The International Journal of Indian Psychology
ISSN 2348-5396 (Online) | ISSN: 2349-3429 (Print)
Volume 14, Issue 3, July- September, 2026
DIP: 18.01.077.20261403, DOI: 10.25215/1403.077
https://www.ijip.in

Research Paper

Normality Ignored: The Misuse of Parametric Statistics in Postgraduate Psychology Research

Pooja Tyagi[1], Ashish Budhwar[2]*

ABSTRACT

Parametric procedures such as the t-test, ANOVA, Pearson's r, and ordinary least squares regression anchor most quantitative work submitted at the postgraduate level in psychology. Students find them familiar, the software makes them painless, and reviewers expect them. Each procedure carries distributional commitments, though, and normality is the one most often invoked and least often examined with care. This commentary sets out how that neglect appears in dissertation practice, why it persists, and what might be done about it. Five recurring problems are described: assumption checks are skipped outright; skewness and kurtosis thresholds are read in isolation; Shapiro–Wilk output is treated as a verdict rather than as evidence; single-item Likert responses are analysed as interval scores without justification; and sample sizes above an arbitrary cut-off are taken to settle the question of distributional fit. These habits are then placed against the institutional pressures that sustain them, including reliance on departmental templates, the way software fluency is mistaken for statistical reasoning, and the continuing premium attached to statistical significance. Recommendations follow at programme level and at supervisor level, with attention to transparent reporting and to the difference between teaching thresholds and teaching judgement. The argument throughout is that assumption checking belongs inside the analysis rather than alongside it, and that postgraduate training in psychology should be redesigned with that point in mind.

Keywords: *Normality Assumption, Parametric Tests, Statistical Literacy, Dissertation Methodology, Postgraduate Research, Robustness*

A handful of inferential procedures account for most of what gets reported in empirical psychology: the independent-samples t-test, one-way analysis of variance, Pearson's correlation coefficient, and multiple linear regression. Each one is well understood, widely taught, and reasonably easy to run on the software students already have. Each one also rests on a set of assumptions, and within that set the normality requirement is the one cited most frequently and misread most often (Ghasemi & Zahediasl, 2012).

What the word "normal" picks out shifts depending on the procedure. A t-test or ANOVA assumes approximate normality of scores within each group. Pearson's r presupposes a

[1]Assistant Professor, Department of Psychology, Guru Nanak Dev University, Amritsar
[2]Senior Research Fellow, Department of Psychology, Guru Nanak Dev University, Amritsar
*Corresponding Author

Received: June 06, 2026; Revision Received: August 02, 2026; Accepted: August 07, 2026

bivariate normal joint distribution between the two variables under study. Ordinary least squares regression places the assumption on the residuals rather than on the raw outcome. When data depart sharply from these conditions, p-values can drift away from their nominal level, confidence intervals can mislead, and parameter estimates can shift in directions that are easy to miss without diagnostic plots (Knief & Forstmeier, 2021).

In dissertation work the choice of a parametric procedure is too often a default rather than a reasoned decision. Students lean on supervisor-supplied templates, copy methodology from the previous cohort's exemplar dissertations, and reproduce whatever the lab has tended to use in the past (Field, 2018). Diagnostic testing is then either skipped or reduced to a single line of SPSS output that nobody returns to. The consequences extend well beyond one bad paper. Studies that ignore distributional checks inflate the rate at which spurious findings make it into the literature, weaken the credibility of student work in the eyes of external examiners, and, more damagingly, teach a generation of new researchers that assumption testing is decoration rather than analysis.

The present commentary sets out the most common shapes the problem takes, the institutional conditions that keep it alive in graduate programmes, and a series of practical adjustments that supervisors, programme leads, and students themselves can make.

Why Normality Matters (and When It Matters Less)

Most classical parametric methods come with a normality clause attached. ANOVA and the t-test require group-wise approximate normality; Pearson's r presumes a bivariate normal joint distribution; ordinary least squares regression assumes residuals are normally distributed with constant variance. None of these conditions holds exactly in practice. The real question, then, is whether the gap between the data and the textbook ideal is large enough to matter for the inference being drawn.

In many studies it is not. The Central Limit Theorem ensures that, given a sample of reasonable size and roughly balanced groups, the sampling distributions of means and regression coefficients are approximately normal even when the underlying variables are skewed or heavy-tailed (Lumley et al., 2002). Monte Carlo work has repeatedly shown that ANOVA holds its nominal Type I error rate across a fairly wide range of non-normal conditions (Schmider et al., 2010), and recent simulation evidence suggests that fitting Gaussian models to non-normal data is, in most situations, less damaging than the more elaborate alternatives often proposed in their place (Knief & Forstmeier, 2021). Pronounced departures, particularly heavy skew combined with influential outliers, can still distort inference in small or unbalanced samples; the appropriate response is to scale the reaction to the severity of the violation rather than to react to any departure at all.

Two attitudes common in student work pull against this calibration. The first dismisses normality outright on the grounds that real data are never normal anyway, and proceeds with parametric tests as though the assumption were a dead letter. The second treats any visible departure as fatal and reaches for a non-parametric test on reflex, regardless of how mild the deviation actually is. Neither stance resembles how applied statisticians weigh distributional evidence.

Common Errors in Practice

The errors that turn up in postgraduate dissertations fall into a small set of recognisable patterns. Table 1 summarises them, and each is discussed in turn below.

Table 1 Common errors in normality assessment and the practice each one substitutes

Common error	What students do	What the error misses	Better practice
Skipping the check	Proceeding from descriptives straight to inferential output, with no diagnostic reported in text or figure.	Even a quick histogram can flag obvious problems; "n is large" is not a substitute for looking at the distribution.	Report at least one numerical and one graphical diagnostic per analysis.
Skewness or kurtosis cut-offs read in isolation	Comparing values to ±1 or ±2 thresholds and stopping there.	Summary indices conceal bimodality, ceiling effects, and outliers; they are also sensitive to sample size.	Combine indices with Q–Q plots and interpret in light of n.
Misreading formal tests	Taking the Shapiro–Wilk p-value as a final verdict on normality.	Power scales with n: large samples flag trivial deviations; small samples miss substantial ones.	Read Shapiro–Wilk alongside graphical evidence and effect magnitude.
Ordinal data treated as interval	Applying parametric tests to single Likert items without argument.	Equal-distance between response categories is unsupported for a single ordinal item.	Use non-parametric methods for single items; justify treatment of multi-item composites.
The "large-n excuses everything" view	Skipping diagnostics for any sample above an arbitrary cut-off, often 30.	The CLT applies to sampling distributions, not raw data; outliers and heteroscedasticity remain.	Run diagnostics regardless of n; let findings, not sample size, drive decisions.

Skipping the Check Entirely

The simplest failure is the most basic one: no check appears at all. Many dissertations move from descriptive statistics straight to inferential output, with no histogram, no Q–Q plot, no formal test, and no narrative line acknowledging that anything was examined. The justification, when one is offered, usually comes back to sample size: the n is "large enough", the writer says, and the question is treated as closed. This trades on a misreading of the Central Limit Theorem that I will come back to below.

Skewness and Kurtosis as Shortcuts

A second pattern is to report values of skewness and kurtosis, compare them to memorised thresholds (usually the textbook cut-offs of ±1 or ±2), and stop. The indices have their uses, but they are not sufficient on their own. A distribution can sit politely within the cut-offs while concealing bimodality, ceiling effects, or one or two influential outliers that a plot would show at a glance. The reverse problem also turns up: with very large samples almost any deviation pushes the indices beyond the textbook threshold (Blanca et al., 2013), which then prompts unnecessary abandonment of parametric procedures that would have worked perfectly well.

Misreading the Formal Tests

The Shapiro–Wilk and Kolmogorov–Smirnov tests are widely used and very often misinterpreted. Both gain statistical power as n grows. In samples of three or four hundred they reliably flag departures so trivial that they have no consequence for the analysis at all. In samples of fifteen or twenty they routinely miss genuine problems. Treating the p-value as a single arbiter therefore produces false alarms in one direction and false reassurance in the other (Ghasemi & Zahediasl, 2012). A Shapiro–Wilk significant at $p < .001$ on $n = 400$ says next to nothing about whether the t-test that follows it is in trouble; a non-significant test on $n = 18$ says even less about whether the analysis is safe.

Ordinal Data Treated as Interval

Likert items are, by construction, ordinal. They are nonetheless subjected as a matter of course to means, standard deviations, and parametric tests, with no discussion of the level of measurement they actually achieve. Multi-item scales with strong internal consistency can approximate interval properties closely enough for parametric work to be defensible (Norman, 2010), and Norman's argument is now widely cited in support of that practice. The same case cannot be made for single items, where the assumption that the psychological distance from "strongly disagree" to "disagree" equals the distance from "agree" to "strongly agree" has no obvious basis and is rarely defended in the text.

The Large-Sample Shrug

Closely related is the assumption that any sample exceeding some round number, with 30 the figure that appears most often, immunises the analysis against distributional concerns. This is the reading of the Central Limit Theorem that I deferred earlier. The theorem guarantees the asymptotic normality of certain sampling distributions, not of the data themselves. It says nothing about leverage, nothing about heteroscedasticity, and nothing about the influence of a single extreme value sitting in the corner of the scatterplot. Once an analysis is more elaborate than a simple difference of means, the $n = 30$ talisman starts to look thin.

Consequences for Reported Inference

When these errors are not corrected, the resulting inferences carry characteristic distortions, and the distortions vary by procedure. In group comparisons, whether using a t-test or ANOVA, skewed distributions combined with unequal group sizes can produce a significant p-value driven by distributional artefacts rather than by a genuine difference in location, yet the result is then reported as straightforward evidence of an effect. In correlational analyses, Pearson's r is chosen by default in places where Spearman's rank coefficient would be the more sensible choice, often because one or both variables are heavily skewed, ordinal, or marked by outliers; the coefficient that emerges can overstate or understate the true association by a noticeable margin (Gravetter & Wallnau, 2017). The pattern in regression is the most exposed of the three. Dissertations frequently report coefficients, R^2, and significance tests with no examination of residuals, no leverage statistics, no comment on homoscedasticity, and no attention to multicollinearity. What looks like a clean model on the page is sitting on a foundation that was never tested.

Why the Pattern Persists

The persistence of these problems is not primarily an information problem. The relevant material appears in every introductory text the students are assigned. The pattern is sustained instead by features of how postgraduate research is taught, supervised, and evaluated.

Imitation Rather Than Reasoning

Students model their methods sections on the dissertations the department holds up as exemplars. Where the exemplars themselves skip diagnostics, the omission replicates intact into the next year's cohort. The implicit lesson is that the method is whatever the previous successful student happened to do, and that doing what was passed will be enough to pass again.

Software Fluency as a Stand-In for Understanding

Most postgraduate programmes invest a substantial share of their methods teaching in showing students how to navigate menus in SPSS or jamovi. Considerably less time is given to the conceptual machinery sitting behind those menus: which test is appropriate to which design, what the assumptions actually imply about the data, and what to do when a diagnostic plot looks awkward (Pallant, 2020). The outcome is competence at producing output paired with limited ability to evaluate it.

The Premium on Statistical Significance

The third factor is incentive. The reward structure surrounding graduate research, in which significant p-values are read as evidence of a successful project, biases analytic choice toward whatever procedure looks most likely to deliver them. Parametric tests are sometimes preferred for the wrong reasons here: they are perceived as more rigorous than their non-parametric counterparts and as more likely to clear the conventional .05 threshold, even where the data make them the less defensible choice.

Recommendations

Combine Numerical, Formal, and Graphical Diagnostics

No single indicator should be allowed to drive the decision. Skewness and kurtosis describe distributional shape; formal tests provide a hypothesis-testing perspective that must be read in light of sample size; histograms, kernel density plots, and Q–Q plots reveal features that the summaries hide. Read together, the three sources produce a defensible reading of the data; read in isolation, each one misleads in its own direction.

Teach Robustness, Not Rules

Students benefit more from understanding that classical parametric procedures are reasonably robust under mild violations than from memorising thresholds they will then misapply. The teaching aim should be judgement: knowing when a deviation can be tolerated, when it warrants a robust alternative, and when it points to something deeper about the data, such as a measurement problem or a mixed population (Schmider et al., 2010).

Use Non-Parametric and Robust Methods on Their Merits

The Mann–Whitney U, the Kruskal–Wallis test, Spearman's rank correlation, bootstrap confidence intervals, and robust regression are not consolation prizes for failed parametric analyses. They are appropriate tools when the data fall outside the conditions for which classical methods were designed. Framing them as second-best, which postgraduate teaching often implicitly does, is itself part of the problem.

Redesign Quantitative Training

Postgraduate methods courses should build conceptual reasoning alongside software training. Justification of analytic choices, interpretation of diagnostic output, and clear communication of uncertainty deserve the same emphasis that is currently given to running

the test in the first place. Supervisors carry the largest share of the responsibility here. By asking students to defend their choices and by modelling thoughtful practice in their own feedback, they shape what counts as acceptable methodology for the next cohort.

Require Transparent Reporting

Dissertation guidelines should require students to document which assumptions were checked, by what means, what the results were, and what decision followed. Such reporting improves the credibility of the work and, through repetition across cohorts, reinforces the message that diagnostics are part of the analysis rather than something appended to it (Pallant, 2020).

CONCLUSION

Normality is neither sacred nor irrelevant. Treating it as either, an inflexible gatekeeping criterion on one side or an outdated formality on the other, produces predictable failures, and postgraduate psychology research displays both. The deeper problem is that assumption checking has come to feel like paperwork: a hurdle to clear before the real analysis begins, rather than part of the reasoning that makes the analysis defensible in the first place.

Repairing the practice is mostly a matter of training. If students are taught to look at their distributions before applying tests to them, to read diagnostic output in context rather than as a verdict, and to report what they did and why, the technical errors largely take care of themselves. The harder work is cultural and falls primarily on programmes and supervisors: to stop treating diagnostics as decoration and to start treating them as the evidence on which the inferential claims rest.

REFERENCES

Blanca, M. J., Arnau, J., López-Montiel, D., Bono, R., & Bendayan, R. (2013). Skewness and kurtosis in real data samples. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 9*(2), 78–84. https://doi.org/10.1027/1614-2241/a000057

Field, A. (2018). *Discovering statistics using IBM SPSS Statistics* (5th ed.). SAGE Publications.

Ghasemi, A., & Zahediasl, S. (2012). Normality tests for statistical analysis: A guide for non-statisticians. *International Journal of Endocrinology and Metabolism, 10*(2), 486–489. https://doi.org/10.5812/ijem.3505

Gravetter, F. J., & Wallnau, L. B. (2017). *Statistics for the behavioral sciences* (10th ed.). Cengage Learning.

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

Lumley, T., Diehr, P., Emerson, S., & Chen, L. (2002). The importance of the normality assumption in large public health data sets. *Annual Review of Public Health, 23*(1), 151–169. https://doi.org/10.1146/annurev.publhealth.23.100901.140546

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

Pallant, J. (2020). *SPSS survival manual: A step by step guide to data analysis using IBM SPSS* (7th ed.). Open University Press.

Schmider, E., Ziegler, M., Danay, E., Beyer, L., & Bühner, M. (2010). Is it really robust? Reinvestigating the robustness of ANOVA against violations of the normal distribution assumption. *Methodology: European Journal of Research Methods for*

the Behavioral and Social Sciences, 6(4), 147–151. https://doi.org/10.1027/1614-2241/a000016

Acknowledgment
The author(s) appreciates all those who participated in the study and helped to facilitate the research process.

Conflict of Interest
The author(s) declared no conflict of interest.

How to cite this article: Tyagi, P. & Budhwar, A. (2026). Normality Ignored: The Misuse of Parametric Statistics in Postgraduate Psychology Research. *International Journal of Indian Psychology*, *14*(3)*,* 921-927. DIP:18.01.077.20261403, DOI:10.25215/1403.077