Misuse, Misreporting, Misinterpretation of Statistical Methods in Usable Privacy and Security Papers

Jenny Tang
Carnegie Mellon University

Lujo Bauer
Carnegie Mellon University

Nicolas Christin
Carnegie Mellon University

Abstract

Null hypothesis significance testing (NHST) is commonly used in quantitative usable privacy and security studies. Many papers use results from statistical tests to assert whether effects or differences exist depending on the resulting p-value. We conduct a systematic review of papers published in 10 editions of the Symposium on Usable Privacy and Security over a span of 20 years to evaluate the field's use of NHST. We code statistical tests for potential statistical validity, reporting, or interpretation issues that may undermine assertions made in the 121 papers that use NHST. Most problematically, tests in 23% of papers inadequately account for non-independence between samples, leading to potentially invalid claims. 58% of papers lack information to verify whether an assertion is supported, such as imprecisely specifying the statistical test conducted. Many papers contain more minor statistical issues or report statistics in ways that deviate from best practice. We conclude with recommendations for statistical reporting and statistical thinking in the field.

1 Introduction

Statistical methods are often used in human-computer interaction research to support assertions about the presence (or absence) of an effect of scientific significance (e.g., some magnitude of difference) accompanied by a measure of statistic significance. Indeed, one of the most common refrains in statistical analysis is that a result is significant because the "p-value" is less than a given threshold, e.g., $p < 0.05$. Despite over half a century of criticism, null hypothesis significance testing (NHST, also known as statistical significance testing)—that is, methods using p-values from inferential statistical tests as evidence to reject a null hypothesis—remains the dominant form of statistical analysis and evaluation [17]. However, simply dichotimizing results into "significant" and "non-significant" through their associated p-values without reporting other information is not in itself sufficient to convey the scientific importance of the claims, nor the richness and complexity of data collected from human subjects. This reliance on p-values to support assertions sometimes leads other information vital to understanding statistical and scientific significance to be omitted.

As a result, complete reliance on p-values is increasingly frowned upon, with some journals banning the reporting of p-values altogether [75, 81]. Most other current guidance is less drastic, and recommends using statistical hypothesis testing as a starting point and providing sufficient context (such as effect sizes, confidence intervals, and underlying data) to convey the scientific significance of the claims [2,13,49,59,80,81]. We use this guidance to evaluate whether the scientific assertions made on the basis of NHST in usable privacy and security (UPS) are accompanied by sufficient reporting for readers to validate whether these assertions are supported by the information present in the paper. We focus on UPS as it is still a fairly young area, with evolving standards, features a considerable amount of quantitative research, and errors or misinterpretations can be detrimental to user safety in the digital world and beyond.

Prior work has also examined the transparency, reporting, and validity of statistical methods in HCI and various subfields [16,25,36,51,62,66,77]. However, the evaluations in these works typically focus on evaluating whether p-values are accurately computed or on whether there may be false negatives (such as due to lack of power) or false positives (such as from inaccurately reported p-values).

In this work, we look beyond statistical significance to examine statistical validity (whether the chosen test is suitable for the data or whether it may produce spurious results), reporting transparency and completeness (whether the reported

USENIX Symposium on Usable Privacy and Security (SOUPS) 2025.
August 10–12, 2025, Seattle, WA, United States.

information is sufficient for an informed reader to evaluate the merit of the assertion), and the accuracy of the interpretation of the result of the statistical test. More specifically, we conduct a systematic review of the use and reporting of statistical tests in all papers published in 10 editions of the Symposium on User Privacy and Security (SOUPS) over a span of 20 years. We examine methods and data in conjunction with statistical procedures, and we evaluate the mapping of results of statistical procedures to assertions, to identify statistical validity issues (i.e., issues with the choice of statistical test) and interpretation issues. We draw upon guidelines from the American Psychological Association (APA) and American Statistical Association (ASA), and consult criteria from prior systematic reviews, to identify reporting issues [2, 25, 62, 66, 77, 80, 81].

We code 479 assertions based on NHST across 121 papers. Because sufficient data to re-run statistical tests is typically not available, and because attempting to re-run statistical tests would significantly limit the scope of our study, we assume that all tests are computed correctly and only evaluate them with respect to the reported information and the assertions they are used to support.

Most problematically, we find that 9% of all assertions, spanning 23% of papers, do not properly account for dependence or independence in the data, leading to potentially invalid claims. We further find that the information reported cannot support 21% of the assertions made based on NHST, across 47% of papers, as the results are not fully substantiated by the reported statistics. For another 23% of assertions, across 58% of papers, information is lacking for readers to judge whether the statistical methods support the assertion. Overall, we identify at least one statistical validity, reporting, or interpretation issue in almost all (∼97%) papers we surveyed. While many reporting issues, such as not reporting the test statistic, are less severe, others, including not reporting the statistical test, affect the reader's ability to evaluate statistical validity and undermine support for the resulting assertions.

Drawing on these findings, we provide recommendations and suggest best practices for statistical reporting and thinking. These recommendations set out ways for authors, reviewers, and readers to move beyond pure reliance on p-values and instead engage with the context necessary to critically evaluate assertions and better reflect the complexity of the people who are at the center of UPS research.

2 Background and Related Work

We give a brief overview of the debates surrounding the scientific use of NHST broadly (§2.1,§2.2), as well as in HCI and UPS (§2.3). We also summarize existing guidance from various fields on best practices for statistical result reporting and research transparency relating to NHST (§2.4).

2.1 Challenges with the Concept of Statistical Significance

The widespread use of NHST and the over-reliance on an "all or nothing" approach with p-value thresholds has spurred much discussion and debate since the 1950s [32, 39, 59, 59, 64, 69–71, 81, 86]. One concern is that this focus on the p-value can lead to inadequate reporting of other relevant information, such as measures of uncertainty or effect sizes when results do not fall under such a threshold [39, 81]. In one case, the journal *Basic and Applied Social Psychology* banned (with mixed results [31]) the use of p-values to spur better and more complete reporting and evaluation of results [75]. Even the concept of NHST itself is controversial, often seen as an incomplete and self-contradictory amalgamation of two distinct schools of thought regarding statistical testing, Fisher and Neyman-Pearson [64, 69].

Despite its shortcomings, NHST remains a useful tool [43, 56], and remains widely used, as p-values are an easily comprehensible (if not always accurate) shorthand to indicate interesting and significant results. The key is thus to provide context beyond just a p-value, such as effect sizes, to provide greater support for the claims being made [2, 79]. Indeed, p-values and statistical models are not by themselves objective measures of truth. For instance, random variation and varying violations of statistical assumptions can lead to different results and p-values in experiment replications [6]. Changes in sample size or statistical test can inflate or deflate p-values [29, 45]. Therefore, adequate reporting of the whole process of statistical selection is critical to allow for validation of scientific accuracy. Unfortunately, investigations of statistical reporting in preclinical research found that inadequate reporting for statistical tests occurs in a large proportion of articles, and there have been concerns that the majority of claims from statistical tests are false across many fields [33, 44]. Questions of transparency, reporting, and replicability have also arisen in the field of Human-Computer Interaction and its subfields [16, 49, 85].

2.2 Other General Statistical Issues

Among the most prominent statistical issues papers have focused on are questions regarding statistical power, corrections for multiple comparisons, and reporting transparency.

Sufficient knowledge regarding the data is necessary before any evaluation can be made on the validity of the statistical test itself. As such, questions about transparency of statistical reporting are at the forefront of much investigation regarding statistical validity [16, 66, 77]. Therefore, we focus on evaluating the adequacy and transparency of statistical reporting in our work as it is the basis upon which other evaluations rest.

Losses of power, such as from small sample sizes or using a two-sample test for paired data, can lead to false negatives—effects that are present not being detected. Conversely, paired

tests are sometimes used for independent data, leading to arbitrary paired comparisons and potential false positives. The multiple comparison problem and concerns about p-hacking (running large numbers of statistical comparisons and only reporting positive results, which may occur by chance) also affect statistical validity. In one striking example, Bennett et al. demonstrated that without corrections for multiple comparisons, results from an fMRI showed statistically significant results supporting that dead salmon could recognize emotions from human facial expressions [11].

However, issues with lack of power or multiple comparisons rest on the recognition of a specific p-value threshold [8]. Hence, while important to consider, we do not explicitly examine these concerns, but rather focus on the reporting surrounding statistical tests. Lack of power resulting in false negatives may still indicate promising directions for further study. False positives may be less impactful when considered within clearly reported scientific context (such as effect size), or when sufficient information is provided for replications of the study.

2.3 Quantitative Methods in HCI and UPS

Concerns about the validity of statistical methods in HCI are not new. In 2007, a meta-analysis of 41 papers in HCI found that only one paper did not contain any problems that undermined the validity of the results [16]. Similarly, comparisons of work in CHI between 2017 and 2022 found that there was little change in transparent reporting practices for quantitative methods [66]. While nearly all papers in both years reported p-values, only around half of papers in 2022 and 37% in 2017 reported effect sizes, and fewer than 20% in either year reported confidence intervals [66]. An analysis of award-winning papers in CHI 2020 found that while the majority reported some form of variability, only 16% reported any kind of effect size [51]. Another 2020 study found that 63% of the papers using NHST published in CHI PLAY between 2014 and 2019 reported effect sizes, yet only 7% reported confidence intervals. These results support previous claims that "many studies rely solely on p-values for their inferences" [77].

Similar trends exist in the usable privacy and security subfield. Prior work examining the use and reporting of statistical methods in UPS revealed that reporting is often incomplete and relies on claims of significant p-values [25]. Groß found that over half of cyber security user studies did not report sufficient information to validate reported p-values [36]. Of those with sufficient information (from the test statistic and degrees of freedom), the recalculated p-value fell on the other side of the $p = 0.05$ threshold in 14% of papers. Other works have uncovered that tests in many UPS user studies are underpowered for detecting small, medium, or sometimes even large effect sizes [35, 62]. This has resulted in concerns that most positive findings in UPS are false [37], echoing worries first raised by Ioannidis in 2005 when examining statistical use across general scientific fields [44].

While these concerns are relevant for many scientific fields, we focus specifically on UPS. We extend prior work by explicitly including regression models in our analyses, which prior work has rarely evaluated. Much prior work also excluded tests where insufficient information was provided or the provided information was unclear, whereas we attempt to evaluate the amount of information that a reader can glean *even when* statistical reporting is incomplete.

2.4 Existing Recommendations and Guidelines

Given the concerns regarding statistical use, various guidelines and checklists exist [3, 5, 24, 34, 42, 57, 59, 63]. Some guidance focuses on specific procedures, such as running corrections for multiple comparisons or sample size selection [12, 15]. Other guidance focuses on reporting. Within HCI, a movement has advocated for transparent reporting and openness, with guidance on "transparent statistics" [4, 22, 41, 49, 66, 77, 78]. Many, including the American Statistical Association (ASA), have suggested moving away from reliance on the $p < 0.05$ threshold [81]. These calls have been echoed in HCI [28, 48, 50]. Other standards are more field- or venue-specific [1, 5, 9, 20, 80]. For example, sharing data is not always feasible due to user data protections or space constraints [52, 66, 78]. Of the many sets of standards, those of the American Psychological Association (APA) are among the most widely known and used, and have been recommended as reporting guidelines both within UPS and HCI [2, 24, 25, 62, 66]. We draw upon these guidelines for our evaluation of statistical use and reporting, and investigate how often reporting supports the assertions based on NHST.

3 Methods

In this systematic review, we examine potential issues in the use of statistical methods in the field of usable privacy and security (UPS) through an evaluation of a sample of papers published at the Symposium on Usable Privacy and Security (SOUPS) between 2005 (the conference's first year) to 2024. We chose the SOUPS conference as it is the largest and longest-running conference whose focus is solely on usable privacy and security. We chose to focus on this research area, as 1) it features a considerable amount of quantitative research, and thus lends itself well to the kind of analysis we propose, 2) the research area is still fairly young, but community standards have greatly evolved over the past 20 years, which warrants scientific exploration, and 3) erroneous assertions or misinterpreted claims can have a direct detrimental impact on users' safety, online and otherwise.

3.1 Sample Selection and Codebook Development

We examine a sample consisting of 121 papers that 1) were published in even-numbered years of the SOUPS conference (10 editions in total) and 2) contain some form of null hypothesis statistical testing (NHST).

We start with the complete set of 223 papers published in even-numbered years and analyze those that use p-values from inferential statistical tests to determine the statistical significance of a result.[1] One author examined all 223 papers in the selected ten editions to determine whether NHST was used in each of them. If there was at least one assertion supported by NHST anywhere in the body of the paper (including tables and figures), the paper was included. This resulted in the 121 papers, a list of which can be found in [72]. We chose to limit ourselves to papers published in even-numbered years, as this produces a stratified sample of statistical use throughout 20 years without being skewed towards earlier or later editions of the conference.

To assemble our dataset, we analyze every use of NHST in each paper. We call each use (or several related uses, as described next) of a test an *instance*. An instance represents either the use and reporting of a single test or several uses of the same test that are all identically reported. Each instance can have more than one issue (of the same type or of different types). For example, suppose a paper reports on the effect sizes and exact p-values of two two-sample t-tests with the variables gender (male vs. non-male) and region (urban vs. rural); and that it does not provide any information on the effect size or p-value of a third two-sample t-test comparing participants with technical background with those without technical background. We would categorize the two tests on gender and region as one instance (because the same test is used and reported identically) and the test on technical background as another instance.

We use both inductive and deductive coding to analyze the use of statistical tests. Initial codes were inductively developed by one author with statistical knowledge coding all 22 papers containing NHST from two years of the SOUPS conference (2014 and 2019) and expanded through referencing the APA guidelines on Reporting Standards for Quantitative Research [2, 7]. The initial set of codes was then refined through discussion among the research team.

The same author (the primary coder) then used this set of codes to code all other papers. For each instance, the primary coder wrote down the details and reasoning for the assigned code, as well as noting edge cases and ambiguities. New codes were iteratively added when previously unencountered issues arose. These cases were resolved through discussion among the research team. The codebook categories were further refined through referencing systematic reviews in HCI and statistical works [16, 24, 49, 62, 66, 77]. After all papers were coded, the primary coder revisited all papers in the sample with the final codebook to ensure consistency across the dataset. The final codebook containing the codes, definitions for each code, and examples can be found in Appendix A. The full list of examples and criteria for the most common tests and models can be found in [73].

For cross-validation, a second coder with a background in statistics coded a subsample (stratified by year) of 19 papers from our full sample of 121 papers. The primary coder trained the second coder by going through and coding two example papers not in the subsample and discussed the rationale for the coding of each instance. The second coder then coded some of the subsample, optionally indicating uncertainty about how to code specific instances. Partway through the sample, the primary and secondary coder met to answer questions and perform an initial resolution of codes. When the full subsample was coded by the second coder, the coders met to discuss codes that were marked by the second coder as unsure. A random subsample of instances was further selected to discuss. Our initial agreement level before reconciliation was 0.51. The two coders discussed each code that differed and came to an agreement on the final code. In the majority of cases (85%), the code was resolved to primary coder's code. In *all* remaining cases, the code was resolved to a code that indicated more severe issues than the primary coder's initial code. This suggests that the codes on which we base our analysis are a conservative lower bound of the issues that arise in statistical methods in UPS papers.

3.2 Code Categories

We next discuss the codes for potential issues with the use, reporting, and interpretation of all inferential statistical tests in the papers that comprise our sample. For brevity, we will use "*tests*" to refer to statistical tests, models, procedures, and techniques unless there is explicit indication otherwise. We divide codes into three main categories: statistical validity issues, reporting issues, and interpretation issues.

3.2.1 Statistical Validity Issues

The first category of codes evaluates whether the choice of statistical test is suitable for the data and would result in valid results. We classify these issues into three main types.

Incorrect. Independence or dependence assumptions for the chosen statistical test are not properly met. While many statistical tests are robust to violations of their assumptions (such as those of normality) at sufficient sample sizes, the independence assumption remains critical [10, 46, 54, 65, 68]. For example, if independent data is arbitrarily paired, the same data can be found to differ in opposite directions (positive *or*

[1] Though not traditionally included in NHST, we include regression models within our systematic review for use of NHST as the significance level of a coefficient can be interpreted as testing whether the coefficient is significantly different from 0.

negative change). Examples of tests we coded as **incorrect** include using a paired t-test on independent samples or not accounting for repeated measures. In the best case, improperly accounting for the independence or dependency between datapoints reduces the power of the test; in the worst, these issues lead to erroneous or invalid results [38, 84], e.g., false positives, false negatives or incorrect correlation direction.

Data Type Mismatch. This code denotes that the statistical test is not suited to the data type. Tests are often structured for specific types of data and perform better when those distribution assumptions are satisfied [16, 55, 58, 61]. For example, we include treating binary data or ordinal data as continuous when used as dependent variables, as well as aggregating multiple Likert scale variables into one independent variable without validation (as this type of aggregation removes any information about variance), as a **data type mismatch**.[2] These issues can lead to a loss of power, leading to false negatives, or testing of a different hypothesis than intended (e.g., difference in means rather in distributions). A description of the types of data considered to be a mismatch for each test can be found in [73].[3]

Unverifiable. There is insufficient information reported about the statistical test or model to determine whether it is suitable for the data. As an example of the **unverifiable** code, some papers simply report p-values without any indication of what statistical method was used to produce these values, report a non-unique name, or a non-unique test statistic [5]. These may be statistical tests that are suitable or not for the data, but without further information, it is impossible to verify.

Each instance is coded with only one statistical-validity-issue code, as the codes are mutually exclusive. Codes in this category provide insight into the whether the statistical test measures what it is intended to investigate in the context of the data, and thus, whether the assertions made on the basis of the instance are supported by the statistical test results.

3.2.2 Reporting Issues

The second category of codes investigates whether sufficient context and evidence are reported to evaluate the *results* from the statistical test. In contrast to the issues related to the *choice of statistical test*, described in Section 3.2.1, the issues we code for in this category do not call into question the choice of test, but instead relate to a reader's ability to verify whether there is evidence that supports the assertion both in terms of statistical significance (e.g., p-value and its context) and scientific significance (e.g., effect magnitude). The APA Publication Manual states that "Because each analytic technique depends on different aspects of the data and assumptions, it is impossible to specify what constitutes a 'sufficient set of statistics' in general terms" when reporting statistical results [2]. The definitions we used of the lower bound for what is acceptable for each test can be found in [73]. Issues with statistical reporting fall into three main categories, described next.

Insufficient Reporting to Evaluate Statistical Significance. Any one of "test statistic, the degrees of freedom, [... and the ...] exact p-value" is not reported when expected [2]. Missing one of these values is generally not severe, as the information provided by these values can sometimes be inferred or calculated from other information. Reporting of these pieces of information aids with contextualization and verification of the results. For example, reporting of degrees of freedom not only allows for (re-)calculation of p-values, but also can allow readers to infer the number of groups, and whether data may have been transformed. Furthermore, large sample sizes can deflate p-values, and degrees of freedom allows a better understanding of what groups are compared and contextualizes significance levels [19, 76].

Insufficient Reporting to Evaluate Scientific Significance. There is insufficient information to understand the "size and direction of the effect" [2]. Without reporting of magnitude or effect size, there is little evidence to support the scientific significance of the assertion. Thus, evaluations of the assertion relies solely on the presence of statistical significance and p-value cutoffs. The following are the types of scientific significance measures we code for.

Effect Size Measures. This includes unitless, standardized effect sizes such as Cohen's d or Cramer's V [23] or effect sizes associated with specific tests and models—such as coefficients or estimates (β) for regressions, or R^2 for model fit. Any measure of effect size constitutes sufficient context to evaluate scientific significance.

Other Context (Descriptive Data). Without effect sizes, information that characterizes the data—whether in numerical (descriptive statistics) or visual format—can also provide sufficient context. These include providing the frequencies in each category (for categorical or ordinal data), aggregated descriptive statistics such as measures of centrality (e.g., means and medians) or distributions (e.g., stacked bar charts or boxplots). The descriptive context must be structured in such a way as to provide adequate information about the groups being compared or evaluated, particularly if visual (see Appendix A). For example, if a statistical test is examining differences between two groups, information about the two groups should be reported separately rather than aggregated into one.

Measure of Variability. If context is reported as point estimates, we further look for measures of variability to understand the confidence in and precision of the reported number [26].[4] For example, means should have context such as

[2] We recognize there remains ongoing debate about the suitability of treating ordinal data as interval data and whether such choices reduce the power of such tests [18, 27, 47, 53]. It is out of scope for this work to settle the debate on the ideal analysis of Likert-type data.

[3] These data types are ones encountered in the coding and not exhaustive. They should not be taken as recommendations for how to analyze each type of data.

[4] This is also commonly termed "uncertainty".

standard deviations to illustrate the distribution of the data. For regression coefficients, we expect confidence intervals (or standard errors) to allow for an evaluation of the precision of the model's estimates.

Insufficient Reporting to Evaluate Validity. Reporting is insufficient to determine the validity of the results of the statistical test. This may be due to not reporting the test used or not providing sufficient information to uniquely identify a test. For instance, a "Wilcoxon test" can indicate either the Mann-Whitney-Wilcoxon test for independent samples or the Wilcoxon Signed-Rank Test for paired data. This code also identifies imprecise model specification, such as not reporting which predictors are used. This code also occurs when reporting regarding the data evaluated by the statistical test is insufficient, such as not explaining what groups are being compared, or failing to explain how values are transformed (e.g., changing a Likert scale variable into a binary variable).

A instance may be coded with more than one reporting issue. For example, the reporting may not be sufficient for evaluating either statistical *or* scientific significance.

3.2.3 Interpretation Issues

This category of codes describes whether the results of statistical tests are consistent with and support the assertions papers make. Sometimes the statistical results (e.g., p-values and effect sizes) are presented as supporting some assertion that is in fact not supported (or not supported in the way described). These we code as interpretation issues, of which there are three major types.

Misinterpretation. Misinterpretations state something that is not directly supported by the results of the test.

Incorrect Interpretation. The interpretation is directly contradicted by the results. For example, results are interpreted on the wrong scale, such as interpreting changes in log odds as changes in odds, which *reverses* the effect for coefficients between 0 and 1.

Improper Statistical Significance Interpretation. The interpretation conflates statistical significance with effect size. This includes interpreting the presence of small p-values to indicate large magnitudes of difference or lack of statistical significance as lack of effect.

Misrepresentation. This indicates that model coefficients are not interpreted adequately in the context of the model. For example, interpretations do not explain results with respect to a baseline categorical variable when necessary or interprets interactions on their own, rather than as additions to main effects. This code also arises when a metric is mentioned that does not match the other context in the paper, such as reporting a correlation for a comparison test.

Not Tested. The interpretation makes claims beyond what is investigated by the statistical test. One example is reporting statistically significant differences between two groups in omnibus tests without pairwise comparisons.

Sub-optimal Effect Interpretation. We use this code when the general direction of the interpretation is correct, but reports results on a sub-optimal scale (or no scale), or describes the test result or its implications in terms that do not fully match what the test actually measures.

Incorrect Scale. Interpretations that are on an incorrect scale but correct direction, such as interpreting changes in likelihoods as linear increases. These result in an interpretation of an effect that is in the right direction (positive or negative) but with an incorrect magnitude.

Sub-optimal or No Scale. This can involve interpreting on a log scale (for logistic or ordinal logistic regressions) rather than transforming first into more intuitively understandable scales. We also classify interpretations that only report the significance or general direction of coefficients without interpreting the magnitude of the effect as sub-optimal.

Interpreting Distributions and Not the Model. Regressions are meant to evaluate the impact or predictive power of specific independent variables on the dependent variable, and not whether there exist differences between levels of the independent variables. For example, logistic regressions do not evaluate whether *more* units with specific traits (predictors) take a specific action but rather how much *more likely* units with these specific traits are to take a specific action (compared with the baseline set of traits). The former (sub-optimal) interpretation is a description of the underlying data, while the latter explanation is an interpretation of the odds ratios resulting from the model. This code only applies to regressions.

No Interpretation. This indicates instances that report on only the presence or absence of statistically significant results. One example is reporting significance levels (or coefficients) in a (regression) table without any explanation in the text of the interpretation of any results.

Each instance is coded with one type of interpretation issue. If a there are multiple interpretations from one test (e.g., different interpretations of coefficients from a regression model), each distinct interpretation is coded as a separate instance.

3.3 Evaluating the Severity of Issues on Assertions

To understand the potential impact of issues in statistical methods, reporting, or interpretation on the generation of knowledge in usable security and privacy, we evaluate whether any instance from the inferential statistical test that we examined was used to support an assertion central to the paper. If the variables or inferences from the results of a test are mentioned in the abstract or introduction of the paper we code that instance of the statistical test as supporting a *main assertion*. We code tests on variables that are not mentioned in the paper abstract or introduction as supporting *peripheral assertions*. We note that not all uses of NHST are necessarily meant to be formal hypothesis tests. Therefore, peripheral claims could be a proxy for exploratory uses of NHST and could have less

complete reporting.

Finally, we evaluate and code whether each assertion about the presence (or absence) of an effect of some size (scientific significance), accompanied by a p-value (statistical significance) based on the results from each NHST instance—whether a main or a peripheral assertion—is supported by the information reported in the paper:

Fully Supported. There are no statistical validity, reporting, or interpretation issues. There is sufficient information to verify the statistical validity, as well as sufficient evidence to evaluate (the statistical and scientific significance of) the assertion beyond a statement of p-value. The assertion is properly interpreted in the context of the statistical test.

Mostly Supported. The issues present are not severe. The reporting supports the assertion's claims regarding statistical significance, as well as effect size and direction. There are no statistical validity issues. There are some light reporting or interpretation issues that do not greatly impact evaluation of the assertion. *Reporting* issues that lead to this code are **insufficient reporting on *statistical* significance** and insufficient reporting of measures of variability (under the **insufficient reporting on *scientific* significance** code). *Interpretation* issues that lead to this code are **No Interpretation** and most subcodes of **sub-optimal effect interpretation** apart from incorrect scale, as that interpretation issue changes the magnitude of results.

Partially Supported. The issues present are somewhat severe and can impact replication. The reporting supports the assertion's claims regarding statistical significance, but there is incorrect or missing information about *scientific* significance (effect size and direction). There are no statistical validity issues. There are reporting or interpretation issues that impact support for the assertion. Instances of NHST that do not report adequate information about relevant characteristics of the data (**insufficient reporting of scientific significance**) lead to this code, as readers have no evidence to further evaluate the assertion (such as a magnitude of effect) apart from a paper's claim that a statistical test resulted in a specific p-value. Misinterpretations that are on an incorrect scale—affecting the magnitude of the effect—also fall into this code.

Not supported. The issues are the most severe, as statistical validity is threatened. These assertions often cannot be replicated or independently verified without the original dataset. The reporting does not support the assertion's claims regarding statistical significance as statistical validity issues or misinterpretations undermine support for the assertion. Instances with a statistical validity issue (apart from unverifiable) fall under this code, as the results of the statistical test on which the assertion is based are not suited for what the test evaluates. This category also identifies assertions that are at odds with the results of the test (**misinterpretations**). This code identifies assertions where the issues are such that the test provides no support either for the claim or for its complement due to issues with statistical validity or to misinterpretations of the results.

Lacking Information. The issues are severe, as there is insufficient information to determine if the assertion is supported. This lack of information also hinders replication and independent verification of assertions. Instances with the statistical validity issue **unverifiable** or the reporting issue **insufficient reporting to evaluate validity** fall under this code. Given lack of or imprecise reporting, the assertions in this category cannot be verified to follow from the results of the reported statistical investigation.

We holistically evaluate whether assertions are supported by coding for three aspects: statistical validity, reporting, and interpretation. We use the most severe issue with the instance to categorize assertions. For instance, if an instance uses an **incorrect** statistical test but fully reports all details of the data, we code the assertion as **not supported**, as the statistical test results are unable to support the assertion.

3.4 Coding Scope and Limitations

Our systematic review focuses on specific types of issues—statistical validity, reporting, and interpretation—and their impact on the assertions in UPS work based on NHST. It does not cover all factors that may impact the statistical or scientific accuracy of results of a statistical test.

We emphasize that we are not coding if something is *true in reality* (i.e., that certain effects or differences are present in the general population). Determination of an objective truth is outside the scope of this work as such determinations involve repeated replications, representative and generalizable samples. Thus, these truths cannot be established within any one statistical test or paper. Indeed, such a task is perhaps instead the overarching goal of scientific endeavor. Rather, we are investigating whether, given the information present within each paper, the assertions made on the basis of results of a statistical investigation are supported.

We do not code for whether tests meet all their underlying assumptions, such as equal variance assumptions. There is often insufficient information to adequately evaluate these assumptions since the robustness of statistical tests to various assumptions is often tied to sample size and presence of outliers, and are best evaluated on a case-by-case basis [87]. In usable privacy and security, often working with data from real human users, almost no test will have all assumptions perfectly satisfied. For example, obtaining truly random samples perfectly representative of the population is difficult, and there will likely always be some statistical noise and multicollinearity within datasets. Hence, we adopt a conservative approach, coding things as incorrect only when we are fully confident that they are. For example, we assume that data is "reasonably" distributed. When evaluating overall validity of assertions, we do not require that papers describe a test for normality for all tested samples when running parametric tests with a normality assumption, nor do we require an

explicit statement of "all else being held constant" for regressions. Other analyses of the use of statistics have been more pedantic [16, 66], and so the results of our analyses should be interpreted as a lower bound of potential issues.

We consider the reporting and interpretation of statistical tests as sufficient as long as the required information appears anywhere in the paper (including figures, methods, appendices, or supplementary material) even if not explicitly referred to in the reporting. As our coding was conducted mainly by one coder and only a subset verified by a second one, some issues may have been overlooked. However, the results from our code resolution process (§3.2) provides some confidence that our dataset is a lower bound of the issues that are present in our sample. We provide the codebook in Appendix A so that readers can independently verify our codes.

Finally, in our codes, even if there is an issue in one category, we still code all other categories with the assumption that the initial category was correct. For example, if a statistical validity issue (e.g., non-independence) is present, we still code the claim for reporting issues and interpretation in the context of the statistical test stated, as the validity of the test does not change the expectation of transparent reporting and accurate interpretation.

4 Results

We identified 879 statistical issues over 479 NHST instances in 121 papers. We find that most usable security and privacy papers within our dataset have at least one instance with a statistical validity (65% of papers; §4.1), reporting (95% of papers; §4.2), or interpretation issue (45% of papers; §4.3), of varying levels of severity (§4.4). Across all assertions based on NHST, 83% are not fully supported by the reported information, with some issues more severe than others. While 21% of assertions are mostly supported by the reporting in the paper, 21% of assertions are not supported, and 23% lack information to judge statistical support, given the reporting in the papers. Table 1 and Table 2 summarize the results.

4.1 Statistical Validity Issues

Statistical validity issues affect 165 instances (34%) across 79 papers (65%). These issues arise from statistical tests that are either unsuitable or whose suitability cannot be verified.

Of those, 45 instances (9%) in 28 papers (23%) use at least one **incorrect statistical test** that does not correctly account for the structure of dependencies between units or measurements. The most common error is failure to account for repeated measures (30 instances), while in 9 instances independent tests were used on dependent (often paired) data, and independent data was arbitrarily paired in 6 instances.

24 instances (5%) across 16 papers (13%) in our sample use tests where there is a **data type mismatch**. The majority of these cases (17 instances) are due to treating ordinal data as continuous or interval data.

Across 96 instances (20%) in 62 papers (51%), at least one statistical test is insufficiently reported, leading to situations where it is **unverifiable whether the test is statistically valid** for the data. This often arises due to a lack of clarity in reporting tests and models.

4.2 Reporting Issues

Reporting issues are the most common, affecting 360 instances (75%) and occurring in 115 papers (95%). These issues interfere with the reader's ability to evaluate the results due to a lack of reported evidence to support statistical or scientific significance of the assertion.

The majority of papers (248 instances, 52%; corresponding to 104 papers, 86%) had **incomplete statistical significance reporting**, meaning the test statistic and associated degrees of freedom, or the exact p-value, were not reported.

256 instances (53%) across 97 papers (80%) provide **incomplete scientific significance context**, hindering evaluation of the scientific significance of the results. Often (69 instances over 46 papers) this is due to solely reporting that a result is (non-)significant without reporting any other evidence. Point estimates (such as measures of centrality or model coefficients) are provided without any variability (e.g., standard deviations or confidence intervals) in 84 instances. Another notable issue is the lack of evaluation of model fit, missing in 41 instances out of 75 instances based on regression models.

122 instances (25%) in 72 papers (60%) report **insufficient information to evaluate statistical validity**. In 58 instances (12%), readers are not provided with information to identify the specific statistical test used. In 64 instances (13%), there is lack of clarity in test or model specification, such as not stating how variables are constructed, or what the independent variables in a regression model are. This type of reporting issue can also impact evaluations of statistical validity. Given that we code each type of issue (statistical validity, reporting, interpretation) assuming correctness in the other categories, there are more instances in this category than in the "Unverifiable" statistical validity category because some instances are classified as "Data Type Mismatch" or "Incorrect" given other information. For example, a paper can use a regression model that does not account for repeated measures ("Incorrect") but also provide no details on the construction of independent variables ("insufficient information to evaluate statistical validity").

4.3 Interpretation Issues

Even when reporting is accurate and the statistical test is suitable for the data, specific interpretations of the results may be misleading or inaccurate. These types of issues occur

Table 1: Summary of results. Our analysis examined 479 NHST instances from 121 papers, which are the totals used to compute the percentages. Percentages may total more than 100% as each paper may use multiple distinct statistical tests or instances.

Issue Type	Num. Papers Affected	% Papers Affected	Num. Instances Affected	% Instances Affected
Statistical Validity Issues	79	65.29	165	34.45
Incorrect	28	23.14	45	9.39
Data Type Mismatch	16	13.22	24	5.01
Unverifiable	62	51.24	96	20.04
Reporting Issues	115	95.04	360	75.16
statSig Reporting Issues	104	85.95	248	51.77
sciSig Reporting Issues	97	80.17	256	53.44
statValidity Reporting Issues	72	59.50	122	25.47
Interpretation Issues	54	44.63	88	18.37
Misinterpretation	32	26.45	40	8.35
Sup-optimal Interpretation	30	24.79	37	7.72
No Interpretation	9	7.44	11	2.30
Some Issue	118	97.52	397	82.88

the least frequently, in 88 instances (18%) across 54 papers (45%).

Misinterpretations of test results arise in 40 instances (8%) from 32 papers (26%), resulting in inferences that are not supported by the data. The most common misinterpretation (13 instances) uses results from an omnibus test to report a statistically significant difference for a pairwise comparison. 10 instances misinterpret statistical significance.

37 instances (8%) across 30 papers (25%) have **sub-optimal interpretations**. The most frequent are interpretations that use an incorrect scale (such as linear scales for logistic regressions), use an unintuitive scale, or fail to interpret magnitudes from regression coefficients at all (26 instances).

The results of 11 statistical tests (2%) in nine papers (7%) have **no interpretation**. While presenting the results is useful, the absence of any further discussion implicitly assumes that statistical significance alone provides sufficient information.

4.4 Evaluating Severity and Impact of Issues

While assertions in ∼97% of papers contain at least one issue, not all these issues are equally severe or have similar impact on the assertions following from NHST. We provide a breakdown in Table 2 and discuss their varying severities. Across all 479 assertions that follow from the instances of NHST, 82 assertions (17%) are **fully supported** by the information reported in the paper. Of the rest, 103 (22%) are **mostly supported**, as the reported context was sufficient for a reader to evaluate the accuracy of the assertion, even if some pieces were not entirely present; or the issues were minor enough that they do not drastically affect the direction and magnitude of the assertion (such as believing a difference exists or that some variables are positively correlated). There is not enough information reported to fully support and contextualize the assertion, but the statistical test chosen is suitable for the data, and the direction and magnitude of the assertion is supported by the reported information. These issues lightly impact independent verification.

81 of assertions (17%) are only **partially supported** by the information given from the papers in which they appear. For many of these assertions, there is only enough information to verify that a specific (suitable) statistical test resulted in some p-value, with insufficient evidence to further contextualize that result in terms of its effect size and direction. In other cases, the interpretation of the result is on an incorrect scale, changing the magnitude (but not direction) of the effect. In these cases, the reported information contains some evidence that supports that the assertion is generally true (e.g., presence of effect) but not enough to evaluate specifics. These issues undermine support for the assertion, as the magnitude of the effects can differ, or cannot be adequately evaluated given the lack of reporting on context. Nonetheless, the reported information supports the statistical significance and general direction (e.g., existence of some effect) of the assertion. These issues are somewhat severe as they may cause further difficulties for comparisons in future replication.

101 assertions (21%) are **not supported** by the reported information, due to issues with statistical validity or misinterpretations of the results. These are the most severe issues. While the assertion that the test attempts to support may still be correct, the test and its reporting provide no positive evidence to support the assertion. A common cause is that the test is not appropriate for the type or structure of data involved. These issues are the most important to address since they can have far-reaching implications for claims and interventions based on these assertions, as the assertions are not properly backed by the statistical methods.

Finally, 112 assertions (23%) are **lacking information** for a reader to judge statistical support. The chosen statistical tests may or may not be suitable for the data, but there is insufficient information provided to tell. This is still a se-

Table 2: Summary of evaluation of the support for 479 assertions from the reported data in each paper. Detailed breakdowns under each category report the prevalence of each issue. The issue and paper counts may sum to more than 100%.

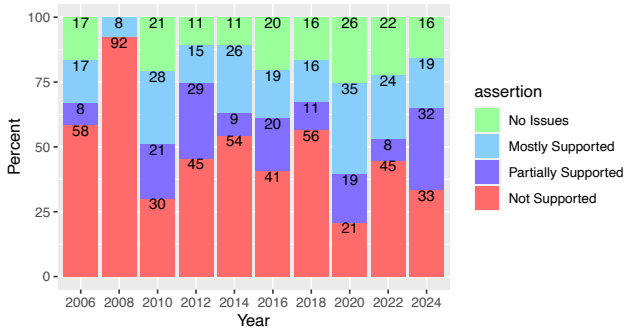
Assertion Evaluation Category	Num. Papers Affected	% Papers Affected	Num. Assert. Affected	% Assert. Affected
Fully Supported	54	44.63	82	17.12
Mostly Supported	70	57.85	103	21.50
statSig Reporting Issues	48	39.67	61	12.73
sciSig Reporting Issues (Insufficient variability)	25	20.66	34	7.1
Sup-optimal Interpretation	14	11.57	16	3.34
No Interpretation	6	4.96	6	1.25
Partially Supported	53	43.80	81	16.91
sciSig Reporting Issues (Insufficient Effect Size or Direction)	51	42.15	76	15.87
Sup-optimal Interpretation (Incorrect Scale)	7	5.79	7	1.46
Not Supported	57	47.11	101	21.09
Incorrect	28	23.14	45	9.39
Data Type Mismatch	16	13.22	24	5.01
Misinterpretation	32	26.45	40	8.35
Lacking Information	70	57.85	112	23.38
Unverifiable	61	50.41	86	17.95
statValidity Reporting Issues	62	51.24	99	20.67

vere issue, as this lack of information undermines support for the assertion and degrades the ability for replication and independent verification of the results.

We now examine the potential broader impact of these statistical issues. The proportion of assertions in each of the above categories differs between main (350 assertions) and peripheral assertions (129 assertions). The proportions of fully supported (17% main, 16% peripheral) and mostly supported assertions (22% main, 20% peripheral) differ by less than 5% between these groups. A slightly higher percentage of peripheral assertions (22%) than main assertions (15%) are partially supported. The same relationship holds for assertions that are lacking information (22% main, 28% peripheral). This suggests that more exploratory hypotheses that may not be envisioned as formal NHST feature less reporting, perhaps due to space constraints. Lastly, more main assertions (24%) than peripheral assertions (14%) are not supported. This may be due to peripheral assertions more commonly lacking information, preventing us from evaluating their suitability.

Finally, we examined the historical breakdown of incorrect assertions and we report the results in Figure 1. We see no discernable pattern indicating any fundamental changes over time or any obvious positive or negative trend.

Figure 1: The percentages of assertions in each category by year. Percentages are also marked on each bar.



5 Discussion

Our analysis of a sample of two decades' worth of SOUPS papers has revealed issues of varying levels of severity in 83% of assertions across most papers (97%). They include relatively minor reporting issues that do not follow best practices, as well as more severe lack of reporting hindering readers from adequately evaluating relevant details regarding the quantitative methods used. The most severe issues are those papers that contain issues with the choice of statistical test, undermining support for assertions, or misinterpretations that make assertions that do not follow from the results of NHST.

There is no one-size-fits-all recommendation for statistical reporting, and recommendations must necessarily differ by field and be guided by the scientific context [60,81]. Grounded in the trends we observed in our study, the statistical literature, and other methodological work in HCI and beyond, we provide a set of recommendations for authors of works that use or intend to use inferential statistical methods for statistical selection (§5.1), reporting (§5.2), and interpretation (§5.3). We reflect on suggestions for the UPS field as a whole, including authors, reviewers, and researchers in §5.4.

5.1 Statistical Selection Recommendations

Statistical tests must suit the data they analyze. This was not the case in 14% of the instances we evaluated. We recommend creating an analysis plan for data prior to collection and analysis and using statistical hypothesis testing "only when you want to test a well-defined hypothesis" [79]. This can involve pre-determining hypotheses and articulating the exact purpose a test for statistical significance is intended to serve. Some [21, 30, 71, 77, 83] have advocated for pre-registering complete protocols before data collection and analysis. While protocol pre-registration may be a best practice, simply making explicit the hypotheses and the corresponding tests ahead of time would already considerably help. In some cases, principled evaluations of potential effect sizes given the data and expert judgment may serve similarly well as statistical testing. As a corollary, testing all variables, such as demographic variables, is unnecessary without a relevant hypothesis [82].

5.2 Reporting Recommendations

The majority of issues we found in statistical use arose from lack of reporting transparency. Based on statistical reporting guidance from various fields, guidelines for reporting transparency [3, 4, 24, 36, 71, 77, 83], and the results from this systematic review, we create a set of recommended "minimally sufficient" information that provide adequate information to for readers to evaluate assertions.

Precise Test Name. The statistical test name reported should provide readers sufficient information to distinguish between non-paired and paired versions of a test, whether the test accounts for repeated measures, and what type of data the test is structured to evaluate.

Data Characteristics. Data collection methods should be clearly described so that readers can evaluate whether the variables analyzed are independent or dependent, and whether statistical methods should account for repeated measures. Stating only whether an experimental design was within or between subjects is not always sufficient. For instance, studies exposing users to multiple scenarios may remain between subjects but analysis must account for repeated measures. Lack of accounting for repeated measures is the most prominent statistical validity issue we identified.

Dependent and Independent Variables. What data (or groups) are being evaluated in the statistical procedure, and how the data is structured should be reported. For categorical variables in regression models, authors should state the baseline. Variable transformation—such as binning continuous variables or aggregating multiple measures—should be clearly described. Which groups or variables are being compared should be stated. Without knowing what was being tested, readers cannot verify whether the test is suitable and whether the results can support the assertions. For regressions, authors should explicitly mention the dependent variable and all independent variables or predictors (including any interactions), to provide context for a knowledgeable reader to evaluate the validity of and confidence in the specific model (and what it seeks to test), as there may be many similarly statistically valid models [14].

To provide sufficient information for readers to evaluate both the statistical significance and scientific significance of a test, we recommend reporting the following information [24, 36, 40, 57, 62, 66, 71].

Magnitude and Direction of Effects. Effects can be reported as standardized effect size measures (e.g., Cohen's d, Pearson's r), in the original units of measurement (e.g., regression coefficients), or as other interpretable forms (e.g., difference in means) [51]. Other context such as descriptive statistics can also be used in lieu of effect size to aid in the evaluation of scientific significance and provide evidence for the assertion. If so, the reporting of context should suit the type of statistical procedure. For example, if comparing multiple groups, the information for each group, rather than in aggregate, should be reported.[5] For categorical measures, frequencies of the categories of interest are also sufficiently illustrative. Data can be reported in visual formats with exact values provided in an appendix or supplementary material or as descriptive statistics. This conveys information beyond a strict statistical significance threshold and has implications for scientific significance as well as provides further evidence to support the assertions made. Provision of a full dataset is the ideal, though not always possible.

Measures of Uncertainty, Variability, or Confidence. Authors should quantify the precision of point estimates, to give readers context about uncertainty. For means, this could include reporting standard deviations; for coefficients, this could mean reporting confidence intervals.

Model Fit. Regression model fit should be evaluated, through R^2 or pseudo-R^2 values, or results from comparison to a null model.[6] Regression models are necessarily reductive estimates of complex natural phenomena. Unmeasured values, small variations, or different variable selections may lead to contradictory—yet similarly statistically valid—conclusions [6, 14]. Some measure of model fit can help readers gauge model usefulness in describing a complex reality.

In the process of writing a research paper, authors often become so familiar with a dataset that they may make (informed) assumptions about the data. Their audience, on the other hand, does not have that ability. Thus, similar to existing recommendations, we recommend having a reader external to the project read the paper [67]. We extend this one step further by recommending that authors ask a statistically literate reader to check over methods and reporting, to ensure reporting is clear and understandable.

[5] This may not always be possible for data with repeated measures.

[6] Certain measures of model fit, such as AIC or BIC, have no meaning in isolation. Such numbers must be reported in relation to a model comparison.

5.3 Interpretation Recommendations

At their core, statistical tests are used to provide some method to investigate the data and provide some value that quantifies the effects. At a high level, we recommend against implicitly equating statistical significance with scientific significance. Underpowered tests may not detect differences that result in a significant p-value, even when these differences exist in the population. Lack of statistical significance of a predictor in a regression model does not indicate that it does not impact the outcome variable, only that it has no statistically significant impact *in the context of the other predictors in the model*. Therefore, we suggest reporting sufficient information to evaluate all hypotheses, even if p-values are above the stated threshold [16, 83]. These can be reported in an appendix or supplementary material, and should not incur unreasonable burden if the number of hypotheses tested is limited, following our recommendations for judicious use of statistical significance testing (§5.1).

Make Results Intuitive. Statistical test results are not always intuitive. Numbers should be converted to interpretations understandable by people with modest statistical background. For example, log-odds can be transformed into odds ratios, which can be further transformed into predicted probabilities. Rather than just leaving numbers in a regression table and expecting the reader to interpret them and understand their statistical and scientific significance, we recommend interpreting the relevant effects and impacts, such as explaining the magnitude of results. There are many different ways of conveying effect size in HCI, and authors can decide what works best in the context of their work [51].

Place Results in (Field-Specific) Context. While standardized effect sizes allow for a quick evaluation of effect magnitude, reporting underlying data allows for a richer understanding of scientific, and practical significance. For example, a 1% increase in conversion rate for a large web advertising platform may add millions of dollars of revenue, while a 1% change in perceived quality of experience in video streaming may be barely noticeable.

5.4 What Is Next?

We are not advocating to stop relying on NHST, but instead are encouraging the UPS field to reflect on the role of statistical expertise, use statistical significance as one of many tools, and embrace uncertainty and complexity [6, 74, 81].

Statistical Expertise and Statistical Thinking. The papers in which we found reason for concern about the use of statistics all underwent peer review, suggesting that program committees may need to increase their commitment to identifying and helping remediate potentially improper use of statistics. Papers using NHST should have at least one reviewer able to evaluate statistics to raise any questions or concerns regarding statistical methods or reporting during the review process [16]. Software packages may also be able to flag potential issues ahead of time for easier identification, or aid PC members in verifying statistical results [36, 66]. Some institutions also have statistical experts who give consultations on applied statistical use for research projects, and we recommend that authors use such resources when available.

Transparent Reporting, Repeated Replications. While the issues that surfaced in this work vary in severity, they were all gleaned from *the information reported in the papers*. Regardless of the severity of the issues, replications and verification of results from NHST can provide further confidence in the conclusions from UPS research. Assertions based on invalid statistical methods may also have less negative impact if studies can be replicated with correct statistical methods to validate (or invalidate) findings. This requires sufficient and transparent reporting to understand the statistical test and evaluate the assertions, and thus we recommend transparent reporting for works using NHST and replication of studies. When facing space constraints, appendices and supplementary material can be used to report full statistical details.

Embrace Complexity and Uncertainty. Statistical methods are an attempt to map a complex reality into a few comprehensible numbers to allow for interpretation and conclusion.While experiments and procedures in physical sciences may be able to strictly control for confounding factors, studies with human users rarely can. There is variation among and within users, and it is difficult—if not impossible—to adequately capture all that complexity. Given this, we call upon UPS and HCI researchers to embrace this variation and uncertainty as a feature of their work [6, 28, 49, 81].

6 Conclusion

Our statistical methods and reporting should reflect the complexity and variety of the people whom we study. Focusing on statistical significance and dichotimizing results and reporting based on an arbitrary statistical cutoff (e.g., "p-values") at the expense of reporting other relevant context reduces the richness inherent in the data and reduces the inferences we are able to draw. In our stratified sample of UPS papers published in the last 20 years, we found statistical validity, reporting, and interpretation issues of varying severity. Many papers relied solely on statistical significance, and failed to report sufficient information for readers to adequately evaluate whether the claims of statistical and scientific significance in the assertions hold. Hence, we advocate for the use of NHST and statistical significance as a starting point, and holistic reporting and clear interpretation to provide information on the scientific significance and practical impact of the research. Instead of reducing human complexity into comparisons against static thresholds such as $p < 0.05$, we argue that we should report all information and context that allow readers to validate, evaluate, and replicate results and assertions in a more complete context, with all its richness and complexity.

Acknowledgments

The authors would like to thank Sakshi Singh for her help in coding papers. We also thank Peggy Wang and Jasmine Li for providing feedback and suggestions on statistical questions.

References

[1] Submission Guidelines - 9th Workshop on Socio-Technical Aspects in Security. https://stast2019.uni.lu/papersubmission.html, 2019. Visited 2025-06-09.

[2] *Publication Manual of the American Psychological Association*. American Psychological Association, Washington, DC, 7th. edition, 2020.

[3] Balazs Aczel, Barnabas Szaszi, Alexandra Sarafoglou, Zoltan Kekecs, Šimon Kucharskỳ, Daniel Benjamin, Christopher D Chambers, Agneta Fisher, Andrew Gelman, Morton A Gernsbacher, et al. A consensus-based transparency checklist. *Nature human behaviour*, 4(1):4–6, 2020.

[4] Lena Fanya Aeschbach, Sebastian A.C. Perrig, Lorena Weder, Klaus Opwis, and Florian Brühlmann. Transparency in measurement reporting: A systematic literature review of chi play. *Proceedings of the ACM on Human-Computer Interaction*, 5(CHI PLAY):1–21, 2021.

[5] Andrew D. Althouse, Jennifer E. Below, Brian L. Claggett, Nancy J. Cox, James A. De Lemos, Rahul C. Deo, Sue Duval, Rory Hachamovitch, Sanjay Kaul, Scott W. Keith, et al. Recommendations for statistical reporting in cardiovascular medicine: a special report from the American Heart Association. *Circulation*, 144(4):e70–e91, 2021.

[6] Valentin Amrhein, David Trafimow, and Sander Greenland. Inferential statistics as descriptive statistics: There is no replication crisis if we don't expect replication. *The American Statistician*, 73(sup1):262–270, 2019.

[7] American Psychological Association. Quantitative Research Design (JARS–Quant). https://apastyle.apa.org/jars/quantitative, 2018.

[8] Peter Bacchetti, Steven G. Deeks, and Joseph M. McCune. Breaking free of sample size dogma to perform innovative translational research. *Science translational medicine*, 3(87):87ps24–87ps24, 2011.

[9] Nick Ballou, Vivek R. Warriar, and Sebastian Deterding. Are you open? a content analysis of transparency and openness guidelines in HCI journals. In *Proceedings of the 2021 CHI Conference on human factors in computing systems*, pages 1–10, 2021.

[10] M.S. Bartlett. The effect of non-normality on the t distribution. In *mathematical proceedings of the cambridge philosophical society*, volume 31, pages 223–231. Cambridge University Press, 1935.

[11] Craig M. Bennett, Michael B. Miller, and George L. Wolford. Neural correlates of interspecies perspective taking in the post-mortem Atlantic salmon: An argument for multiple comparisons correction. *Neuroimage*, 47(Suppl 1):S125, 2009.

[12] Craig M. Bennett, George L. Wolford, and Michael B. Miller. The principled control of false positives in neuroimaging. *Social cognitive and affective neuroscience*, 4(4):417–422, 2009.

[13] Rebecca A. Betensky. The p-value requires context, not a threshold. *The American Statistician*, 73(sup1):115–117, 2019.

[14] Leo Breiman. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3):199–231, 2001.

[15] Kelly Caine. Local standards for sample size at CHI. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 981–992, 2016.

[16] Paul Cairns. HCI... not as it should be: inferential statistics in HCI research. In *Proceedings of HCI 2007 The 21st British HCI Group Annual Conference University of Lancaster, UK*. BCS Learning & Development, 2007.

[17] Robert J. Calin-Jageman and Geoff Cumming. The new statistics for better science: Ask how much, how uncertain, and what else is known. *The American Statistician*, 73(sup1):271–280, 2019.

[18] James Carifio and Rocco J. Perla. Ten common misunderstandings, misconceptions, persistent myths and urban legends about Likert scales and Likert response formats and their antidotes. *Journal of social sciences*, 3(3):106–116, 2007.

[19] Chris Chatfield. Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 158(3):419–444, 1995.

[20] Fanny Chevalier, Lewis Chuang, Pierre Dragicevic, Shion Guha, Steve Haroz, Matthew Kay, and Chat Wacharamanotham. Transparent Statistics in Human–Computer Interaction. https://transparentstatistics.org/, 2020. Visited 2025-06-09.

[21] Lewis L. Chuang and Ulrike Pfeil. Transparency and openness promotion guidelines for HCI. In *Extended abstracts of the 2018 CHI conference on human factors in computing systems*, pages 1–4, 2018.

[22] Andy Cockburn, Carl Gutwin, and Alan Dix. Hark no more: On the preregistration of CHI experiments. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–12, 2018.

[23] Jacob Cohen. *Statistical power analysis for the behavioral sciences*. Routledge, 2013.

[24] Kovila Coopamootoo and Thomas Groß. A codebook for evidence-based research: The nifty nine completeness indicators v1. 1. *School of Computing Technical Report Series*, 2017.

[25] Kovila Coopamootoo and Thomas Groß. A systematic evaluation of evidence-based methods in cyber security user studies. *School of Computing Technical Report Series*, 2019.

[26] Geoff Cumming and Sue Finch. Inference by eye: confidence intervals and how to read pictures of data. *American psychologist*, 60(2):170, 2005.

[27] Joost C.F. De Winter and Dimitra Dodou. Five-point Likert items: t test versus Mann-Whitney-Wilcoxon. *Practical assessment, research & evaluation*, 15(11):1–12, 2010.

[28] Pierre Dragicevic. Fair statistical communication in HCI. *Modern statistical methods for HCI*, pages 291–330, 2016.

[29] Morten W. Fagerland. t-tests, non-parametric tests, and large studies—a paradox of statistical practice? *BMC medical research methodology*, 12:1–7, 2012.

[30] Wolfgang Forstmeier, Eric-Jan Wagenmakers, and Timothy H. Parker. Detecting and avoiding likely false-positive findings–a practical guide. *Biological Reviews*, 92(4):1941–1968, 2017.

[31] Ronald D. Fricker Jr, Katherine Burke, Xiaoyan Han, and William H. Woodall. Assessing the statistical analyses used in basic and applied social psychology after their p-value ban. *The American Statistician*, 73(sup1):374–384, 2019.

[32] Steven N. Goodman. Why is getting rid of p-values so hard? Musings on science and statistics. *The American Statistician*, 73(sup1):26–30, 2019.

[33] Romain-Daniel Gosselin. Insufficient transparency of statistical reporting in preclinical research: a scoping review. *Scientific Reports*, 11(1):3335, 2021.

[34] Sander Greenland, Stephen J. Senn, Kenneth J. Rothman, John B. Carlin, Charles Poole, Steven N. Goodman, and Douglas G. Altman. Statistical tests, p values, confidence intervals, and power: a guide to misinterpretations. *European journal of epidemiology*, 31(4):337–350, 2016.

[35] Thomas Groß. Statistical reliability of 10 years of cyber security user studies. In *International Workshop on Socio-Technical Aspects in Security and Trust*, pages 171–190. Springer, 2020.

[36] Thomas Groß. Fidelity of statistical reporting in 10 years of cyber security user studies. In *Socio-Technical Aspects in Security and Trust: 9th International Workshop, STAST 2019, Luxembourg City, Luxembourg, September 26, 2019, Revised Selected Papers 9*, pages 3–26. Springer, 2021.

[37] Thomas Groß. Why most results of socio-technical security user studies are false–and what to do about it. In *International Workshop on Socio-Technical Aspects in Security and Trust*, 2022.

[38] Xavier A. Harrison. Using observation-level random effects to model overdispersion in count data in ecology and evolution. *PeerJ*, 2:e616, 2014.

[39] Norbert Hirschauer, Sven Grüner, and Oliver Mußhoff. The p-value and statistical significance testing. In *Fundamentals of Statistical Inference: What is the Meaning of Random Error?*, pages 63–96. Springer, 2022.

[40] Kasper Hornbæk. Some whys and hows of experiments in human–computer interaction. *Foundations and Trends® in Human–Computer Interaction*, 5(4):299–373, 2013.

[41] Kasper Hornbæk, Søren S. Sander, Javier Andrés Bargas-Avila, and Jakob Grue Simonsen. Is once enough? On the extent and content of replications in human-computer interaction. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 3523–3532, 2014.

[42] Douglas W. Hubbard and Alicia L. Carriquiry. Quality control for scientific research: Addressing reproducibility, responsiveness, and relevance. *The American Statistician*, 73(sup1):46–55, 2019.

[43] Raymond Hubbard. Will the ASA's efforts to improve statistical practice be successful? Some evidence to the contrary. *The American Statistician*, 73(sup1):31–35, 2019.

[44] John P.A. Ioannidis. Why most published research findings are false. *PLoS medicine*, 2(8):e124, 2005.

[45] John P.A. Ioannidis. What have we (not) learnt from millions of scientific papers with p values? *The American Statistician*, 73(sup1):20–25, 2019.

[46] P.K. Ito. 7 robustness of ANOVA and MANOVA test procedures. *Handbook of statistics*, 1:199–236, 1980.

[47] Susan Jamieson. Likert scales: How to (ab) use them? *Medical education*, 38(12):1217–1218, 2004.

[48] Maurits Kaptein and Judy Robertson. Rethinking statistical analysis methods for CHI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1105–1114, 2012.

[49] Matthew Kay, Steve Haroz, Shion Guha, and Pierre Dragicevic. Special interest group on transparent statistics in HCI. In *Proceedings of the 2016 CHI conference extended abstracts on human factors in computing systems*, pages 1081–1084, 2016.

[50] Matthew Kay, Gregory L. Nelson, and Eric B. Hekler. Researcher-centered design of statistics: Why Bayesian statistics better fit the culture and incentives of HCI. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 4521–4532, 2016.

[51] Yea-Seul Kim, Jake M. Hofman, and Daniel G. Goldstein. Putting scientific results in perspective: Improving the communication of standardized effect sizes. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2022.

[52] Jan H. Klemmer, Juliane Schmüser, Byron M. Lowens, Fabian Fischer, Lea Schmüser, Florian Schaub, and Sascha Fahl. Transparency in usable privacy and security research: Scholars' perspectives, practices, and recommendations. In *In 46th IEEE Symposium on Security and Privacy*, 2025.

[53] Thomas R. Knapp. Treating ordinal scales as interval scales: an attempt to resolve the controversy. *Nursing research*, 39(2):121–123, 1990.

[54] Ulrich Knief and Wolfgang Forstmeier. Violating the normality assumption may be the lesser of two evils. *Behavior Research Methods*, 53(6):2576–2590, 2021.

[55] William Kuzon, Melanie Urbanchek, and Steven McCabe. The seven deadly sins of statistical analysis. *Annals of plastic surgery*, 37:265–272, 1996.

[56] Kyle M. Lang, Shauna J. Sweet, and Elizabeth M. Grandfield. Getting beyond the null: Statistical modeling as an alternative framework for inference in developmental science. *Research in Human Development*, 14(4):287–304, 2017.

[57] Thomas A. Lang and Douglas G. Altman. Statistical analyses and methods in the published literature: The SAMPL guidelines. *Guidelines for reporting Health Research: A User's manual*, pages 264–274, 2014.

[58] Ivy Liu and Alan Agresti. The analysis of ordered categorical data: An overview and a survey of recent developments. *Test*, 14:1–73, 2005.

[59] Blakeley B. McShane, Eric T. Bradlow, John G. Lynch Jr, and Robert J. Meyer. "Statistical significance" and statistical reporting: Moving beyond binary. *Journal of Marketing*, 88(3):1–19, 2024.

[60] Blakeley B. McShane, David Gal, Andrew Gelman, Christian Robert, and Jennifer L. Tackett. Abandon statistical significance. *The American Statistician*, 73(sup1):235–245, 2019.

[61] Kane Nashimoto and F.T. Wright. Nonparametric multiple-comparison methods for simply ordered medians. *Computational statistics & data analysis*, 51(10):5068–5076, 2007.

[62] Anna-Marie Ortloff, Christian Tiefenau, and Matthew Smith. SoK: I have the (developer) power! sample size estimation for Fisher's exact, chi-squared, McNemar's, Wilcoxon rank-sum, Wilcoxon signed-rank and t-tests in developer-centered usable security. In *Nineteenth Symposium on Usable Privacy and Security (SOUPS 2023)*, pages 341–359, 2023.

[63] Georgios D. Panos. Violating independence assumption in medical statistics. *Lancet (London, England)*, 404(10456):934–935, 2024.

[64] Jose D. Perezgonzalez. Fisher, Neyman-Pearson or NHST? A tutorial for teaching data testing. *Frontiers in psychology*, 6:223, 2015.

[65] Fred Ramsey and Daniel Schafer. *The statistical sleuth: A course in methods of data analysis. 3rd*. Cengage Learning, 2013.

[66] Kavous Salehzadeh Niksirat, Lahari Goswami, Pooja S.B. Rao, James Tyler, Alessandro Silacci, Sadiq Aliyu, Annika Aebli, Chat Wacharamanotham, and Mauro Cherubini. Changes in research ethics, openness, and transparency in empirical studies between CHI 2017 and CHI 2022. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–23, 2023.

[67] Stuart Schechter. Common pitfalls in writing about security and privacy human subjects experiments, and how to avoid them. Technical Report MSR-TR-2013-5, Microsoft, January 2013.

[68] Emanuel Schemider, Matthias Ziegler, Erik Danay, Luzi Beyer, and Markus Bühner. Is it really robust? Reinvestigating the robustness of ANOVA against violations of the normal distribution assumption. *European Journal of Research Methods for Behavioural and Social Sciences*, 6:147–151, 2010.

[69] Philip M. Sedgwick, Anne Hammer, Ulrik Schiøler Kesmodel, and Lars Henning Pedersen. Current controversies: Null hypothesis significance testing. *Acta Obstetricia et Gynecologica Scandinavica*, 101(6):624–627, 2022.

[70] Jonah Stunt, Leonie van Grootel, Lex Bouter, David Trafimow, Trynke Hoekstra, and Michiel de Boer. Why we habitually engage in null-hypothesis significance testing: A qualitative study. *Plos one*, 16(10):e0258330, 2021.

[71] Denes Szucs and John P.A. Ioannidis. When null hypothesis significance testing is unsuitable for research: a reassessment. *Frontiers in human neuroscience*, 11:390, 2017.

[72] Jenny Tang, Lujo Bauer, and Nicolas Christin. Misuse, misreporting, misinterpretation of statistical methods in usable privacy and security papers: List of papers. https://kilthub.cmu.edu/, 2025. DOI: 10.1184/R1/29260508.

[73] Jenny Tang, Lujo Bauer, and Nicolas Christin. Misuse, misreporting, misinterpretation of statistical methods in usable privacy and security papers: Tests and statistics considered. https://kilthub.cmu.edu/, 2025. DOI: 10.1184/R1/29275457.

[74] Christopher Tong. Statistical inference enables bad science; statistical thinking enables good science. *The American Statistician*, 73(sup1):246–261, 2019.

[75] David Trafimow and Michael Marks. Editorial. *Basic and Applied Social Psychology*, 37(1):1–2, 2015.

[76] Mark van der Laan, Jiann-Ping Hsu, Karl E. Peace, and Sherri Rose. Statistics ready for a revolution: Next generation of statisticians must build tools for massive data sets. *AMSTAT news: the membership magazine of the American Statistical Association*, (399):38–39, 2010.

[77] Jan B. Vornhagen, April Tyack, and Elisa D. Mekler. Statistical significance testing at CHI play: Challenges and opportunities for more transparency. In *Proceedings of the annual symposium on computer-human interaction in play*, pages 4–18, 2020.

[78] Chat Wacharamanotham, Lukas Eisenring, Steve Haroz, and Florian Echtler. Transparency of CHI research artifacts: Results of a self-reported survey. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI 2020, page 1–14, 2020.

[79] Larry Wasserman. *All of statistics: a concise course in statistical inference*. Springer Science & Business Media, 2013.

[80] Ronald L. Wasserstein and Nicole A. Lazar. The ASA statement on p-values: context, process, and purpose. *The American Statistician*, 70(2):129–133, 2016.

[81] Ronald L. Wasserstein, Allen L. Schirm, and Nicole A. Lazar. Moving to a world beyond "p< 0.05". *The American Statistician*, 73(sup1):1–19, 2019.

[82] Miranda Wei, Jaron Mink, Yael Eiger, Tadayoshi Kohno, Elissa M. Redmiles, and Franziska Roesner. SoK (or SoLK?): On the quantitative study of sociodemographic factors and computer security behaviors. In *33rd USENIX Security Symposium (USENIX Security 24)*, pages 7011–7030, August 2024.

[83] Jelte M. Wicherts, Coosje L.S. Veldkamp, Hilde E.M. Augusteijn, Marjan Bakker, Robbie C.M. Van Aert, and Marcel A.L.M. Van Assen. Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in psychology*, 7:1832, 2016.

[84] Wolfgang Wiedermann and Alexander von Eye. Robustness and power of the parametric t test and the nonparametric Wilcoxon test under non-independence of observations. *Psychological Test and Assessment Modeling*, 55(1):39–61, 2013.

[85] Max L.L. Wilson, Paul Resnick, David Coyle, and Ed H. Chi. Replichi: the workshop. In *CHI'13 Extended Abstracts on Human Factors in Computing Systems*, pages 3159–3162. 2013.

[86] Frank Yates. The influence of statistical methods for research workers on the development of the science of statistics. *Journal of the American Statistical Association*, 46(253):19–34, 1951.

[87] Donald W. Zimmerman. A note on preliminary tests of equality of variances. *British Journal of Mathematical and Statistical Psychology*, 57(1):173–181, 2004.

A Codebook

A.1 General

Instance. Each entry is a instance (or set of instances) using the same statistical method and reported in the same way. A paper can have more than one set of instances present, and a set of instances can have more than one issue (of the same type or of different types). If a paper conducts the same statistical test multiple times, but uses different reporting formats, each distinct type of reporting is coded as a separate instance, even though the same statistical method was used.

Example: If a paper reports on the effect size and exact p-value for three two-sample t-test results, but then gives no information on the effect size or p-value for another result evaluated using a two-sample t-test, the first set of three results is categorized as one instance, and the result reported without an effect size or p-value is categorized as another instance. All results of a regression model are one instance (unless certain coefficients are interpreted differently.)

A.2 Statistical Test

The test used in the paper for each instance. See table in [73].

A.3 Statistical Issues

This code evaluates whether the choice of statistical test is suitable for the type of data tested [57].
Incorrect. Test does not properly account for dependencies between data. Expected dependencies listed in table in [73].

Examples:

- Used test suited for paired data for non-paired data.
- Used test that did not account for repeated measures when multiple datapoints were from the same unit.

Data Type Mismatch. The test is not well-suited for the data type. See table in [73] for expected data types.

Examples:

- Treating Likert data as continuous when it is a dependent variable.
- Sum (or other aggregation) of multiple Likert data points into one independent variable.

Unverifiable. There is not enough information reported to determine whether the statistical test is suitable.

Examples:

- Non-unique test statistic (e.g., t-statistic can be from a two-sample t-test, paired t-test, or a linear regression coefficient).
- Non-exact test name (e.g., "Wilcoxon test" can indicate Mann-Whitney-Wilcoxon test or Wilcoxon rank-sum test; "correlation" can indicate Pearson's correlation or Spearman's rank correlation).

  Exception: assume a two-sample t-test if text states "t-test" and assume a Pearson's χ^2 if text states "χ^2" in the text of the paper given the prevalence of these tests.
- Lack of information about the variables in the model.
- Unclear what predictors are included in a regression model.
- Unclear how ordinal variables are transformed into binary variables.

Other. Anything that does not seem to fall into the above categories (make a note and discuss).
No Issues. No statistical validity issues.

A.4 Reporting Issues

Codes under this category investigate whether sufficient context is provided to evaluate the validity, statistical significance, and scientific significance of the results from the statistical test. This assumes that the choice of statistical test is correct (whether or not it is) and only evaluates the reporting. There are multiple subcategories under "Reporting Issues."

A.4.1 statSigSufficient

Code whether the reporting includes the appropriate test statistic, degrees of freedom, and exact p-value. See table for acceptable test statistics.
Yes. Reports test statistic, degrees of freedom, and exact p-value (acceptable to say $p < 0.001$ for very small p-values, need exact value if > 0.001; so $p < 0.01$ is not acceptable) For regression coefficients, does not need anything except p-value threshold.

Examples:

- We found a statistically significant association between X and Y, $t(336) = 4.17, p < 0.001$

No. Missing any of test statistic, degrees of freedom, exact p-value.

Examples:

- We found a statistically significant association between X and Y, $p = 0.037$

A.4.2 sciSigReportingType

Check what type of reporting (if any) exists to allow a reader to evaluate the scientific significance of the results of the test.
Effect Size. One of the standard effect sizes (see table in [73]).

Counts. Counts or frequencies in each category (sufficient information to reconstruct counts is also acceptable, such as percentages when the total number is given).
Sufficient A sufficient reporting of context (see spreadsheet in [73] for what is acceptable for each test and whether a measure of variability is needed for specific measures).
Examples:

- Centrality with variability.
- Exact values usually sufficient (e.g., if dataset provided).
- Allow for reporting of underlying data in aggregate for paired tests (because otherwise too difficult to report).
- If a "sufficient" reporting of context can be constructed from visuals and figures (e.g., bar charts with exact counts).

No Variability Gives a point measurement as context without information on variability or uncertainty (e.g., coefficient without CI or SE, mean without SD); see table in [73].
Insufficient. Insufficient information on context.
Examples:

- Only gives coefficients for significant results.
- Distribution but not of all the groups tested (e.g. give frequency of 4–5 on Likert scale but the statistical test uses all 5 Likert buckets).
- Gives information that is by each variable (axes) rather than as a scatterplot (paired) for correlations.

Absent. No presentation of any effect size or context.

A.4.3 Reporting Issue

The high level evaluation of whether there was enough information reported to evaluate scientific significance or if there were other issues with the amount of information presented about the test.
statSig Reporting Issues. There is insufficient context to evaluate statistical significance. This code occurs when statSigSufficient is not "Yes."
sciSig Reporting Issues. There is insufficient context to evaluate scientific significance.
Examples:

- If sciSig is not any of Effect Size, Counts, or Sufficient then it is insufficient context.
- No information is reported beyond the model or test conducted.
- When there is no significant result from p-value (paper simply says something similar to "the result was not significant").

No Model Fit. For regressions only, no evaluation of model fit (R^2, pseudo-R^2 etc.). This falls under *sciSig Reporting Issues*.
Examples:

- No evaluation of model fit—evaluation is not done or reported at all, with no evaluation statistics.
- Insufficient evaluation of model fit—the paper discusses some model fitting but no statistics or quantitative results are given for model fit.

No Test or Model. No information provided about the statistical test or model to accurately evaluate the conclusions. This falls under *Statistical Validity Reporting Issues*.
Examples:

- Only gives a p-value or states that there is no significant difference.
- No unique test can be inferred from name or test statistic.

Unclear Model. Unclear what the IVs or DVs are. This falls under *Statistical Validity Reporting Issues*.
Examples:

- Unclear data aggregation, such as no information on how a variable is bucketed or transformed.
- Unclear IVs, Does not give information on what variables are in a model or whether interactions are present.
- Unclear baseline in regression for categorical variables that are used as independent variables.

Other. All other reporting issues (make a note and discuss).
No Issue. No issues with reporting.

A.5 Interpretation Issues

Codes for whether the conclusion or instance in the paper is in fact supported by the statistical results (or supported in the way described in a paper).
Incorrect Effect Interp. An interpretation that is directly contradicted by the results. This falls under *misinterpretation*.
Examples:

- Interpreting log odds as odds (different scales, may reverse direction of effect).
- Interpretation implies impact of DV on IV in regression (reverse direction of relation).

Improper statSig Interp. Conflates statistical significance with effect size or makes claims that p-values prove or disprove an assertion. This falls under *misinterpretation*.
Examples:

- Claims of highly significant differences or effects when the p-value is small, but the data itself is not presented (or not sufficiently presented).

- Assumes rejection of the null means the null is not true or assumes lack of rejection of the null means the null is true.

Misrepresentation. Model coefficients are not interpreted adequately in the context of the model or test results are interpreted on some other metric that does not match between different parts of the reporting. This falls under *misinterpretation*.

Examples:

- Not taking into account the baseline for categorical variables.
- Does not explain what baseline is set as for predicted probabilities in logistic regressions.
- Reporting results of interactions alone, without interpreting it as an additional effect to main effects.

Not Tested. Claiming a result that was not directly measured by the test. This falls under *misinterpretation*.

Examples:

- Use of omnibus test to claim a pairwise difference.
- Reporting an interaction without testing it or putting it in the model.

Incorrect scale. Interpretations of results are on an incorrect scale but the correct direction. This falls under *sub-optimal effect interpretation*.

Examples:

- Interpreting log odds changes as linear increases or as changes in DV by one point (correct direction, wrong magnitude).
- Interpreting linear regression coefficients as changes in likelihood (usually correct direction, wrong magnitude).

Sub-optimal or no scale. Lack of determination of magnitude in interpretation or interpretation on an unintuitive scale. This falls under *sub-optimal effect interpretation*.

Examples:

- Only gives vague directional interpretations (e.g., positive or negative) when coefficients are presented in a table with no sense of magnitude.
- Interprets on log scale (for logistic or ordinal logistic regressions) rather than transformed (i.e., scale is correct but unintuitive).

Interprets distributions. For regressions only, interpreting in terms of the dataset rather than result of the model. This falls under *sub-optimal effect interpretation*.

Examples:

- Interprets using only underlying data (distributions of responses) and not regression result (impact of IV on DV).
- Logistic regression reported as response distributions or rates rather than likelihoods.

No Interpretation. No in-text interpretation of the results, only reports on presence of statistical significance for tests that also evaluate magnitudes.

Examples:

- Reports all tests in a table or figure without talking about any results from tests in the text.
- Indicating significant coefficients in a table without any explanation (e.g., size, direction etc).

No Issue. No issues with interpretation. Note, this code assumes that the test is valid as reported. In other words, if the test can be used to claim X (e.g., for a two-sample test, presence or absence of a statistically significant difference), it falls under this code, *even if* the test is not suited for the data.

Examples:

- Test "did not find a statistically significant result."
- Test "found a difference between these groups."