OPEN EDUCATIONAL RESOURCE

How to Write the Results Section: Quantitative Results in a Dissertation

Derek Jansen, MBA
Grad Coach
Reviewed by Kerryn Warren, PhD · November 2026

Abstract

The results section of a quantitative dissertation or thesis reports what the data show. Two problems recur in it: output from the statistics software reported without selection, and interpretation that belongs in the discussion. This article sets out what a quantitative results section contains and in what order: the flow of participants and the description of the sample, the reliability of the measures, the checks on missing data, outliers and the shape of the data, the descriptive statistics, and the inferential tests and hypotheses. It explains how each element is reported, with reference to the American Psychological Association's reporting standards and its rules for numbers and statistics, including exact *p*-values, effect sizes and confidence intervals. Extracts from four master's and doctoral theses, in languages and cultures, business, clinical psychology and human-environment relations, show how a sample, a reliability check, a normality check, a descriptive table and a hypothesis test were reported in practice. An illustrative write-up shows how to report a supported and an unsupported hypothesis. The article is intended for students and early-career researchers writing up a quantitative or survey study, and for the educators who supervise and teach them.

Keywords: how to write the results section of a dissertation; dissertation results section example; results chapter dissertation; how to present survey results in a dissertation; reporting statistics; hypothesis testing; effect size; open educational resource

Suggested citation: Jansen, D. (2026). *How to write the results section: Quantitative results in a dissertation*. Grad Coach. https://doi.org/10.5281/zenodo.24198378

About this article: This is the formal edition of an article first published at gradcoach.com/quantitative-results-analysis/, where the figures can also be downloaded for reuse. Further open educational resources on research methods and statistics are listed at gradcoach.com/open-educational-resources/.

© 2026 Grad Coach. Free to share and adapt under CC BY 4.0 with credit to Grad Coach and a link to gradcoach.com/quantitative-results-analysis/. Quoted thesis extracts remain their authors'; each thesis keeps its own license.

1. Introduction

The results section is where an examiner, committee or reviewer checks whether the data answer the research questions. In a thesis it is often a chapter of its own, also called the findings or analysis chapter, and in a conventional structure it sits between the methodology and the discussion. Two kinds of results section are hard to examine. One reproduces the output of the statistics software, test after test, and leaves the reader to find the answers. The other begins explaining the findings before they have been fully reported, so that it reads as a discussion chapter.

The analysis itself is seldom the difficult part of writing up. The harder decisions are what to report and in what order. This article addresses both for a quantitative study, including a survey. Section 2 separates the results from the discussion. Section 3 sets out how to choose what to report and the usual order of a results section. Sections 4 to 7 take that order in turn: the sample, the checks on the measures and the data, the descriptive statistics, and the inferential tests and hypotheses. Section 8 summarizes the conventions for writing the section. Throughout, extracts from four theses show how each element was reported in practice. Qualitative findings are reported differently and are not covered here.

2. Results and Discussion

The results section **presents the findings**: the numbers and what they show, no more and no less. The discussion **interprets them**, relating them to prior research, the research questions and the limitations of the study. The American Psychological Association's journal article reporting standards for quantitative research (JARS–Quant) draw the same line. They list the statistics to report under the results, and under the discussion they list the interpretation of the results, the comparison with the work of others and the generalizability of the findings (Appelbaum et al., 2018, p. 8).

A master's thesis in human-environment relations from Cornell University shows the difference clearly. The study compared a traditional luxury store with a "phygital" store, a physical store with digital elements, in an experiment with 280 participants. One hypothesis, H2a, proposed that the store's perceived usefulness would explain (mediate) the effect of store type on purchase intention. The results chapter reports the outcome as follows:

> The mediating effect for perceived usefulness was not significant, as the 95% confidence interval (CI) for the indirect effect included zero (b = .05, BootSE = .04, BootCI = ([-0.01 to 0.16]). Thus, H2a is not supported.
>
> Verdiguel (2024, pp. 30–31)

The discussion chapter returns to the same finding:

> Conversely, perceived usefulness, despite being higher in phygital stores, did not significantly mediate this relationship. This suggests that the practical benefits of phygital elements are secondary to their hedonic appeal in influencing luxury purchase intentions. Previous research has found that while utilitarian motives may motivate consumers to visit a store, the fulfillment of hedonic motives has a greater impact on purchasing behaviors (Dawson et al., 1990; Khan & Dhar, 2006).
>
> Verdiguel (2024, p. 33)

© 2026 Grad Coach · CC BY 4.0 · gradcoach.com

The results version gives the statistic and a verdict on the hypothesis, with no citations and no "this suggests". The discussion version drops the numbers, proposes what the finding might mean and brings in prior research. A practical test follows: **a sentence that explains why a result occurred belongs in the discussion**.

Some programs ask for a combined "Results and Discussion" chapter, as does a thesis from Purdue University quoted in Section 4. Even then, the findings are best reported in full before they are interpreted, so that a reader can see what the data showed before reading what the author makes of it. The program's handbook is the first place to check which structure is expected.

3. Deciding What to Report and in What Order

3.1 The research questions as the filter

Analysis produces far more output than a results section needs. Exploratory runs, alternative tests and dozens of charts were useful for understanding the data, but they do not all belong in the write-up. The filter is the purpose of the study: a result belongs in the section if it helps answer a research question or test a hypothesis. A practical way to apply the filter is to list each research question and, beside it, the analyses that address it, then to outline the section before drafting it. Analyses that answer no question are candidates for cutting or for an appendix.

The filter has one limit. JARS–Quant asks that the results of all inferential tests conducted be reported, and that planned analyses be distinguished from exploratory ones (Appelbaum et al., 2018, p. 8). Selecting what to report therefore means choosing which analyses to run and present as answers, rather than hiding tests that were run and gave unwelcome results. A test that was run and is relevant to a hypothesis is reported, whatever it showed.

The results also lay the foundation for the discussion. Anything the discussion will interpret must first appear in the results. If the discussion is to emphasize a strong relationship between two variables, the correlation or regression that shows it belongs in the results.

3.2 The usual order

Most quantitative results sections follow a similar sequence, because each element gives the reader what is needed to trust the next:

1. **An overview** of what the section reports and in what order, ideally tied to the research questions. A paragraph or two is enough.
2. **The sample**: how many people were approached, how many took part, and who they were.
3. **Checks on the measures and the data**: the reliability of any multi-item scales, missing data, outliers and the shape of the data.
4. **Descriptive statistics** for the main variables.
5. **Inferential statistics and hypothesis tests**, where the research questions call for them.
6. **A short summary** of the main findings, leading into the discussion.

© 2026 Grad Coach · CC BY 4.0 · gradcoach.com

Not every study needs every element. A descriptive study may end at the descriptive statistics, and not every quantitative study tests hypotheses. Programs also differ in where they place the first elements. JARS–Quant lists the sample's demographic characteristics and the reliability coefficients for the scores analyzed under the method section (Appelbaum et al., 2018, pp. 6–7), and many theses describe the sample in the methodology chapter for that reason. The order of the elements matters more than which chapter holds them.

4. Describing the Sample

The first data a reader needs are about the participants. For a survey, the description starts with the participant flow: how many people were invited, how many began, how many responses were usable and why the others were excluded. JARS–Quant asks for the percentage of the sample approached that actually participated and the number of participants at each stage of a study (Appelbaum et al., 2018, pp. 6–7). A master's thesis from Purdue University's School of Languages and Cultures, which surveyed students of Japanese at the university, reports its flow in a few sentences: about 400 students were invited, 124 started the survey, and 100 remained after incomplete responses and native speakers of Japanese were removed (Osumi, 2019, p. 38).

The description of the sample then covers the characteristics relevant to the study, typically age, gender and other demographics, and any topic-specific characteristics such as level of study or experience. Its main purpose is to show **how closely the sample resembles the population** about which the study aims to draw conclusions. The comparison requires figures for the population, so the survey needs to collect the characteristics that will be compared, and the population figures need a source, such as enrollment records. Figure 1 shows the comparison for an illustrative survey.

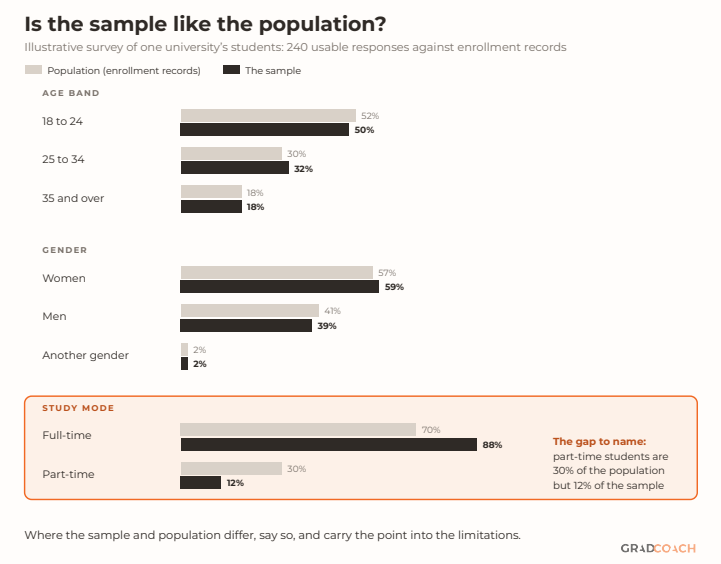

Figure 1. An illustrative sample compared with its population; study mode differs by 18 percentage points.

© 2026 Grad Coach · CC BY 4.0 · gradcoach.com

In this example, age and gender lie within two percentage points of the population, but part-time students make up 30% of the population and only 12% of the sample. A write-up names a gap of that kind plainly instead of leaving it out. The Purdue thesis does this in its methodology chapter:

> English constituted the largest group with 58 speakers, and Chinese (either Mandarin or Cantonese, or both) was the second-largest group with 26 participants. [...] Considering the fact that Chinese native speakers constituted a half of the students in Japanese first year-level as shown in Figure 4 and 5 in Chapter 2, it should be noted that this result might not represent the actual population of Japanese learners at the study site.
>
> Osumi (2019, p. 39)

The extract sets the sample beside the thesis's own figures for first-year classes and states the mismatch. The thesis then carries the point into its limitations, noting that the number of Chinese-speaking learners was too small given their share of the population (Osumi, 2019, p. 81). That is the pattern to follow: a group that is under- or over-represented is named in the results and reconsidered as a limitation in the discussion or conclusion.

A description of the sample matters even when a study does not aim to generalize. It gives the context in which the findings should be read. The Cornell experiment recruited its 280 participants through Amazon Mechanical Turk, an online platform, and 94.60% of them identified as white (Verdiguel, 2024, pp. 23–24). The thesis later notes that this sample may not fully represent the wider population of luxury consumers (Verdiguel, 2024, p. 37). A reader who knows who took part can judge how far the findings might apply beyond the sample.

5. Checking the Measures and the Data

Before any statistical results are reported, the section shows that the data are fit for the methods used. Most of this work concerns three things: the reliability of composite measures, missing data and outliers, and the shape of the data.

5.1 Reliability of composite measures

Many surveys measure a concept, or construct, with several items whose scores are combined into one composite score, such as four Likert-type items (statements that respondents rate on a scale, such as from strongly disagree to strongly agree) that should each capture job satisfaction in a slightly different way. If the items measure the same thing, respondents should answer them consistently. That property is internal consistency, and it is not guaranteed, particularly for items written for the study. JARS–Quant asks for internal consistency coefficients for composite scales, estimated on the study's own sample (Appelbaum et al., 2018, p. 7).

The usual coefficient is **Cronbach's alpha**, which typically ranges from 0 to 1 (Tavakol & Dennick, 2011, p. 53). Values reported as acceptable range from .70 to .95, and a very high alpha may indicate that some items are redundant (Tavakol & Dennick, 2011, p. 54). A low alpha can result from having few items, from items that correlate poorly or from a scale that measures more than one construct. If the items correlate poorly, some should be revised or discarded (Tavakol & Dennick, 2011, p. 54). Because alpha is affected by the number of items, it should be calculated for each construct rather than for a whole questionnaire (Tavakol & Dennick, 2011, p. 54).

© 2026 Grad Coach · CC BY 4.0 · gradcoach.com

A doctoral thesis in business from the University of Lincoln, which surveyed 440 online shoppers in Vietnam about collecting parcels from pickup points, reported its reliability check in four sentences:

> Cronbach's alpha was calculated for each construct to verify whether the scales met the recommended reliability threshold of 0.7 or higher, as suggested by Hair et al. (2010). To improve reliability, one item from the Subjective Experience (SBE) scale was removed, while the number of items in other constructs remained unchanged. The Cronbach's alpha values ranged from 0.755 (for SBE) to 0.933 (for CPL), indicating a high level of internal consistency across the constructs. This confirms that the measurement scales used in the questionnaire are statistically reliable and suitable for further analysis.
>
> Luong (2025, p. 81)

The extract names the threshold and its source, admits the item that was removed and why, and gives the range of values (CPL is the study's complexity scale). A table beside it lists every scale's items and alpha, so the reader can see each one. Its final sentence claims more than alpha can show, however. A high alpha indicates that the items are interrelated, but internal consistency is necessary and not sufficient for showing that the items measure a single construct, so alpha alone does not establish unidimensionality (Tavakol & Dennick, 2011, pp. 53–54). Nor does it establish validity: an instrument can be reliable without measuring what it is intended to measure (Tavakol & Dennick, 2011, p. 53). A more careful closing sentence would say that the scales showed acceptable internal consistency.

5.2 Missing data and outliers

JARS–Quant asks the results to report how much data are missing, what is known about why, and how the missing data were handled (Appelbaum et al., 2018, p. 8). The criteria for excluding participants and the definition and handling of outliers are planned in advance and described in the method (Appelbaum et al., 2018, p. 7). The results then report what those rules found: how many cases were flagged, and what was done with them.

The decision matters because a single value can change the descriptive statistics on its own. In the illustrative data in Figure 2, one value, 58, lies beyond the upper fence that a box plot would draw. That fence sits 1.5 interquartile ranges above the third quartile, the value below which three-quarters of the data fall; the interquartile range is the spread of the middle half of the data. Here the first and third quartiles are 21 and 33, so the fence is $33 + 1.5 \times 12 = 51$. Removing 58 moves the mean from 27.9 to 26.3 while the median moves only from 27 to 26, and the standard deviation falls from 10.5 to 7.9, by about a quarter. Different rules flag different cases. The z-score of 58, its distance from the mean in standard deviations, is about 2.9: inside the ±3.29 criterion used in the extract below. This is another reason to state the rule that was applied. Where an outlier decision could change a conclusion, reporting the key results both with and without it lets the reader judge.

© 2026 Grad Coach · CC BY 4.0 · gradcoach.com

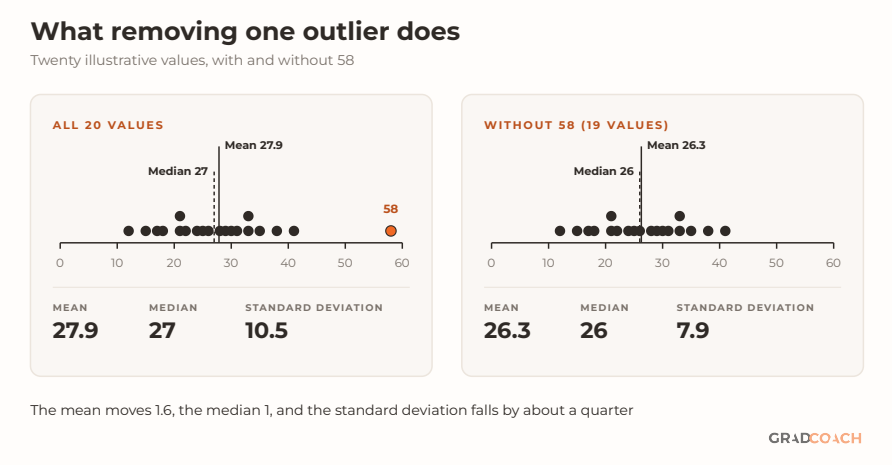

Figure 2. What removing one outlier does to the mean, median and standard deviation of 20 values.

An outlier is not an error by default. A doctoral thesis in clinical psychology from the University of Lincoln validated a measure of vicarious trauma, the lasting negative impact on helping professionals of cumulative exposure to clients' trauma (Tunstall, 2026), with 418 helping professionals. It followed the COSMIN Study Design Checklist (COSMIN-SDC: the COnsensus-based Standards for the selection of health Measurement Instruments Study Design Checklist) and reported its screening and its decision as follows:

> Univariate outliers were examined using standardized z-scores (±3.29 criterion; Tabachnick & Fidell, 2019). Some cases (N=23) exceeded the ±3.29 threshold but all values were below ±4 and fell within plausible participant variability ranges for psychological scale scores (Mokkink et al., 2019). No values represented data-entry errors or impossible totals as these had been removed during initial data cleansing. In line with COSMIN-SDC guidance to retain valid participant responses and avoid artificially constraining variability, these cases were not removed, and all data was retained for subsequent analyses.
>
> Tunstall (2026, pp. 63–64)

The passage gives the rule and its source, the number of cases it flagged, the reason they were judged genuine, and the decision with its justification.

5.3 The shape of the data

Many common inferential tests, including t-tests, analysis of variance and linear regression, are parametric tests: they rest on assumptions about how the data are distributed, of which approximate normality is the best known. A normal distribution is bell-shaped, with most values near the middle and progressively fewer toward both ends, in the same pattern on each side. For a t-test, the assumption applies to each group being compared (T. K. Kim, 2015). For a regression, it applies to the errors, which are checked through the residuals, the differences between observed and predicted values, using a histogram or a normal quantile-quantile (Q-Q) plot, which compares the residuals with the values a normal distribution would give (H.-Y. Kim, 2019). Once samples are large, the t-test and linear regression do not require normally distributed outcomes; this holds for almost any distribution of the outcome (Lumley et al., 2002). A modest departure from normality therefore matters less as the sample grows. When data are markedly skewed in a small sample, a non-parametric alternative may be the better choice.

© 2026 Grad Coach · CC BY 4.0 · gradcoach.com

Symmetry is part of normality but not the same thing: a flat spread of values is symmetric and far from normal. The check therefore starts with a plot of the data, read together with descriptive statistics such as the mean, median and skewness, a measure of how asymmetric a distribution is. Figure 3 shows the two shapes met most often: a symmetric distribution with one peak, and a right-skewed distribution whose peak sits low and whose tail runs toward the high values.

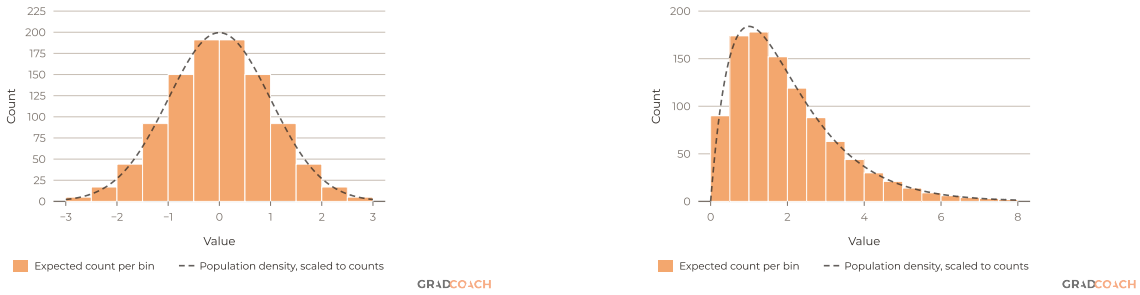

Figure 3. A symmetric histogram (left) and a right-skewed one (right). From Jansen (2026a, 2026b), CC BY 4.0.

The clinical psychology thesis in Section 5.2 reported its check in two short paragraphs. Its abbreviations name its questionnaires: the Trent Measure of Vicarious Trauma (TMVT), the Professional Quality of Life Scale–5 (ProQOL-5), the Oldenburg Burnout Inventory (OLBI), the Short Warwick–Edinburgh Mental Wellbeing Scale (SWEMWBS) and the Depression Anxiety Stress Scales–21 (DASS-21). CFA is confirmatory factor analysis, which tests whether a measure's items group together as intended. Kurtosis describes how heavy the tails of a distribution are compared with a normal distribution; Section 6 returns to both measures.

> Parametric correlations and CFA with large sample sizes (N>300) are considered robust to deviations of normality (Comrey & Lee, 1992; Fagerland & Sandvik, 2012; Kline, 2016). However, COSMIN-SDC and Boateng et al.'s (2018) guidelines both emphasise model-based evaluation over perfect distributional assumptions or reliance on standalone normality or significance testing (e.g. Kolmogorov-Smirnov & Shapiro-Wilk, Field, 2018).
>
> Consistent with guidance, continuous scales and subscales (TMVT, ProQOL-5, OLBI, SWEMWBS, DASS-21) were evaluated through visual inspection of histograms and boxplots, and Skewness/Kurtosis calculations. Values were all within acceptable visual and statistical ranges for approximate univariate normality (skewness between –1 and +1; kurtosis between –2 and +2) and therefore appropriate for parametric and CFA analyses.
>
> Tunstall (2026, p. 63)

The passage follows a clear sequence: why perfect normality is not required at this sample size, what was inspected, the cut-offs used, and the decision that followed. One element is missing. The extract gives no source for its skewness and kurtosis cut-offs, and guidance on such cut-offs varies between fields and authors, so a write-up should name the source of any threshold it applies.

6. Presenting the Descriptive Statistics

With the sample, the measures and the data described, the statistical results begin with the descriptive statistics for the main variables. JARS–Quant asks for the number of cases, the means, the standard deviations and other measures that characterize the data, for the total sample and

© 2026 Grad Coach · CC BY 4.0 · gradcoach.com

each subgroup (Appelbaum et al., 2018, p. 8). For a numerical variable, including a scale score, the usual measures are the following:

1. **Central tendency:** the mean (the arithmetic average) and, where the data are skewed or contain outliers, the median (the middle value when the values are sorted).
2. **Spread:** the standard deviation, a measure of how far the values typically lie from the mean, so that a small standard deviation indicates tightly grouped scores.
3. **Shape:** skewness, which describes whether the longer tail of the distribution runs toward the low or the high values, and kurtosis, which describes how heavy the tails of the distribution are compared with a normal distribution.

A single table that reports these measures for several variables at once is economical. Table 1 shows four of the eleven rows from the descriptive statistics table in the Purdue thesis, in which students rated their motivation on scales with a maximum score of 6. In the table, M is the mean, SD the standard deviation and N the number of respondents (lowercase *n*, as in Box 1, is the number in one group).

Table 1. Four rows of a table of descriptive statistics. Adapted from Osumi (2019, p. 56), CC BY 4.0.

Motivational factor	*M*	*SD*	*N*	Skewness	Kurtosis
Classroom support	5.25	0.73	97	−2.22	9.55
Attitudes to community	5.39	0.71	100	−2.20	8.06
Instrumentality-promotion	3.49	1.21	98	0.16	−0.71
Ought-to L2 self	2.78	1.02	98	0.14	−0.54

The table is worth reading closely. The two factors rated closest to the maximum of 6 also carry the most extreme skewness and kurtosis of all eleven. When most respondents choose the top of a scale, the scores cluster at one end and the tail runs toward the low values: the mirror image of the right-skewed histogram in Figure 3, a pattern often called a ceiling effect. Data of this shape are a warning against tests that assume normality. The thesis analyzed its relationships with Spearman's rank-order correlation, a non-parametric method that does not assume normality, although it does not give the shape of the data as the reason (Osumi, 2019, p. 48). It also ran one-way analyses of variance (Osumi, 2019, p. 48), which do assume normality. The varying values of N also show, row by row, how many respondents answered each set of items.

A table should not stand alone. The Purdue thesis follows its table with a sentence that says what it shows: "This table indicates that participants gave a high rating to most of the motivational factors" (Osumi, 2019, p. 56). The text states the main findings and the table holds the detail. APA Style asks that statistics not be repeated in both the text and a table (American Psychological Association, 2024), so the prose points to the table and picks out the few figures that matter, rather than copying every value.

For categorical data, such as gender or job role, the descriptive statistics are counts and percentages, presented in the text, in a table or in a bar chart. In a survey write-up these often make up much of the descriptive results. Every figure and table needs a clear title and labeled axes or columns, so that it can be understood without searching the text.

© 2026 Grad Coach · CC BY 4.0 · gradcoach.com

Some research questions are answered fully by descriptive statistics, for example a question about what proportion of a population holds a view or behaves in a certain way. For such a study, the statistical results may end here. Where the research questions ask whether groups differ or variables are related in a wider population, inferential statistics follow.

7. Reporting Inferential Statistics and Hypothesis Tests

7.1 What to report for each test

Inferential statistics use the sample to draw conclusions about a population, whereas descriptive statistics describe only the sample. Broadly, inferential tests either compare groups, as t-tests compare two means and analysis of variance compares three or more, or assess relationships between variables, as correlation and regression do. Within each family, parametric tests carry the assumptions discussed in Section 5.3, and non-parametric tests are available for data that do not meet them. The tests reported are those that answer the research questions, and each is reported in full. A results section with three well-reported tests is easier to examine than one with fifteen partly explained ones.

JARS–Quant sets out what "in full" means. For each inferential test it asks for the results, including the exact *p*-value where significance testing is used and the statistics needed to reconstruct the test, such as the degrees of freedom; and it asks for effect-size estimates and confidence intervals corresponding to each test, when possible (Appelbaum et al., 2018, p. 8). It also asks for a clear distinction between primary, secondary and exploratory hypotheses (Appelbaum et al., 2018, p. 8). An effect size is a measure of how large a difference or relationship is. Reporting it matters because statistical significance does not by itself show that size.

7.2 The logic of a hypothesis test

In practice, hypothesis tests are not a separate set of analyses: they are the inferential tests described in Section 7.1. What differs is the write-up, which takes one hypothesis at a time and gives each a verdict. A hypothesis is a prediction that a test can support or fail to support, usually of a difference between groups or a relationship between variables. A significance test of a hypothesis runs in five steps:

1. **State the null hypothesis**, usually that there is no difference or relationship in the population.
2. **State the alternative hypothesis**, the difference or relationship the study predicts.
3. **Set the significance level**, often called alpha (not to be confused with Cronbach's alpha) and conventionally .05, before the data are analyzed. The significance level is part of the study design and is not changed in light of the data, whereas the *p*-value is a result computed from the data (Greenland et al., 2016, pp. 339–340).
4. **Run the test and obtain the *p*-value**: the probability that the test statistic would be at least as extreme as the one observed if the null hypothesis and every other assumption of the test were correct (Greenland et al., 2016, p. 339).

© 2026 Grad Coach · CC BY 4.0 · gradcoach.com

5. **Compare the *p*-value with the significance level.** If it is at or below the level, the null hypothesis is rejected and the result is consistent with the alternative; if it is above, the null hypothesis is not rejected.

The *p*-value is calculated on the assumption that the null hypothesis is true, so it is not the probability that the null hypothesis is true (Greenland et al., 2016, p. 340). Figure 4 works through the five steps with the first hypothesis of the Cornell experiment, H1a, which predicted that a store with phygital elements would produce greater luxury purchase intention than a traditional store (Verdiguel, 2024, p. 10).

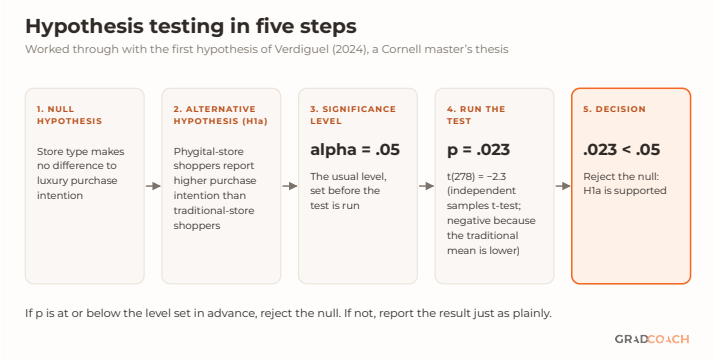

Figure 4. The five steps of a significance test, worked through with H1a of Verdiguel (2024).

The thesis reports the test as follows:

> Three independent t-tests were conducted to evaluate the effect of phygital design elements in luxury retail environments on luxury purchase intention (H1a), as well as the perceived usefulness (H1b) and playfulness (H1c) of the store. Participants in the phygital store condition (M = 5.83, SD = 0.82) reported significantly higher luxury purchase intention than those in the traditional store condition (M = 5.56, SD = 1.11; t(278) = -2.3, p = .023). Thus, H1a was supported.
>
> Verdiguel (2024, p. 30)

The extract names the test and the hypothesis, gives both groups' means and standard deviations, then the test statistic, t, with its degrees of freedom in parentheses, t(278), and the exact *p*-value, and ends on a one-line verdict. The negative sign reflects only the order in which the two groups were compared; the traditional store had the lower mean. Two elements that JARS–Quant asks for are absent: an effect size and a confidence interval. Cohen's *d* expresses the difference between two means in standard deviation units. When only the total sample size is known, *d* for two independent groups can be approximated as $2t/\sqrt{N}$ (Lakens, 2013, p. 3). For this test, with 280 participants, that gives $d \approx 0.27$, and the means and standard deviations give a similar value, about 0.28, assuming two groups of equal size. Cohen's conventional benchmarks would call this a small effect, although the benchmarks are arbitrary and should not be interpreted rigidly (Thompson, 2007, as cited in Lakens, 2013, p. 3). The *p*-value shows that the difference is statistically significant; the effect size reported beside it shows that the difference is modest: just over a quarter of a standard deviation. In raw terms, the two means differ by 0.27 points on a 7-point scale.

The extract also follows the thesis's own conventions. APA Style would report the t statistic to two decimals (American Psychological Association, 2024), and JARS–Quant lists the statement of support or nonsupport for each hypothesis under the discussion rather than the results (Appelbaum

© 2026 Grad Coach · CC BY 4.0 · gradcoach.com

et al., 2018, p. 8). Many theses give the verdict in the results, as this one does, and the program's guidance decides which is expected.

7.3 Results that are not significant

A hypothesis that is not supported is reported as plainly as one that is, with the same statistics. The wording needs care. A *p*-value above .05 does not show that there is no effect or no difference. It means only that the null hypothesis is one of many hypotheses that would give a *p*-value above .05; unless the estimate exactly equals the null value, some difference or association is present in the sample data (Greenland et al., 2016, p. 341). The appropriate phrase is "not statistically significant", not "no effect".

A confidence interval shows which effects remain compatible with the data when a result is not significant. If 95% confidence intervals are calculated repeatedly in valid applications, about 95% of them will contain the true effect size; that coverage is a property of the procedure across many studies, not of any single interval (Greenland et al., 2016, p. 340). The values inside an interval are those more compatible with the data, under the statistical model, than the values outside it (Greenland et al., 2016, p. 340). The H2a result in Section 2 illustrates the point: its interval for the indirect effect ran from −0.01 to 0.16, which includes zero but also includes positive effects.

Box 1 brings these elements together in an illustrative write-up of two hypotheses, one supported and one not.

Box 1. An illustrative write-up of two hypothesis tests, written for this article; the data are invented.

> H1 predicted that students who attended a study-skills workshop would score higher on the end-of-term examination than students in a control group. An independent-samples *t*-test showed that workshop students ($M = 68.4$, $SD = 10.2$, $n = 60$) scored higher than control students ($M = 63.1$, $SD = 11.0$, $n = 60$), $t(118) = 2.74$, $p = .007$, $d = 0.50$. The mean difference was 5.3 points, 95% CI [1.5, 9.1]. H1 was supported.
>
> H2 predicted that the difference would remain at a follow-up test six weeks later. Workshop students ($M = 65.2$, $SD = 10.5$) and control students ($M = 63.6$, $SD = 10.3$) did not differ significantly, $t(118) = 0.84$, $p = .40$, $d = 0.15$. The mean difference was 1.6 points, 95% CI [−2.2, 5.4]. H2 was not supported.

Each paragraph of Box 1 states the hypothesis, names the test, gives the descriptive statistics for both groups, then the test statistic, the exact *p*-value, the effect size and the confidence interval, and ends with the verdict. The second result does not show that the workshop's effect had disappeared. Its interval runs from a difference of 2.2 points in favor of the control group to 5.4 points in favor of the workshop, so the data are compatible with a small disadvantage, no difference or a moderate benefit. Saying what that range means for the study is a task for the discussion.

7.4 Presenting a tested model

Where a study develops and tests a conceptual framework, the results usually end with the tested model: the framework drawn again, with each path's result marked, so that the reader can see at a glance which relationships the data supported. What the pattern of results means is again left to the discussion.

© 2026 Grad Coach · CC BY 4.0 · gradcoach.com

8. Writing and Presenting the Section

A few conventions make a results section easier to read and examine:

1. **Write in the past tense** when reporting what was found ("Participants in the experimental group scored higher"). A sentence that points to a table often reads more naturally in the present tense ("Table 2 shows").
2. **Follow the style guide's rules for statistics.** In APA Style, statistical symbols such as M, SD, t and *p* are italicized and need no definition. The symbol is used with a mathematical operator (M = 5.8), and the word is used in running text ("the means were"). A statistic that cannot exceed 1, such as a correlation or a *p*-value, has no zero before the decimal point. Means and standard deviations for data on integer scales are reported to one decimal, and inferential statistics to two. Exact *p*-values are given to two or three decimals, with values below .001 written as $p < .001$ (American Psychological Association, 2024).
3. **Build tables and figures for the reader** rather than pasting software output, which is cluttered and rarely follows an academic program's formatting rules.
4. **State the findings in the text.** Tables and figures support the writing; the important values and patterns are stated in the prose.
5. **Order the material logically.** If one analysis or experiment informed the next, they are reported in that order.
6. **Close with a short summary** of the main findings (a paragraph or two) that leads into the discussion without beginning it.

A final check against the research questions completes the draft. Each question should be answerable from the results as written, and anything that answers none of them is a candidate for cutting or for an appendix. Universities and departments may follow a style other than APA, so the relevant handbook is the final authority on the details.

9. Conclusion

A quantitative results section reports what the data show and leaves their interpretation to the discussion. What it includes is decided by the research questions, and its order follows what a reader needs to trust each result: the flow of participants and the description of the sample, the reliability of the measures, the handling of missing data and outliers and the shape of the data, the descriptive statistics, and then the inferential tests and hypotheses. Each test is reported in full, with its test statistic, degrees of freedom and exact *p*-value, an effect size and a confidence interval, and every hypothesis, supported or not, is reported with the same care. A *p*-value is read as a statement about the data under the null hypothesis and the test's assumptions, and a non-significant result is never reported as evidence of no effect. Written in this way, the results section gives the discussion a complete and checkable foundation.

© 2026 Grad Coach · CC BY 4.0 · gradcoach.com

References

American Psychological Association. (2024). *APA Style numbers and statistics guide*. https://apastyle.apa.org/instructional-aids/numbers-statistics-guide.pdf

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board task force report. *American Psychologist*, 73(1), 3–25. https://doi.org/10.1037/amp0000191

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, P values, confidence intervals, and power: A guide to misinterpretations. *European Journal of Epidemiology*, 31(4), 337–350. https://doi.org/10.1007/s10654-016-0149-3

Jansen, D. (2026a). *Histogram shape - skewed right* [Figure]. Grad Coach. https://gradcoach.com/open-educational-resources/right-skewed-histogram-example/

Jansen, D. (2026b). *Histogram shape - symmetric, unimodal* [Figure]. Grad Coach. https://gradcoach.com/open-educational-resources/symmetric-histogram-example/

Kim, H.-Y. (2019). Statistical notes for clinical researchers: Simple linear regression 3 – residual analysis. *Restorative Dentistry & Endodontics*, 44(1), Article e11. https://doi.org/10.5395/rde.2019.44.e11

Kim, T. K. (2015). T test as a parametric statistic. *Korean Journal of Anesthesiology*, 68(6), 540–546. https://doi.org/10.4097/kjae.2015.68.6.540

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, 4, Article 863. https://doi.org/10.3389/fpsyg.2013.00863

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, 151–169. https://doi.org/10.1146/annurev.publhealth.23.100901.140546

Luong, V. D. (2025). *Factors influencing e-commerce consumer's adoption intention towards collection-and-delivery points (CDPs): The case of Vietnam* [Doctoral thesis, University of Lincoln]. University of Lincoln Research Repository. https://doi.org/10.24385/lincoln.30655538

Osumi, A. (2019). *Exploring motivation for learning Japanese as a foreign language* [Master's thesis, Purdue University]. Purdue University Graduate School. https://doi.org/10.25394/PGS.8023421.v1

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, 2, 53–55. https://doi.org/10.5116/ijme.4dfb.8dfd

Tunstall, C. M. (2026). *Confirmatory factor analysis and further validation of the Trent Measure of Vicarious Trauma (TMVT)* [Doctoral thesis, University of Lincoln]. University of Lincoln Research Repository. https://doi.org/10.24385/lincoln.33113120

Verdiguel, N. R. (2024). *Phygital design: Boosting luxury purchase intention by enhancing consumer perceptions of in-store usefulness and playfulness* [Master's thesis, Cornell University]. eCommons. https://doi.org/10.7298/egd2-5q14

© 2026 Grad Coach · CC BY 4.0 · gradcoach.com