 (10.99)
	Scatter plot	4	4.35 (1.05)
	Boxplot	3	3.26 (0.79)
	Violin plot	1	1.09 (0.26)
	Residual plot	7	7.61 (1.83)
Analytical (k = 196)		*290*	*100 (75.92)*
Moment-based		162	55.86 (42.41)
	Skewness	86	29.66 (22.51)
	Kurtosis	76	26.21 (19.90)
GoF-based		128	44.14 (33.51)
	Shapiro-Wilk test	67	23.10 (17.54)
	Kolmogorov-Smirnov test	61	21.03 (15.97)
Combined analytical and graphical methods *(k = 26)*		75	

Notes: The counts (K) individual approach exceeds the number of studies (k) as some studies applied multiple normality checking tests.

https://doi.org/10.1017/S0272263126101600 Published online by Cambridge University Press

same investigation. Since the specific tests used in these studies are already included in the respective "Analytical" and "Graphical" summaries, the "Combined" cases were excluded from the percentage calculations to prevent double-counting. The results presented in Table 3 are further elaborated below.

The "Combined" category indicates that a single study performed both analytical and graphical normality checking. Because the specific tests used in these studies have already been counted in the "Analytical" and "Graphical" summary sections, respectively, the "Combined" cases are not included in the percentage calculations to avoid double-counting.

Graphical tests

Among the 66 studies that performed graphical normality checking, the Q-Q plot is the most popular method used in L2 research, accounting for 45.65% of all graphical normality tests, followed by histograms (32.61%), residual plots (7.61%), P-P plots (5.43%), scatterplots (4.35%), and boxplots (3.26%). As explained in Graphical Measures Section, Q-Q plots excel at detecting tail deviations, whereas P-P plots are more sensitive to central deviations but may fail to clearly show extreme outliers in the tails (see Figure 2). Regarding histograms, they are considered appropriate or reliable for assessing normality only when the dataset is relatively large—specifically, when the sample size is greater than 200 (Neter et al., 2005; Oppong & Agbedra, 2016). Among the studies reviewed here, only 12 out of 28 studies using histograms met or approached this requirement (sample sizes ranging from 150 to 24231), while many of the remaining studies had considerably smaller samples (24–100). Boxplots are primarily used to identify outliers, which is accurately applied in our reviewed case.

The use of imprecise terminology such as "scatterplot" for normality assessment in some of the reviewed studies introduces potential ambiguity, as it does not distinguish between bivariate scatterplots, which do not evaluate normality, and probability plots such as P-P or Q-Q plots, which are specifically designed for this purpose. Similarly, "residual plots" represent an umbrella term encompassing various diagnostic visualizations. Plots of residuals versus fitted values primarily evaluate linearity and homoscedasticity assumptions (Altman & Krzywinski, 2016), whereas Q-Q plots of residuals are specifically employed to assess normality of the error distribution. It is possible that when researchers report using "residual plots" for normality assessment, they are actually referring specifically to Q-Q plots of residuals rather than alternative residual diagnostics. Since none of the reviewed studies displayed the actual plots in their manuscripts or appendices, which may be due to word or space limits, and only briefly described their approach, we cannot definitively determine how scatterplots and residual plots were employed for normality checking. Stem-and-leaf plots were not used in any of the L2 studies included in this review. One study used a violin plot to show distribution (Hanzawa & Suzuki, 2023), which was not commonly used as a normality test. This visualization is similar in nature to histograms for showing raw data distribution. However, to follow the expression of the original study, we retained it as a violin plot.

Analytical tests

Regarding analytical normality checking, both moment-based and GoF-based approaches are commonly used in the L2 studies reviewed. Among these, skewness

https://doi.org/10.1017/S0272263126101600 Published online by Cambridge University Press

is the most frequently employed, accounting for 29.66% of all statistical checks, followed by kurtosis (26.21%), the SW test (23.10%), and the KS test (21.03%). Kurtosis was often calculated alongside skewness but not vice versa; no study applied kurtosis alone without skewness. It appears that these studies likely computed and reported excess kurtosis rather than raw kurtosis, because they used conventional statistical software such as SPSS. However, they did not explicitly state this in their reporting.

As explained in Analytical Measures Section, analytical normality checking can be sensitive to sample size. Some tests may be overly sensitive when dealing with large sample sizes, such as the SW test, while others require different criteria to accommodate varied sample sizes, such as skewness and kurtosis measures.

GOF-based measures

This sample size sensitivity represents one of the common issues found in current L2 practice. In the L2 studies that employed the SW test, their sample sizes ranged between 8 and 5,600, with 17 out of 67 SW tests having sample sizes of 50 or fewer, while 28 exceeded 100, and 22 cases had sample sizes between 50 and 100, as demonstrated in Figure 6. The SW test is highly recommended for small sample sizes because it uses predetermined coefficients that are fixed for each sample size (3-50) (Royston, 1982, 1992), with established critical values (González-Estrada et al., 2022; Khan & Ahmad, 2017). However, once the sample size exceeds 100, the original SW test can easily flag nonnormality for minor deviations (Fu & Li, 2022). For larger samples ($50 < n \leq 2000$), the test may also be reliable if Royston's approximation method is properly implemented in the software employed. In other words, once the sample size exceeds 50, the reliability of the test depends critically on the specific computational implementation used by each software package. Under these circumstances, researchers should verify the computational method employed by their software or consider alternative approaches. If such verification is not feasible, the SF test, which is a modification of the SW test, is recommended as an alternative. The SF test does not require predetermined coefficients, making it potentially more computationally

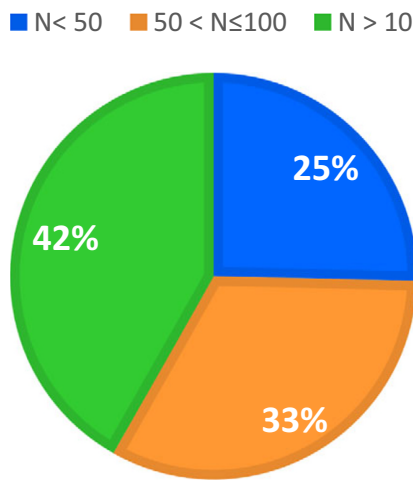


Figure 6. Distribution of sample size in L2 studies using the Shapiro-Wilk (SW) test.

https://doi.org/10.1017/S0272263126101600 Published online by Cambridge University Press

efficient for larger samples. However, no studies with large sample sizes used the SF test for normality checking in the L2 research we reviewed, possibly because the SF test is less accessible than the SW test in statistical software commonly used by L2 researchers, such as earlier versions of SPSS, Jamovi, and JASP (Yap & Sim, 2011).

In our reviewed cases, only 25% of studies had sample sizes under 50. However, this does not mean that the remaining 75% of studies were performing normality tests incorrectly. Rather, as we mentioned, when the sample size exceeds 50, and especially exceeds 100, certain computational approximation methods should be used to ensure the reliability of the SW test.

The KS test, although long criticized for normality checking (D'Agostino et al., 1990; Schoder et al., 2006), remains commonly used in L2 research (21.03% of analytical checks and 15.97% of all checking approaches). Without requiring a normal distribution prerequisite, the KS test is more versatile but less reliable for normality checking. Therefore, when the sample size exceeds the range for which SW test coefficients are available, the KS test could serve as a quick alternative. It effectively filters out clear normality violations; as a lenient measure, data that fail the KS test are unlikely to pass more stringent normality checks, suggesting nonparametric alternatives (e.g., Bylund et al., 2024). However, relying solely on the KS test for normality decisions proves problematic, particularly when subsequent analyses lack robustness against nonnormality.

Moment-based measures

Moment-based approaches including skewness and kurtosis appeared frequently, either together or with skewness alone, as these indicators capture different aspects of normality. Unlike GoF-based measures that provide clear p-values (usually $p < 0.05$) for significance determination, skewness and kurtosis criteria are more complex due to sample size effects. As previously discussed, critical values vary in both magnitude (3.29 or 1.96) and type (absolute value or z-score). While no fixed standards exist, researchers typically recommend z-scores for small ($n < 50$, ±1.96) and medium samples ($50 < n < 300$, ±3.29 or ±2.58), with absolute values for large samples ($n > 300$, ±2). More conservative absolute thresholds ranging from 0.5 to 3 when subsequent models accommodate moderate nonnormality (Kline, 2016, for structural equation modeling).

Among our 85 reviewed cases, 39 studies used absolute values: 29 applied absolute values of ±2 for skewness with sample sizes ranging from 40 to 61,200; 5 studies used more stringent absolute values of ±1.5 or ±0.5 with samples ranging from 60 to 243, and 4 studies used more lenient absolute values of ±3 for skewness with sample sizes ranging from 190 to 863. Thirteen studies reported z-scores with samples from 36 to 906, of which 2 failed to report specific z-score criteria. Among the 29 studies using absolute skewness values of ±2, a total of 10 met the recommended sample size requirement ($n > 300$). The remaining 19 studies could have benefited from employing z-scores to address the unreliable variances in small samples that can bias skewness calculations. Regarding the 11 studies that used and reported z-scores, critical values of 1.96, 2.58, and 3.29 were adopted. Among these studies, most of them met the sample size requirement ($n < 300$) to apply z-scores of skewness for evaluating normality. However, there was a tendency to use a strict critical value of 1.96, which is recommended for samples under 50 (Kim, 2013). For some studies, for instance, Otwinowska et al. (2020), with a sample size of 83, which falls within the 50-300 range where Kim (2013) recommends a more lenient criterion of 3.29, the data were evaluated using the stricter

https://doi.org/10.1017/S0272263126101600 Published online by Cambridge University Press

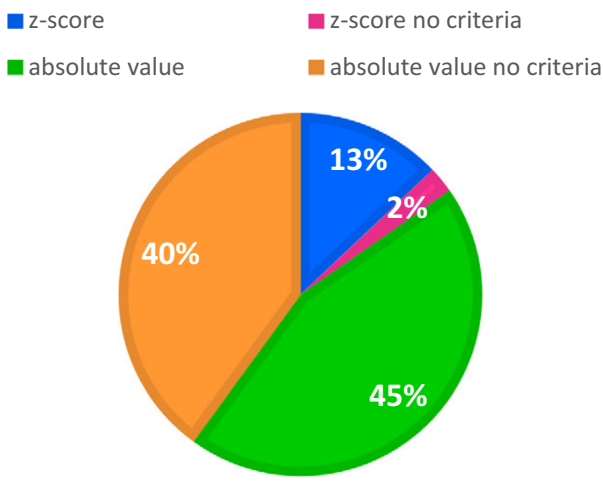


Figure 7. Normality reporting practices in using skewness.

threshold of 1.96. Notably, their data met the normality assumption even under the stricter 1.96 threshold, which supports the robustness of their results.

Figure 7 presents the proportion of studies using z-scores or absolute values and the proportion of cases that failed to report any specific criteria. We did not break down the results by specific criteria, as there are no common standards for such criteria, as we explained above, and authors should consider their subsequent analysis when selecting the criteria they use. It is worth noting that 42% of the cases in our reviewed studies did not report their criteria, which may limit the reproducibility and evaluation of outlier identification procedures. Studies that do not report criteria for the interpretation of skewness and kurtosis create evaluation challenges. Johannessen et al. (2025) described skewness values ranging from –1.5 to –0.55 as "highly skewed," while Wang and Halenko (2024) interpreted skewness from –1.229 to 1.802 (z-scores from –1.714 to 1.597) as normally distributed. While both interpretations may be valid within their respective frameworks, the absence of clearly stated thresholds can lead to ambiguity. In some published work, misunderstandings about distributional criteria can lead to reporting errors. For example, there are cases where datasets were described as normally distributed because skewness values exceeded ±2, when in fact values should fall within that range to support the assumption of normality. Kurtosis was often reported alongside skewness, with 41 (53.95%) studies applying absolute thresholds such as ±2, ±3, or ±7, or a z-score of 3.29 or 1.96, while 35 (46.05%) did not report any criteria. These inconsistencies reflect both minor and major issues in the interpretation of skewness and kurtosis—specifically, the use of ambiguous thresholds for identifying distributions that do not substantially deviate from normality, and the reliance on fixed absolute cutoffs, which may not be appropriate for small sample sizes.

Review summary

In summary, while normality is extensively researched among statisticians, our results suggest that L2 studies published in the selected L2-focused journals have given relatively insufficient attention to this critical assumption, with low checking rates

https://doi.org/10.1017/S0272263126101600 Published online by Cambridge University Press

documented, which is in line with previous L2 reviews (e.g., Hu & Plonsky, 2021). Among those who assessed data normality in their studies, approximately half failed to report which specific tests they employed (282 out of 519 studies). Many of these 282 studies assumed or claimed their data would achieve normality following certain transformations without conducting normality checks on the transformed data. Among the 237 studies that did report normality testing, 26 studies performed both analytical and graphical analysis, while 170 reported only analytical normality checking results. The remaining 41 relied on graphical measures to determine normality. While graphical methods provide valuable visual information about distributional properties, combining them with analytical tests may offer a more comprehensive assessment of normality assumptions (Hernandez, 2021).

The remaining studies that employed analytical measures showed limited diversity, using primarily two GoF-based tests (SW and KS tests) and two moment-based measures (skewness and kurtosis). However, these studies gave inadequate consideration to the sample size sensitivity of different tests and the appropriate selection of criteria values and types. Problematic practices include: (1) SW tests were frequently applied to large sample sizes where reliability diminishes beyond n = 50 if no proper approximation method is used; (2) KS tests were used inappropriately for small samples; (3) critical values for moment-based measures were often unreported or improperly selected for different sample sizes. For example, researchers frequently applied absolute values for small sample sizes when z-scores would be more appropriate to account for the less reliable standard deviations characteristic of smaller samples.

Recommendations for normality checking

Building on the findings from the two previous sections, while multiple normality checking approaches exist to address various violations of univariate distribution across different sample sizes, L2 researchers have adopted a relatively small subset. In some cases, these methods have been inappropriately applied in terms of approach selection and criteria setting. To address these methodological gaps, we propose a set of practical recommendations for L2 research and statistical training for researchers. A summary of the recommendations is provided in Table 4.

First, our review reveals that L2 researchers perform transformations on theoretically nonnormally distributed data with many not assessing normality of either the original or transformed data. While transformation is a useful and efficient strategy, not all transformations successfully convert nonnormality to normality, as demonstrated by Nakata et al. (2023). It would be important to re-assess normality after applying data transformation or correction to the original dataset.

Second, regarding normality checking practices, we recommend beginning with graphical exploration as a preliminary step. It is useful to start with histograms or stem-and-leaf plots to visualize raw data distribution patterns and understand data characteristics. As Tukey (1977) famously remarked, "There is no excuse for failing to plot and look." For instance, distributions exhibiting bimodality (i.e., two distinct peaks) violate the assumption of unimodality underlying most univariate normality tests, indicating that such tests are inappropriate in these cases (Thode, 2002). Boxplots can then identify potential outliers and help determine whether extreme values should be removed. Alternatively, researchers can explore Q-Q plots for highly skewed distributions with numerous extreme values, as these excel at detecting tail deviations, while P-P plots better reveal central deviations in distributions with pronounced peaks. Not

https://doi.org/10.1017/S0272263126101600 Published online by Cambridge University Press

https://doi.org/10.1017/S0272263126101600 Published online by Cambridge University Press

Table 4. Practical recommendations for normality assessment in L2 research by sample size, method, and rationale.

Recommendation Area	Applicable Sample Size	Relevant Methods/ Tests	Specific Action	Rationale	Reference
Initial assessment	All	Histogram, stem-and-leaf plot	Works better at larger sample, but too large might cause invisibility in stem-and-leaf plot	Visualize raw data and identify patterns	Tukey (1977)
Outlier detection	All, especially small samples	Boxplot, Q-Q plot	—	Detect extreme values and tail deviations	Gan and Koehler (1990)
Test selection	n <50	Shapiro–Wilk test	—	Robust performance in small n	Khan and Ahmad (2017)
		Skewness/kurtosis z-scores	Apply z-scores with ±1.96 for small samples	Accounts for sampling error in small datasets	Kim (2013)
	n <300	Lilliefors test	—	Lilliefors test is effective for small to medium samples	Abdi and Molin (2012)
	50 < n < 300	Skewness/kurtosis z-scores	Use z-score cutoffs of ±2.58 (n < 175) or ±3.29	Flexibility based on analytical context	Kim (2013); Mayers (2013)
	n >50	Shapiro–Wilk test	Check if the computational method of coefficient estimation is Royston's approximation	Test reliability depends on computational methods	Royston (1992)
		Shapiro–Wilk test	—	More robust options with increasing n if Royston's approximation cannot be verified and computational efficient	Shapiro and Francia (1972)
	n >300	Skewness/kurtosis (raw)	Use absolute skewness ±2 and kurtosis ±7	Practical for large samples	Kim (2013)
	Large samples	Q-Q, P-P plots, Histogram	Avoid over-relying on analytical tests in large n	Large samples inflate sensitivity	Bayoud (2021); Demir (2022)
Posttransformation validation	All	Refer to test selection	Re-assess normality after data transformation	Not all transformations restore normality	Nakata et al. (2023)

all visualizations need to be reported in the main body of the paper, but they are helpful for deciding which analytical approaches to adopt and thus can be reported in appendices or supplemental figures. For instance, when outliers are present and graphical assessment is insufficient for removal decisions, analytical tests such as CVM test and AD test that are sensitive to tail deviations may help determine whether these extreme values genuinely violate normality assumptions. Visually, if data appear clearly skewed in raw distribution plots, skewness should be calculated.

Beyond distribution characteristics, sample size should be considered in selecting a normality test selection. For small samples ($n < 50$), we recommend the SW test as the primary choice given its robust design for normality assessment (Khan & Ahmad, 2017). Skewness and kurtosis may supplement this analysis but should employ z-scores with critical values of ±1.96 to account for unreliable standard deviations in small samples (Kim, 2013). The Lilliefors test also performs well with small datasets. Medium sample sizes ($50 < n < 200–300$) offer more options, including the SF test, Lilliefors test, and skewness and kurtosis measures. Z-score criteria here are flexible, with options including ±2.58 and ±3.29 (Kim, 2013; Mayers, 2013), and researchers need to select more lenient or stringent standards based on their analytical context and sample size. Regarding large samples ($n > 300$), normality checking can use absolute value criteria for skewness (±2) and kurtosis (±7) (Kim, 2013). Additionally, researchers should consider the robustness of their subsequent statistical analyses. For analyses with robustness toward moderate nonnormality, more lenient normality checking methods, such as the KS test or preliminary graphical checking, may suffice. Otherwise, we recommend following the above suggestions.

Finally, it would be important to recognize that all analytical normality measures can become oversensitive with very large samples, flagging trivial deviations from normality as statistically significant problems. As previously discussed, the Central Limit Theorem (CLT) suggests that normality concerns diminish with sufficiently large samples, as sampling distributions of means approach normality regardless of the underlying population distribution. While this scenario remains uncommon in participant-based L2 research, graphical assessment can provide valuable complementary insights in such cases. If data appear visually distributed in a bell-shaped pattern without obvious deviations in P-P and Q-Q plots, conversion to nonparametric analyses may be unnecessary even when analytical normality tests indicate nonnormal distribution. Overly strict adherence to normality standards risks inappropriate application of nonparametric analyses, which potentially lowers statistical power and leads to misleading conclusions.

The recommendations that follow are offered as practical guidance for navigating univariate normality testing in applied research. While researchers with extensive statistical training may already have established workflows, these suggestions could be especially helpful for those who wish to strengthen their quantitative methods or refine their approach to examining data distributions before analysis. We emphasize that these are recommendations rather than prescriptive requirements, recognizing that different research contexts and analytical goals may call for adapted approaches. Additionally, researchers have access to various data transformation techniques that can improve distributional properties when needed (Pek et al., 2018). As we have noted throughout this review, many statistical procedures demonstrate reasonable robustness to moderate departures from normality, and normality assessment should be understood as part of thoughtful analytical decision-making rather than as a binary gate that mechanically determines whether parametric or nonparametric approaches are permissible (Knief & Forstmeier, 2021).

https://doi.org/10.1017/S0272263126101600 Published online by Cambridge University Press

Discussion

The study is not without limitations. First, we focused on a short time window and on 10 Q1 journals that contain L2 or related terms in their titles, while there are also other well-known journals like *Language Teaching Research* and *Language Testing*, which also include L2 quantitative research. Although including more leading journals would draw a more comprehensive picture, as we reported in the PRISMA diagram, even with these time and journal restrictions, the initial searching yielded over 2,000 papers. Unlike other systematic reviews where unrelated studies can be easily screened out by titles, this review required reading into the methodology of each quantitative study, which would lead to an infeasible workload when further expanding the scope. The limited scope of this review means that the findings may not capture the full range of practices in L2 research.

Second, our research provides a detailed examination of normality checking practices in L2 research, but we acknowledge that positioning these findings relative to adjacent fields remains incomplete. Low rates of normality checking have been documented across multiple disciplines, including psychology, education, and medicine (Ernst & Albers, 2017; Nielsen et al., 2019; Shatz, 2023), suggesting this is not a limitation unique to applied linguistics. However, existing studies in other fields have typically focused on whether normality is checked rather than on how it is checked. We were therefore unable to locate comparable research examining the specific methods, criteria, and reporting practices that researchers employ when conducting normality assessments. Consequently, we cannot determine whether the patterns we observed—such as the predominance of certain normality tests or the underreporting of specific criteria—reflect practices unique to L2 research or represent broader trends across other empirical education disciplines. In the absence of comparative evidence, we offer several tentative explanations for the patterns documented in our review.

One explanation could be the limited space available in manuscripts. Since normality checking is one of many assumptions and it is not feasible for researchers to devote too much space to assumption checking, they may tend to reserve space for the main data analysis. This could also partially explain why analytical normality checking is mentioned more frequently than graphical analysis, as compared to placing a plot, a one-line report of analytical normality checking results saves more space while providing more objective evidence. However, the underreporting of assumption checking, regardless of the reason, reduces methodological transparency, undermining the interpretability and replicability of findings (Plonsky, 2024). This lack of transparency creates barriers for researchers attempting to evaluate study quality or build upon previous work.

A second possible explanation concerns the availability of normality checking approaches in commonly used statistical analysis tools or packages, especially in their early versions as reported in Table 2. Software design may also play a role. For example, as of November 2025, Jamovi offers several graphical approaches for checking normality such as histograms, along with skewness and kurtosis, all available in the *Descriptives* tab under the *Exploration* module. However, the SW test is embedded within many general linear modeling modules, making it only a click away during inferential analyses. In contrast, accessing the *Descriptives* module to run additional checks requires extra steps that may discourage users from doing so. Consequently, it is unsurprising that those using Jamovi tend to rely more heavily on the SW test.

A third explanation could relate to familiarity with statistical methods. Researchers may gravitate toward normality checking approaches they have encountered in their

https://doi.org/10.1017/S0272263126101600 Published online by Cambridge University Press

own training or in frequently cited methodological sources, particularly if exposure to alternative methods has been limited. This pattern would be consistent with broader observations about methodological practices being shaped by disciplinary conventions and pedagogical emphases. However, this remains speculative, as although several studies have examined statistical knowledge among applied linguistics researchers and graduate students (Gonulal et al., 2017; Loewen et al., 2014; Zhang & Han, 2024), none of them provided detailed information about their understanding of specific normality checking approaches.

Therefore, we recommend that future researchers focus on different subfields of L2 research or other fields to examine whether normality checking practices differ across fields and to explore whether there are any field-specific practices or extend the scope to journals of varying ranks and regional profiles in a shorter time window to provide a more comprehensive perspective. Moreover, researchers could review the inclusion of these practices in statistical textbooks or course materials to better understand why only a subset of these checking approaches is commonly used.

Conclusion

Normality remains a foundational statistical assumption in quantitative L2 research, yet it is insufficiently addressed or misapplied. Despite the existence of over 60 graphical and analytical methods for normality checking, our review found that L2 researchers rely on a narrow subset of these tools—primarily skewness, kurtosis, the Shapiro-Wilk (SW) test, and the Kolmogorov-Smirnov (KS) test—often without fully addressing the assumptions or constraints underlying their use. Many studies in the selected journals we examined did not report re-assessing normality after data transformation, paid limited attention to the influence of sample size on normality test sensitivity, or applied unclear or potentially inconsistent thresholds for interpreting results. In particular, critical missteps include the use of the SW test in large samples without verifying the coefficient estimation methods, the reliance on the KS test without graphical support, and the inconsistent or unstated application of criteria for skewness and kurtosis.

This disconnect between available statistical methods and current practices in L2 research suggests a need for clearer methodological guidance. By examining both graphical and analytical tests in detail and evaluating their sensitivity to various forms of non-normality, this study provides a nuanced understanding of how different methods function and under what conditions they yield meaningful results. Moreover, statistical software like SPSS v30+ and R offer all of the 12 univariate normality tests reviewed in this paper, making accessibility no longer a barrier for future researchers.

Although we initially expected that focusing on higher-ranking, Q1 journals would yield an overly positive picture of normality-checking practices, we observed practices that diverge from current methodological recommendations in quite a few cases. However, we must clarify that these departures from recommended practices may not necessarily invalidate the analysis results. A considerable number of the studies applied more stringent criteria in normality checking, meaning that if their data met these stricter standards, they would also satisfy the more lenient criteria recommended in the methodological literature.

Finally, this study adopted an AI-assisted review approach, where ChatGPT-4o was involved in the preliminary paper selection and re-annotation of the data. Though it did not achieve a 100% success rate as reported in previous research (Hampson et al., 2025) and has limitations in dealing with data sample sizes involved in analysis when multiple

https://doi.org/10.1017/S0272263126101600 Published online by Cambridge University Press

sources of data were reported, it was still effective and reliable, achieving near-perfect inter-coder reliability with the human annotator and high absolute agreement rates. This AI-assisted approach could inform future studies with small research teams dealing with large data annotations.

We hope that our recommendations offer L2 researchers a useful framework for selecting and interpreting normality tests based on sample size, data characteristics, and analytical goals. Ultimately, improving the transparency and rigor of normality assessments will enhance the validity of statistical inferences in L2 research and reduce the risk of misinterpretation or inappropriate model choice.

Supplementary material. The supplementary material for this article can be found at http://doi.org/10.1017/S0272263126101600.

References

Abdi, H., & Molin, P. (2007). Lilliefors test of normality. In N. J. Salkind (Ed.), *Encyclopedia of measurement and statistics* (pp. 508–510). Sage.

Ahmadiyeh, F., Sajedi-Amin, S., Kafili-Hajlari, T., & Naseri, A. (2023). Roadmap for outlier detection in univariate linear calibration in analytical chemistry: Tutorial review. *Journal of Chemometrics*, *37*(1), e3460. https://doi.org/10.1002/cem.3460

Ahad, N.A., Yin, T.S., Othman, A.R., & Yaacob, C.R. (2011). Sensitivity of normality tests to non-normal data. *Sains Malaysiana*, *40*(6), 637–641. https://core.ac.uk/download/pdf/11491563.pdf

Ahn, H. (2021). From Interlanguage grammar to target grammar in L2 processing of definiteness as uniqueness. *Second Language Research*, *37*(1), 91–119. https://doi.org/10.1177/0267658319868003

Ahn, S., & Jiang, N. (2023). Can adult learners sense L2 emotional words automatically? The role of L2 use on the emotional Stroop effect. *Second Language Research*, *39*(4), 1265–1278. https://doi.org/10.1177/02676583221131256

Ajao, I. O., Obafemi, O. S., & Bolarinwa, F. A. (2018). Effects of sample size and dispersion on quantile based plots for detecting normality. In *Edited Proceedings of 2nd International Conference* (Vol. *2*, pp. 432–438). Professional Statisticians Society of Nigeria.

Akbarian, I., & Elyasi, F. (2023). Developing EFL learners' collocational knowledge in self-regulated flipped classroom: A mixed-methods study. *Chinese Journal of Applied Linguistics*, *46*(4), 562–586. https://doi.org/10.1515/CJAL-2023-0405

Altman, N., & Krzywinski, M. (2016). Regression diagnostics. *Nature Methods*, *13*(5), 385–386. doi:10.1038/nmeth.3854

Altman, D. G., & Bland, J. M. (1995). Statistics notes: The normal distribution. *BMJ*, *310*, 298.

Anderson, T. W., & Darling, D. A. (1954). A test of goodness-of-fit. *Journal of the American Statistical Association*, *49*(268), 765–769. https://doi.org/10.1080/01621459.1954.10501232

Armitage, P., & Colton, T. (Eds.). (1998). *Encyclopedia of biostatistics* (Vol. *2*, pp. 1759–1760). Wiley.

Arnastauskaitė, J., Ruzgas, T., & Bražėnas, M. (2021). An exhaustive power comparison of normality tests. *Mathematics*, *9*(7), 788. https://doi.org/10.3390/math9070788

Aryadoust, V., & Raquel, M. R. (Eds.). (2019). *Quantitative data analysis for language assessment, Volume I: Fundamental techniques*. Routledge.

Baghban, A. A., Younespour, S., Jambarsang, S., Yousefi, M., Zayeri, F., & Jalilian, F. A. (2013). How to test normality distribution for a variable: A real example and a simulation study. *Archives of Advances in Biosciences*, *4*(1). https://doi.org/10.22037/jps.v4i1.3974

Balanda, K. P., & MacGillivray, H. L. (1988). Kurtosis: A critical review. *The American Statistician*, *42*, 111–119. https://doi.org/10.2307/2684482

Bayoud, H. A. (2021). Tests of normality: New test and comparative study. *Communications in Statistics - Simulation and Computation*, *50*(12), 4442–4463. https://doi.org/10.1080/03610918.2019.1643883

Berghoff, R. (2022). L2 processing of filler-gap dependencies: Attenuated effects of naturalistic L2 exposure in a multilingual setting. *Second Language Research*, *38*(2), 373–393. https://doi.org/10.1177/0267658320945757

https://doi.org/10.1017/S0272263126101600 Published online by Cambridge University Press

Brown, J. D., & Rodgers, T. (2002). *Doing second language research*. Oxford University Press.
Bylund, E., Samuel, S., & Athanasopoulos, P. (2024). Crosslinguistic differences in food labels do not yield differences in taste perception. *Language Learning*, *74*(S1), 20–39. https://doi.org/10.1111/lang.12641
Chatterjee, S. (2021). A new coefficient of correlation. *Journal of the American Statistical Association*, *116*(536), 2009–2022. https://doi.org/10.1080/01621459.2020.1758115
Cain, M. K., Zhang, Z., & Yuan, K.-H. (2017). Univariate and multivariate skewness and kurtosis for measuring nonnormality: Prevalence, influence and estimation. *Behavior Research Methods*, *49*(5), 1716–1735. https://doi.org/10.3758/s13428-016-0814-1
Choonpradub, C., & McNeil, D. (2005). Can the box plot be improved. *Songklanakarin Journal of Science and Technology*, *27*(3), 649–657. thaiscience.info/journals/Article/SONG/10986896.pdf
Correll, M., Li, M., Kindlmann, G., & Scheidegger, C. (2019). Looks good to me: Visualizations as sanity checks. *IEEE Transactions on Visualization and Computer Graphics*, *25*(1), 830–839. https://doi.org/10.1109/tvcg.2018.2864907
Cramér, H. (1928). On the composition of elementary errors: First paper: Mathematical deductions. *Scandinavian Actuarial Journal*, *1928*(1), 13–74. https://doi.org/10.1080/03461238.1928.10416862
D'Agostino, R. B., Belanger, A., & D'Agostino, R. B. (1990). A suggestion for using powerful and informative tests of normality. *The American Statistician*, *44*(4), 316–321. doi/abs/10.1080/00031305.1990.10475751
Darling, D. A. (1957). The Kolmogorov-Smirnov, Cramer-von Mises tests. *The Annals of Mathematical Statistics*, *28*(4), 823–838. https://doi.org/10.1214/aoms/1177706788
DeCarlo, L. (1997). On the meaning and use of kurtosis. On the meaning and use of kurtosis. *Psychological methods*, *2*(3), 292. doi:10.1037/1082-989X.2.3.292
Delignette-Muller, M. L., & Dutang, C. (2015). fitdistrplus: An *R* package for fitting distributions. *Journal of Statistical Software*, *64*(4). https://doi.org/10.18637/jss.v064.i04
Demir, S. (2022). Comparison of normality tests in terms of sample sizes under different skewness and kurtosis coefficients. *International Journal of Assessment Tools in Education*, *9*(2), 397–409. https://doi.org/10.21449/ijate.1101295
Doane, D. P., & Seward, L. E. (2011). Measuring skewness: A forgotten statistic? *Journal of Statistics Education*, *19*(2), 3. https://doi.org/10.1080/10691898.2011.11889611
Durbin, J., & Knott, M. (1972). Components of Cramér–Von Mises statistics. *I.* Journal of the Royal Statistical Society Series B: Statistical Methodology, *34*(2), 290–307. https://doi.org/10.1111/j.2517-6161.1972.tb00908.x
Emmanuel O. B., T. Maureen, N., & Wonu, N. (2020). Detection of non-normality in data sets and comparison between different normality tests. *Asian Journal of Probability and Statistics*, 1–20. https://doi.org/10.9734/ajpas/2019/v5i430149
Ernst, A. F., & Albers, C. J. (2017). Regression assumptions in clinical psychology research practice-a systematic review of common misconceptions. *PeerJ*, *5*, e3323. https://doi.org/10.7717/peerj.3323
Farrell, P. J., & Rogers-Stewart, K. (2006). Comprehensive study of tests for normality and symmetry: extending the Spiegelhalter test. *Journal of Statistical Computation and Simulation*, *76*(9), 803–816. https://doi.org/10.1080/10629360500109023
Field, A. (2018). *Discovering statistics using SPSS*. Sage Publications.
Fu, M., & Li, S. (2022). The effect of immediate and delayed corrective feedback on L2 development. *Studies in Second Language Acquisition*, *44*(1), 2–34. https://doi.org/10.1017/S0272263120000388
Gan, F. F., & Koehler, K. J. (1990). Goodness-of-Fit tests based on probability plots. *Technometrics*, *32*(3), 289–303. https://doi.org/10.1080/00401706.1990.10484682
Gass, S. (2009). A survey of SLA research. In W. Ritchie & T. Bhatia (eds.), *Handbook of second language acquisition* (pp. 3–28). Emerald.
George, D., & Mallery, P. (2001). *SPSS for Windows: 10.0 update*. Allyn & Bacon.
Gonulal, T., Loewen, S., & Plonsky, L. (2017). The development of statistical literacy in applied linguistics graduate students. *ITL - International Journal of Applied Linguistics*, *168*(1), 4–32. https://doi.org/10.1075/itl.168.1.01gon
González-Estrada, E., Villaseñor, J. A., & Acosta-Pech, R. (2022). Shapiro-Wilk test for multivariate skew-normality. *Computational Statistics*, *37*(4), 1985–2001. https://doi.org/10.1007/s00180-021-01188-y
Hair J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). *Multivariate data analysis* (8th edition). Cengage.

https://doi.org/10.1017/S0272263126101600 Published online by Cambridge University Press

Hampson, T., Cargos, K., & McKinley, J. (2025). Automate the 'boring bits': An assessment of AI-assisted systematic review (AIASR). *Research Methods in Applied Linguistics*, *4*(3), 100258. https://doi.org/10.1016/j.rmal.2025.100258

Hanzawa, K., & Suzuki, Y. (2023). Does automaticity in lexical and grammatical processing predict utterance fluency development?: A six-month longitudinal study in Japanese EFL context. *Journal of Second Language Studies*, *6*(2), 290–318. https://doi.org/10.1075/jsls.22007.han

Henderson, A. R. (2006). Testing experimental data for univariate normality. *Clinica Chimica Acta*, *366*(1–2), 112–129. https://doi.org/10.1016/j.cca.2005.11.007

Hanusz, Z., Enomoto, R., Seo, T., & Koizumi, K. (2017). A Monte Carlo comparison of Jarque–Bera type tests and Henze–Zirkler test of multivariate normality. *Communications in Statistics - Simulation and Computation*, *47*(5), 1439–1452. https://doi.org/10.1080/03610918.2017.1315771

Hatem, G., Zeidan, J., Goossens, M., & Moreira, C. (2022). Normality testing methods and the importance of skewness and kurtosis in statistical analysis. *BAU Journal - Science and Technology*, *3*(2). https://doi.org/10.54729/KTPE9512

Hernandez, H. (2021). Testing for normality: What is the best method. *ForsChem Research Reports*, *6*, 1–38. https://doi.org/10.13140/RG.2.2.13926.14406

Ho, A. D., & Yu, C. C. (2015). Descriptive statistics for modern test score distributions: Skewness, kurtosis, discreteness, and ceiling effects. *Educational and Psychological Measurement*, *75*(3), 365–388. https://doi.org/10.1177/0013164414548576

Hu, Y., & Plonsky, L. (2021). Statistical assumptions in L2 research: A systematic review. *Second Language Research*, *37*(1), 171–184. https://doi.org/10.1177/0267658319877433

Indira, V., Vasanthakumari, R., Sakthivel, N. R., & Sugumaran, V. (2011). A method for calculation of optimum data size and bin size of histogram features in fault diagnosis of mono-block centrifugal pump. *Expert Systems with Applications*, *38*(6), 7708–7717. https://doi.org/10.1016/j.eswa.2010.12.140

Jarque, C. M., & Bera, A. K. (1987). A test for normality of observations and regression residuals. *International Statistical Review / Revue Internationale de Statistique*, *55*(2), 163–172. https://doi.org/10.2307/1403192

Jelito, D., & Pitera, M. (2021). New fat-tail normality test based on conditional second moments with applications to finance. *Statistical Papers*, *62*(5), 2083–2108. https://doi.org/10.1007/s00362-020-01176-2

Jiménez-Gamero MD (2023) Testing normality of a large number of populations. *Statist Pap*. https://doi.org/10.1007/s00362-022-01384-y

Johannessen, J. B., Lundquist, B., Rodina, Y., Tengesdal, E., Kaldhol, N. H., Türker, E., & Fyndanis, V. (2025). Cross-linguistic effects in grammatical gender assignment and predictive processing in L1 Greek, L1 Russian, and L1 Turkish speakers of Norwegian as a second language. *Second Language Research*, *41*(1), 217–259. https://doi.org/10.1177/02676583241227709

Keren G, Lewis C (2014) A handbook for data analysis in the behaviorial sciences: volume *1* methodological issues volume 2: statistical issues. Psychology Press.

Keselman, H. J., Othman, A. R., & Wilcox, R. R. (2013). Preliminary testing for normality: Is this a good practice? *Journal of Modern Applied Statistical Methods*, *12*(2), 2–19. https://doi.org/10.22237/jmasm/1383278460

Khan, R. A., & Ahmad, F. (2017). Power comparison of various normality tests. *Pakistan Journal of Statistics and Operation Research*, *11*(3). https://doi.org/10.18187/pjsor.v11i3.1082

Khatun, N. (2021). Applications of normality test in statistical analysis. *Open Journal of Statistics*, *11*(01), 113–122. https://doi.org/10.4236/ojs.2021.111006

Kim H. Y. (2013). Statistical notes for clinical researchers: assessing normal distribution (2) using skewness and kurtosis. *Restorative dentistry & endodontics*, *38*(1), 52–54. https://doi.org/10.5395/rde.2013.38.1.52

Kim, N. (2021). A Jarque-Bera type test for multivariate normality based on second-power skewness and kurtosis. *Communications for Statistical Applications and Methods*, *28*(5), 463–475. https://doi.org/10.29220/CSAM.2021.28.5.463

Kline, R. (2016). *Principles and practice of structural equation modeling* (4th ed.). The Guilford Press.

Knief, U., & Forstmeier, W. (2021). Violating the normality assumption may be the lesser of two evils. *Behavior Research Methods*, *53*(6), 2576–2590. https://doi.org/10.3758/s13428-021-01587-5

Kwak, S. G., & Kim, J. H. (2017). Central limit theorem: The cornerstone of modern statistics. *Korean Journal of Anesthesiology*, *70*(2), 144. https://doi.org/10.4097/kjae.2017.70.2.144

https://doi.org/10.1017/S0272263126101600 Published online by Cambridge University Press

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, *33*(1), 159. https://doi.org/10.2307/2529310

Lanzante, J. R. (2021). Testing for differences between two distributions in the presence of serial correlation using the Kolmogorov–Smirnovand Kuiper's tests. *International Journal of Climatology*, *41*(14), 6314–6323. https://doi.org/10.1002/joc.7196

Le Cessie, S., Goeman, J. J., & Dekkers, O. M. (2020). Who is afraid of non-normal data? Choosing between parametric and non-parametric tests. *European Journal of Endocrinology*, *182*(2), E1–E3. https://doi.org/10.1530/eje-19-0922

Lilliefors, H. W. (1967). On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American statistical Association*, *62*(318), 399–402. https://doi.org/10.2307/2283970

Lindstromberg, S. (2016). Inferential statistics in *Language Teaching Research*: A review and ways forward. *Language Teaching Research*, *20*, 741–768. https://doi.org/10.1177/1362168816649979

Loewen, S., Lavolette, E., Spino, L. A., Papi, M., Schmidtke, J., Sterling, S., & Wolff, D. (2014). Statistical Literacy Among Applied Linguists and Second Language Acquisition Researchers. *TESOL Quarterly*, *48*(2), 360–388. https://doi.org/10.1002/tesq.128

Mala, I., Sladek, V., & Bilkova, D. (2021). Power comparisons of normality tests based on L-moments and classical tests. *Mathematics and Statistics*, *9*(6), 994–1003. https://doi.org/10.13189/ms.2021.090615

Marange, C. S., & Qin, Y. (2021). An empirical likelihood ratio-based omnibus test for normality with an adjustment for symmetric alternatives. *Journal of Probability and Statistics*, *2021*, 1–18. https://doi.org/10.1155/2021/6661985

Mayers, A. (2013). *Introduction to statistics and SPSS in psychology*. Pearson Education Limited.

Mbah, A. K., & Paothong, A. (2015). Shapiro–Francia test compared to other normality test using expected *p*-value. *Journal of Statistical Computation and Simulation*, *85*(15), 3002–3016. https://doi.org/10.1080/00949655.2014.947986

Mishra, P., Pandey, C., Singh, U., Gupta, A., Sahu, C., & Keshri, A. (2019). Descriptive statistics and normality tests for statistical data. *Annals of Cardiac Anaesthesia*, *22*(1), 67. https://doi.org/10.4103/aca.ACA_157_18

Nakata, T., Suzuki, Y., & He, X. (Stella). (2023). Costs and benefits of spacing for second language vocabulary learning: Does relearning override the positive and negative effects of spacing? *Language Learning*, *73*(3), 799–834. https://doi.org/10.1111/lang.12553

Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (2005). *Applied Linear Statistical Models*. (5th ed.). McGraw-Hill Education.

Nielsen, E. E., Nørskov, A. K., Lange, T., Thabane, L., Wetterslev, J., Beyersmann, J., De Unã-Álvarez, J., Torri, V., Billot, L., Putter, H., Winkel, P., Gluud, C., & Jakobsen, J. C. (2019). Assessing assumptions for statistical analyses in randomised clinical trials. *BMJ Evidence-Based Medicine*, *24*(5), 185–189. https://doi.org/10.1136/bmjebm-2019-111174

Nosakhare, U. H., Bright, A. F. (2017). Evaluating of techniques for univariate normality test using monte carlo simulation. *American Journal of Theoretical and Applied Statistic*, *6*(5-1), 51–61. 10.11648/j.ajtas.s.2017060501.18

Okeniyi, J. O., Okeniyi, E. T., & Atayero, A. A. (2020). *Implementation of data normality testing as a Microsoft Excel® library function by Kolmogorov-Smirnov goodness-of-fit statistics.* The 23rd International Business Information Management Association Conference, IBIMA 2014: Vision 2020: Sustainable Growth, Economic Development, and Global Competitiveness.

OpenAI. (2024, May 13). *Hello GPT-4o.* https://openai.com/index/hello-gpt-4o/

Oppong, F. B., & Agbedra, S. Y. (2016). Assessing univariate and multivariate normality, a guide for non statisticians. *Mathematical Theory and Modeling*, *6*(2), 26–33. files.core.ac.uk/download/pdf/234680353.pdf

Otwinowska, A., Foryś-Nogala, M., Kobosko, W., & Szewczyk, J. (2020). Learning orthographic cognates and non-cognates in the classroom: Does awareness of cross-linguistic similarity matter? *Language Learning*, *70*(3), 685–731. https://doi.org/10.1111/lang.12390

Öztuna, D., Elhan, A. H., & Tüccar, E. (2006). Investigation of four different normality tests in terms of type 1 error rate and power under different distributions. *Turkish Journal of Medical Sciences*, *36*(3), 171–176. https://journals.tubitak.gov.tr/medical/vol36/iss3/7/

Park, H.M. (2008). *Univariate analysis and normality test using SAS, Stata, and SPSS.* Technical Working Paper. The University Information Technology Services (UITS) Center for Statistical and Mathematical Computing, Indiana University.

https://doi.org/10.1017/S0272263126101600 Published online by Cambridge University Press

Parsons, N. R., Price, C. L., Hiskens, R., Achten, J., & Costa, M. L. (2012). An evaluation of the quality of statistical design and analysis of published medical research: Results from a systematic survey of general orthopaedic journals. *BMC Medical Research Methodology*, *12*(1), 60. https://doi.org/10.1186/1471-2288-12-60

Patrício, M., Ferreira, F., Oliveiros, B., & Caramelo, F. (2017). Comparing the performance of normality tests with ROC analysis and confidence intervals. *Communications in Statistics - Simulation and Computation*, *46*(10), 7535–7551. https://doi.org/10.1080/03610918.2016.1241410

Pek, J., Wong, O., & Wong, A. C. M. (2018). How to address non-normality: A taxonomy of approaches, reviewed and illustrated. *Frontiers in Psychology*, *9*, 2104. https://doi.org/10.3389/fpsyg.2018.02104

Pek, J., Wong, O., & Wong, A. C. M. (2017). Data transformations for inference with linear regression: Clarifications and recommendations. *Practical Assessment, Research, and Evaluation*, *22*(1), Article 9. https://doi.org/10.7275/2w3n-0f07

Pettitt, A. N. (1977). Testing the Normality of Several Independent Samples Using the Anderson-Darling Statistic. *Applied Statistics*, *26*(2), 156. https://doi.org/10.2307/2347023

Plonsky, L, & Ghanbar, H. (2018). Multiple regression in L2 research: A methodological synthesis and guide to interpreting R2 values. *Modern Language Journal*, *102*, 713–31. https://doi.org/10.1111/modl.12509

Plonsky, L. (2024). Study quality as an intellectual and ethical imperative: A proposed framework. *Annual Review of Applied Linguistics*, *44*, 4–18. doi:10.1017/S0267190524000059

Rani Das, K. (2016). A brief review of tests for normality. *American Journal of Theoretical and Applied Statistics*, *5*(1), 5. https://doi.org/10.11648/j.ajtas.20160501.12

Raymaekers, J., & Rousseeuw, P. J. (2024). Transforming variables to central normality. *Machine Learning*, *113*(12), 4953–4975. https://doi.org/10.1007/s10994-021-05960-5

Razali, N. M., and Wah, Y. B. (2011). Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and Anderson-darling tests. *Journal of Statistical Modeling and Analytics 2*, 21–33. nrc.gov/docs/ml1714/ml17143a100.pdf

Rodu, J., & Kafadar, K. (2022). The q–q Boxplot. *Journal of Computational and Graphical Statistics*, *31*(1), 26–39. https://doi.org/10.1080/10618600.2021.1938586

Royston, J. P. (1982). An extension of Shapiro-Wilk's W test for non-normality to large samples. *Applied Statistics*, *31*, 115–124. https://doi.org/10.2307/2347973

Royston, P. (1992). Approximating the Shapiro-Wilk W-test for non-normality. *Statistics and computing*, *2*(3), 117–119. https://doi.org/10.1007/BF01891203

Saito, K. (2020). Multi- or single-word units? The role of collocation use in comprehensible and contextually appropriate second language speech. *Language Learning*, *70*(2), 548–588. https://doi.org/10.1111/lang.12387

Schoder, V., Himmelmann, A., & Wilhelm, K. P. (2006). Preliminary testing for normality: some statistical aspects of a common concept. *Clinical and experimental dermatology*, *31*(6), 757–761. https://doi.org/10.1111/j.1365-2230.2006.02206.x

Shapiro, S. S., & Francia, R. S. (1972). An approximate analysis of variance test for normality. *Journal of the American Statistical Association*, *67*(337), 215–216. https://doi.org/10.1080/01621459.1972.10481232

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/10.2307/2333709

Shatz, I. (2023). Assumption-checking rather than (just) testing: The importance of visualization and effect size in statistical diagnostics. *Behavior Research Methods*, *56*(2), 826–845. https://doi.org/10.3758/s13428-023-02072-x

Steinskog, D. J., Tjøstheim, D. B., & Kvamstø, N. G. (2007). A Cautionary Note on the Use of the Kolmogorov–Smirnov Test for Normality. *Monthly Weather Review*, *135*(3), 1151–1157. https://doi.org/10.1175/MWR3326.1

Tabachnick B. G., Fidell L. S. (2013) *Using multivariate statistics*, 6th ed. Boston, MA: Pearson.

Thadewald, T., & Büning, H. (2007). Jarque–Bera Test and its Competitors for Testing Normality – A Power Comparison. *Journal of Applied Statistics*, *34*(1), 87–105. https://doi.org/10.1080/02664760600994539

Thode, H. C. (2002). *Testing for normality*. Marcel Dekker, Inc.

Tukey, J. W. (1977). *Exploratory data analysis*. Addison-Wesley.

Uhm, T., & Yi, S. (2021). A comparison of normality testing methods by empirical power and distribution of *P*-values. *Communications in Statistics - Simulation and Computation*, *52*(9), 4445–4458. https://doi.org/10.1080/03610918.2021.1963450

https://doi.org/10.1017/S0272263126101600 Published online by Cambridge University Press

Ventura-León, J., Peña-Calero, B. N., & Burga-León, A. (2023). The effect of normality and outliers on bivariate correlation coefficients in psychology: A Monte Carlo simulation. *The Journal of General Psychology*, *150*(4), 405–422. https://doi.org/10.1080/00221309.2022.2094310

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

Wang, J., & Halenko, N. (2024). Developing second language mandarin fluency through pedagogic intervention and study abroad: Planning time, speech rate, and response duration. *Language Learning*, *74*(S2), 177–206. https://doi.org/10.1111/lang.12694

Weinberg, S. L., & Abramowitz, S. K. (2002). *Data analysis for the behavioral sciences using SPSS*. Cambridge University Press.

West, S. G., Finch, J. F., & Curran, P. J. (1995). Structural equation models with nonnormal variables: Problems and remedies. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 56–75). Sage Publications.

Wijekularathna, D. K., Manage, A. B. W., & Scariano, S. M. (2019). Power analysis of several normality tests: A Monte Carlo simulation study. *Communications in Statistics - Simulation and Computation*, *51*(3), 757–773. https://doi.org/10.1080/03610918.2019.1658780

Wilcox, R. R. (2010). *Fundamentals of modern statistical methods: Substantially improving power and accuracy* (2nd ed.). Springer.

Wilk, M. B., & Gnanadesikan, R. (1968). Probability plotting methods for the analysis of data. *Biometrika*, *55*(1), 1. https://doi.org/10.2307/2334448

Yap, B. W., & Sim, C. H. (2011). Comparisons of various types of normality tests. *Journal of Statistical Computation and Simulation*, *81*(12), 2141–2155. https://doi.org/10.1080/00949655.2010.520163

Zhang, P., & Han, C. (2024). Examining statistical literacy, attitudes toward statistics, and statistics self-efficacy among applied linguistics research students in China. *International Journal of Applied Linguistics*, *34*(2), 433–449. https://doi.org/10.1111/ijal.12500

Cite this article: Aryadoust, V., & Jia, Y. (2026). Univariate normality checking practices in L2 research: An AI-assisted systematic review. *Studies in Second Language Acquisition,* 48: 535–570. https://doi.org/10.1017/S0272263126101600

https://doi.org/10.1017/S0272263126101600 Published online by Cambridge University Press