Numeracy
Advancing Education in Quantitative Literacy

Volume 16 | Issue 1 Article 6

2023

Beyond Statistical Significance: A Holistic View of What Makes a Research Finding "Important"

Jane E. Miller
Rutgers, The State University of New Jersey, jem@rutgers.edu

Recommended Citation

Miller, Jane E.. "Beyond Statistical Significance: A Holistic View of What Makes a Research Finding "Important"." *Numeracy* 16, Iss. 1 (2023): Article 6. DOI: https://doi.org/10.5038/1936-4660.16.1.1428

Beyond Statistical Significance: A Holistic View of What Makes a Research Finding "Important"

Abstract

Students often believe that statistical significance is the *only* determinant of whether a quantitative result is “important.” In this paper, I review traditional null hypothesis statistical testing to identify what questions inferential statistics can and cannot answer, including statistical significance, effect size and direction, causality, generalizability, and changeability of the independent variable. I illustrate these issues with examples from an empirical study of the association between how much time teenagers spent playing video games and time spent reading. I describe how study design and context determine each of those aspects of “importance,” and close by summarizing how to provide a holistic view of importance when writing about a quantitative analysis. I also include exercises to guide students through applying these concepts to articles in newspapers and scholarly journals.

Keywords

inferential statistics, substantive significance, causality, confounding, generalizability

Cover Page Footnote

Jane E. Miller is a Professor at the Edward J. Bloustein School of Planning and Public Policy at Rutgers University, where she specializes in research communication, numeracy, and quantitative literacy. Her latest book is *Making Sense of Numbers: Quantitative Reasoning for Social Research* (Sage). Her two previous books *The Chicago Guide to Writing about Numbers* and *The Chicago Guide to Writing about Multivariate Analysis* are both in their second editions and are also available in Chinese translation. She has also authored a series of articles in teaching and research journals about on how to communicate about quantitative research. She earned her B.A. in economics from Williams College and her Ph.D. in population studies from the University of Pennsylvania.

Introduction[1]

Students who have recently learned about inferential statistics often believe that statistical significance is the *only* determinant of whether a result is "important." Although it is a necessary part of conveying the uncertainty associated with results based on a sample instead of an entire population (Utts and Heckard 2014), statistical significance alone is *not* a sufficient basis for assessing the importance of a numeric estimate (Thompson 2004; Ziliak and McCloskey 2008; McShane et al. 2019; Miller 2021).

Contributing to the confusion around the meaning of statistical significance is that researchers, journalists, and others frequently are not very precise in their use of the term "significant" or "significance," failing to specify whether they mean *statistical* significance or *substantive* (practical) significance (Daza 2020; Miller 2021). Both are valuable aspects of the importance of a numeric result, and each provides a different perspective on what that finding means and how it can be applied to address a real-world question.

As a consequence, in 2019 *The American Statistician* published an entire Supplement titled "Statistical Inference in the 21st Century: A World Beyond *'p* < 0.05'" (Wasserstein et al. 2019) to address the pitfalls of that approach to assessing statistical significance, and to suggest ways to improve communication about the "significance" of results. Also in that vein, top peer-reviewed journals are beginning to strengthen requirements for reporting and interpreting effect size as well as statistical significance (Amrhein et al. 2019).

An essential further point, however, is that assessing the practical meaning of a research finding involves several criteria beyond effect size and statistical significance, including factors that affect the extent to which that finding can be applied to understanding and generating solutions to real-world problems.

To distinguish between statistical significance and other factors that determine the importance of a numeric finding, I examine what questions inferential statistics can and cannot answer. I illustrate these points with examples from a study of the association between the amount of time teenagers spent playing video games and how much time they spent on other activities—a topic that many students will find relatable and easy to grasp.

Brief Review of Statistical Significance Testing

To lay the groundwork for a discussion of what questions statistical significance can answer, here is a brief overview of the purpose and steps involved in conducting

[1] An earlier version of this paper was presented as the opening keynote address at the annual meetings of the National Numeracy Network on March 4, 2022.

and interpreting an inferential statistical test. For a more thorough treatment of these concepts, see a standard statistics textbook such as Utts and Heckard (2014) or Salkind (2016).

Threshold-based Null Hypothesis Testing

To acknowledge the uncertainty associated with numeric estimates based on samples, the field of statistics developed a set of procedures called "hypothesis testing," also known as "statistical significance testing." In the traditional null hypothesis significance testing (NHST) paradigm (McShane et al. 2019), the first step is to state the research question as a hypothesis, putting it into a form that can be tested using statistical methods. Two forms of the hypothesis are written: A null hypothesis (denoted H_0) and a research (or alternative) hypothesis (denoted H_A). When investigating an association between an independent variable (IV) and a dependent variable (DV), often the null hypothesis posits no relationship between the IV and DV, whereas the research hypothesis predicts a non-zero relationship between those two variables.

Next, a sample statistic, such as a measure of association, is calculated from the sample data, along with the standard error of that estimate—the amount of variation in the sample statistic based on different samples drawn from the same population. A *p*-value is then obtained, representing the probability of obtaining a result at least as large as the sample statistic if there is *no association* (a null value) between those variables in the population. Researchers then select a significance level (α), the probability of *incorrectly* concluding that the true population value is *not* equal to the null value, based on the estimate from the sample. A significance level of 0.05 is the conventional cutoff for assessing statistical significance.

Using the traditional "threshold approach" (McShane et al. 2019), if the *p*-value is less than the significance level (threshold), the result is termed "statistically significant." A statistically significant result is one that has a very low probability (*p*-value) that the sample estimate could be as large as it is solely due to chance associated with how the sample was drawn.

A different way of conveying the uncertainty around a numeric estimate based on a sample involves calculating a confidence interval, which is the estimated range of values within which the true population value falls, with the degree of confidence specified by the confidence level: confidence level = (1 – significance level) × 100. For instance, a 95% confidence interval (95% CI) is associated with a significance level (α) of 0.05. A CI is computed from the sample statistic and its standard error; a larger standard error results in a wider confidence interval.[2]

[2] Confidence interval = sample statistic ± (critical value × standard error of the estimate). The critical value depends on the nature of the statistical test, sample size, and the confidence level selected by the researcher. See Utts and Heckard (2014) or another statistics textbook for more on critical values.

DOI: https://doi.org/10.5038/1936-4660.16.1.1428

Under the NHST approach, assessing statistical significance using a confidence interval involves comparing it to the null value. If the CI around the point estimate (sample statistic) does *not* encompass the null value, the result is said to be statistically significant.

Both NHST approaches to assessing statistical significance contribute to a misleading picture of empirical patterns in the overall body of literature on a topic. This occurs because for a given-sized standard error, the confidence interval around a high point estimate is less likely to include the null [no difference] value than the CI for a lower point estimate, therefore larger effect sizes are more likely than smaller ones to be deemed statistically significant. In combination with the historical bias in favor of statistically significant results, the use of a NHST approach to drawing inferences about statistical significance leads to publication bias: an *upward bias* in the effect sizes that are reported in the published literature (Amrhein et al. 2019).

Non-threshold Interpretation of an Inferential Statistical Result

An alternative approach that averts some of the drawbacks of the dichotomous (statistically significant or not) approach involves interpreting measures of uncertainty of statistical estimates in continuous fashion instead of comparing them against a cutoff like $p<0.05$ or evaluating whether the CI overlaps the null value (McShane et al. 2019).

In line with that approach, Amrhein et al. (2019) recommend replacing the term "confidence interval" with "*compatibility* interval" to reflect the fact that any of the values within that interval are compatible (consistent) with the data used to calculate the estimate, and therefore, focusing on just one particular value in that interval can be misleading.[3]

What Questions Can Statistical Significance Answer?

Building on what we learned above, we see that statistical significance answers a very specific question: "How likely would it be to obtain an estimate at least as large as the one based on the *sample* if the true value of that statistic was the null value in the *population* from which that sample was drawn?" The *p*-value measures the probability of incorrectly concluding, based on the sample statistic, that the true population parameter equals the null value, so we want *p* to be as small as possible.

[3] However, the point estimate is the most compatible with the sample data, and values near the point estimate are more compatible than those near the upper or lower limits of the compatibility interval (Amrhein et al. 2019).

Example: Based on an analysis of observational data from the Panel Study of Income Dynamics (PSID)–a large representative sample of teenagers in the United States–Cummings and Vandewater (2007) found that for every hour boys played video games, they read on average *just two minutes less* ($p < 0.01$).

Comments: $p < 0.01$ simply means that the 99% confidence interval around the point estimate of the effect of gaming time on reading time did not include the null value (0, indicating no difference between groups; left-most bar in Figure 1: 99% CI: -1 to -4 minutes). However, because of the large sample used in the gaming study (n = 425), the standard error of that estimate was very small, meaning that even the trivially small effect reached conventional levels of statistical significance.

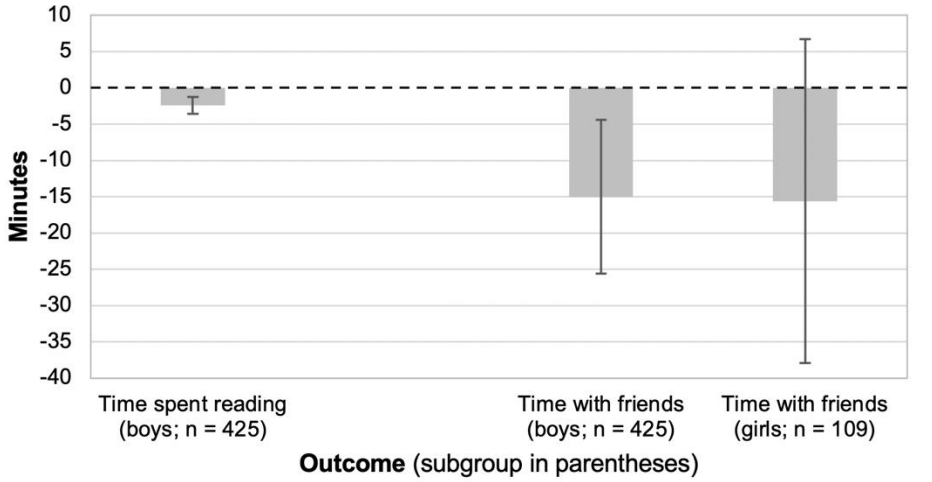

Figure 1. Estimated change in outcome for each additional hour spent playing video games, US teens, 2002–2003. Calculated from estimates provided in Cummings and Vandewater (2007) using data from the 2002–2003 US Panel Study of Income Dynamics (PSID). "Whiskers" indicate 95% confidence intervals unless otherwise noted. 99% CI: for reading time (boys): -1 to -4 minutes. "Time with friends" refers to non-video gaming activities.

Example: The same study found that each hour teenagers spent playing video games was associated with a 15-minute decrease in time spent with friends in activities other than video games. Although that association *was* statistically significant for boys, the same size effect was *not* statistically significant for girls (two right-hand bars in Figure 1).

Comments: The 95% confidence interval for boys did not include the "no difference" value (0; 95% CI: -26 to -4 minutes), therefore that result was considered statistically significant. The 95% CI for girls was –38 to +7 minutes, thus including the null value, which authors using a threshold-based approach often interpret simply as "not statistically significant," overlooking a substantial part of the picture.

DOI: https://doi.org/10.5038/1936-4660.16.1.1428

The statistical significance (or lack thereof) of those two results merely tells us the probability that the empirical associations observed in the sample could have arisen based solely on chance in how those samples were drawn from the population. That is important information, but it is not the *only* thing needed to evaluate the "importance" of the numeric results.

Example: A substantial portion of the confidence interval for girls (from –38 to 0) is compatible with the conclusion that, as for boys, time girls spent playing video games was inversely associated with time spent with friends. Moreover, the wider confidence interval for girls arose because of a larger standard error for girls than for boys, reflecting at least in part that in the PSID sample, far fewer girls (n = 109) than boys were gamers (n = 425).

Comments: Using the non-threshold approach provides a much more nuanced interpretation of the same set of inferential statistical information, incorporating both the size and direction of the associations between gender, gaming, and time with friends. By pointing out a reason for the wider confidence interval around the girls' estimate, the explanation provides valuable contextual information for interpreting that statistical evidence.

What Questions Can't Statistical Significance Answer?

Although information about the degree of uncertainty of a numeric estimate is an expected part of analysis of data from a random sample, inferential test results *cannot* answer several other equally valuable questions for gauging the importance of the results. These include questions about factors affecting the practical meaning of those results, including their substantive significance and applicability (Daza, 2020; McShane et al., 2019).

Substantive Significance

Substantive significance encompasses both the size and direction of a numeric result.

Size of the association. The first question that inferential statistics cannot answer concerns whether the effect is substantively significant. In other words, is it big enough to matter for that topic and context (Daza 2020; Miller 2021)? Put differently, substantive significance pertains to whether the result is sufficiently large to be meaningful educationally, politically, clinically, or in whatever domain the topic fits.

Example: As noted previously, teenage boys read *just two more minutes* for each one-hour reduction in time spent playing video games. Is that a big

enough difference to attract serious interest from parents or teachers who seek to increase boys' reading time?

Comments: Would two additional minutes of reading per day appreciably improve boys' cognitive function or enjoyment of literature? It is difficult to imagine that an intervention to decrease time spent playing video games would be a worthwhile approach to increasing reading time, based on such a small effect.

Standard deviations and other empirical measures of distribution are useful benchmarks for assessing whether an observed change in the dependent variable is big enough to matter. The same size effect is considered more substantively important if it is equivalent to a substantial share of a standard deviation of the dependent variable than if it corresponds to only a small fraction of a standard deviation (Miller 2021).

Example: In Cummings and Vandewater (2007), the standard deviation for reading time was 23 minutes, therefore the two-minute increase associated with a one-hour reduction in gaming time is trivially small.

Comments: For reading time, the observed difference (two minutes) is less than one-tenth of a standard deviation (2 ÷23 = 8.6%), which is too small of a difference to be of substantive importance for that topic and context.

Direction of the association. Second, results of inferential statistics often do not address whether that association is in the *expected direction*. The dependent variable could show statistically significant variation with the independent variable, but *opposite* from the hypothesized direction—a very important point about that numeric result!

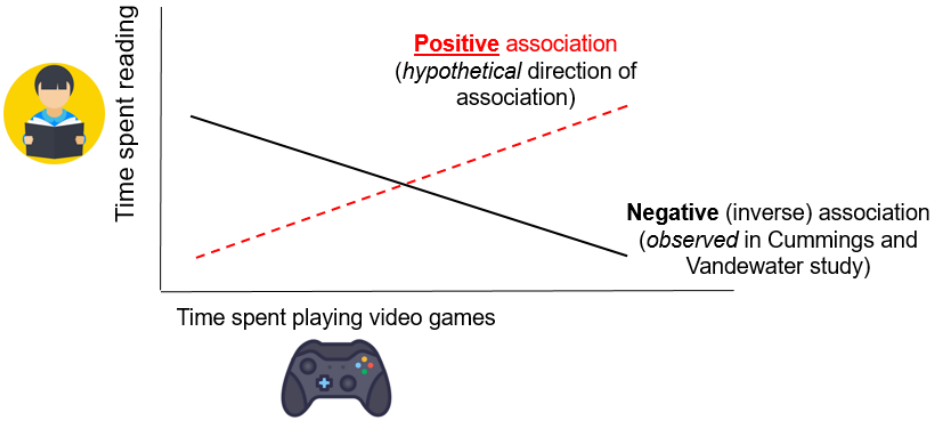

Figure 2. Illustration of positive and negative associations.

Example: What if Cummings and Vandewater (2007) had found that—contrary to what was expected—less video-game-playing time was statistically significantly associated with *less* reading time (red dashed line in Figure 2)?

Comments: Instead of the expected inverse association, this hypothetical situation found a positive association. If the goal is to increase reading time, we certainly wouldn't want to encourage teenagers to cut back on playing video games because doing so would also reduce the amount of time they spent reading. Thus, the sign (direction) of the association is an essential piece of information about its substantive significance.

If a researcher poses a *directional* hypothesis and reports the result of a 1-sided test, then statistical significance *does* shed light on whether the association was in the hypothesized direction. For instance, if Cummings and Vandewater (2007) had reported and interpreted the *p*-value for a one-sided test, that information *would* answer whether the observed association was inverse as predicted. Thus, when interpreting results of a test statistic, it is important to determine whether researchers reported result of a two-sided (*non-directional)* or one-sided (directional) test. However, the default setting in most software is a two-sided test, so that is what most researchers report unless they have deliberately changed that setting to specify a directional test and interpreted the *p*-value accordingly.

Applicability of Results

Another set of questions that inferential statistics cannot answer pertain to whether and how the results can be translated and applied beyond the set of cases included in the study sample. These issues include whether an association can be interpreted as cause-and-effect, whether an intervention can be designed to manipulate the independent variable, and the generalizability of the results, all of which determine the practical implications of those findings for addressing social, health, or other real-world questions.

Whether the association is causal. Statistical significance of an association between two variables does *not* tell us whether that relationship can be interpreted as cause and effect. As students learn in introductory statistics courses, "association does not equal causation" or "correlation does not necessarily imply causation." In other words, the fact that there is statistically significant variation in the values of one variable according to values of some other variable does *not*, *by itself*, prove that the independent variable actually caused a change in the dependent variable.

We say that a relationship between an independent variable (*x*) and a dependent variable (*y*) is causal if changing the values of *x* leads to a change in the values of *y*, all else being equal. In other words, if a relationship is cause-and-effect, then

altering the variable that we hypothesize is the *cause* will produce a response in values of the variable that we think is the *effect* (Miller 2021).

Investigating whether an association between an independent and a dependent variable can be interpreted as causal is a crucial step for any study whose results are intended to inform decisions, programs, or policies to affect that outcome.

There are three main criteria for assessing causality: (1) the presence of an empirical association; (2) time order; and (3) non-spuriousness (Schneider and Lilienfeld, 2015). Evidence for an empirical association pertains to its size, direction, and statistical significance, which are discussed above.

Causal order of independent and dependent variables. Statistical significance *cannot* be used to determine the causal order of the variables *x* and *y*: whether what we think is the cause is actually the effect, otherwise known as "reverse causation" or the "cart before the horse" problem.

Example: What if boys who became more interested in reading cut back on video-game-playing to make more time for their reading? That would also produce an inverse association between reading and playing video games, but with the *opposite* of our hypothesized causal order (see Fig. 3).

Comments: If video-game-playing time depends on reading time, then what we thought was the independent variable (cause) is really the dependent variable (effect). If reverse causation is occurring, then intervening to decrease what we believed was the cause (video-game-playing time) would not have the desired positive impact on what we thought was the effect (reading time).

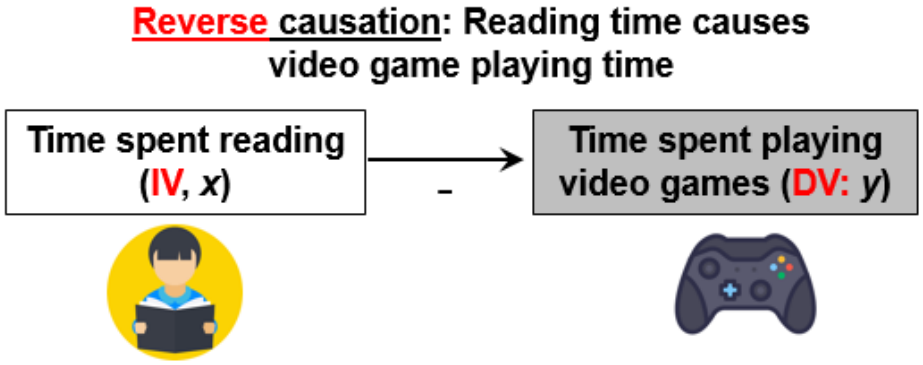

Figure 3. Illustration of reverse causation.

Whether the association is non-spurious. Statistical significance of a bivariate association cannot shed light on whether an observed association is non-spurious. Sometimes a third variable (z) explains an observed association between *x* and *y*, in which case we say that the association is "confounded" by the third variable. Confounding is also known as the "possibility of alternative explanations" (Michael

et al. 1984) because *something other than* the hypothesized independent variable (*x*) is the true cause of variation in the dependent variable (*y*). In such situations, the association between *x* and *y* is termed "spurious." If an observed association is spurious, then intervening to change what we thought was the cause will not result in the desired effect.

Example: Would getting boys to cut down on playing video games *cause* them to increase the time they spend reading? Suppose that further investigation revealed that an increase in time spent participating in drama club or school government was associated with *more* time reading (e.g., documents such as scripts of plays or policy documents) but *less* time playing video games (lower part of Figure 4).

Comments: In this hypothetical situation, video-game-playing time (x) is not a real cause of reading time (y); that association is confounded by time spent on those extracurricular activities (z). In other words, the observed inverse association between video-game-playing time (the hypothesized independent variable) and reading time (the dependent variable) is spurious—explained entirely by the association of each of those variables with a third factor (e.g., time spent on drama club or student government). As a consequence, inducing boys to reduce their video-game-playing time would not yield the desired increase in reading time. A better intervention to increase reading time might be to encourage boys to spend more time on those extracurricular activities (the true cause).

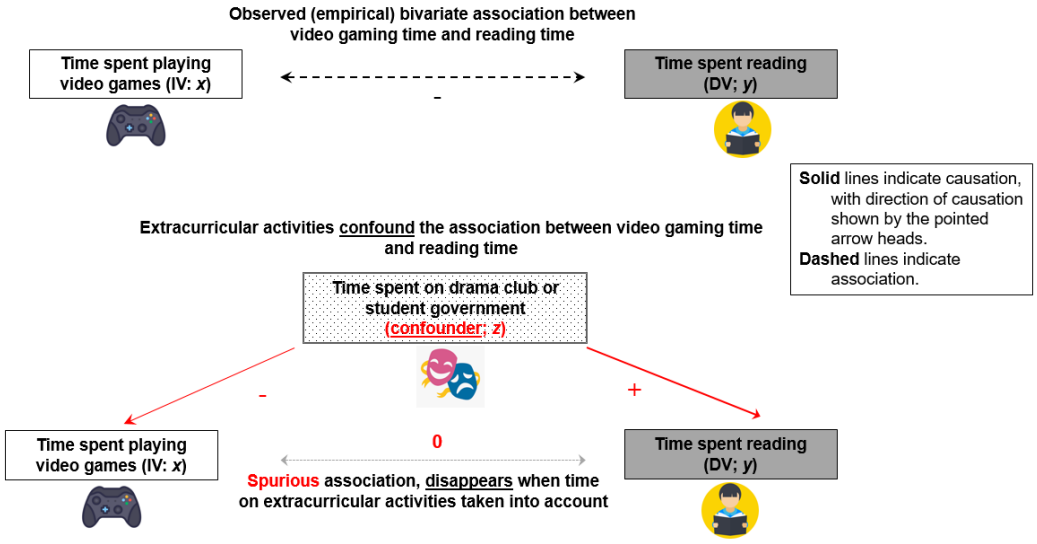

Figure 4. Illustration of confounding.

Modifiability of the Independent Variable

Another consideration that affects how results can be applied is whether the independent variable (sometimes known as the "risk factor") is modifiable—another question that statistical significance cannot answer. If the risk factor cannot be changed, then even an effect that is statistically significant, big enough to matter, in the expected direction, and causal, is *not* a good basis for an intervention to change the dependent variable. Identifying the degree to which a particular characteristic is malleable requires information about the nature of that trait in its real-world context.

Example: How easy is it to get teenage boys to substantially cut back on playing video games and to sustain that change?

Comments: As anyone familiar with adolescents is probably painfully aware, prying a gaming device out of the hands of a teenage boy for an hour every day will probably require so much effort on the part of his parents and generate so much conflict that such a strategy is not likely to succeed in the long run. Better to find a different cause of increasing reading time that is easier to change and maintain.

Generalizability

Finally, statistical significance does *not* tell us about the extent to which we can generalize (apply) the findings of the study beyond the set of cases that were used to produce the estimates. There are two types of generalizability: *Sample* generalizability refers to the ability to apply results based on a sample to the larger population from which that sample was drawn. *Cross-population* generalizability refers to the ability to apply conclusions based on a sample to a population that has *different* characteristics than the one from which the sample was drawn (Chambliss and Schutt 2019).

Example: The Panel Study of Income Dynamics (PSID) data used for the analysis of video-game-playing and reading time were from a nationally-representative sample of 10- to 19-year-olds in the United States from 2002 to 2003 (Cummings and Vandewater 2007).

Comments: The PSID sample has good sample generalizability, meaning that we can have high confidence about extrapolating the findings to all US teenagers in those years (see Fig. 5a). The extent of cross-population generalizability, such as to other age groups or countries (Fig. 5b), depends on whether those other populations had different distributions of video-game-playing and reading, or factors influencing the relationship between those variables.

DOI: https://doi.org/10.5038/1936-4660.16.1.1428

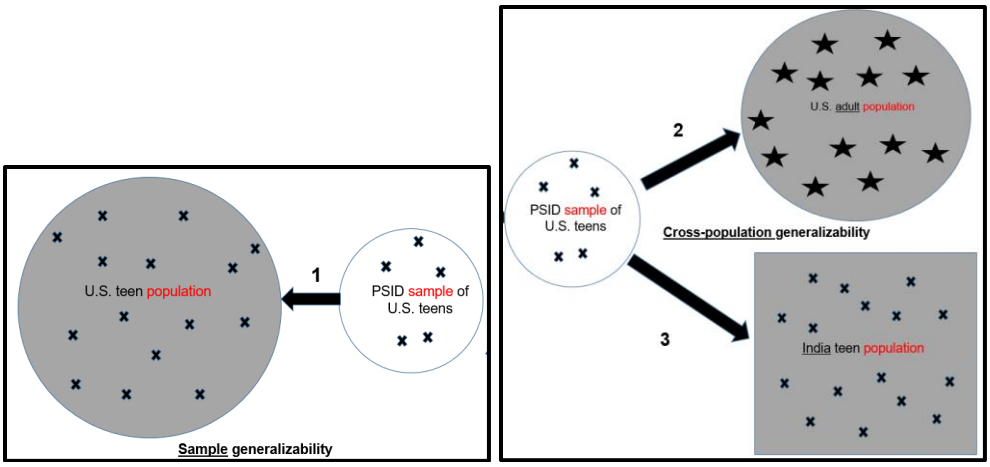

Figure 5. (a) Sample generalizability; and (b) cross-population generalizability. PSID: Panel Study of Income Dynamics.

How Study Design Affects Elements of "Importance"

Different aspects of context and study design affect whether an association can be interpreted as cause-and-effect, the extent to which results can be generalized, and statistical significance of those findings. That means that even if one of those conditions for "importance" is satisfied for a particular study, one or more of the other conditions might *not* be. Most of the criteria for evaluating those facets of importance are determined by aspects of study design, including which methods were used to select the cases, whether the study is cross-sectional or longitudinal, whether it is observational or experimental, and sample size (Miller 2021). Daza (2021) refers to factors that affect the quality of the evidence from a study as affecting its *scientific* (as distinct from *statistical*) significance.

Causal Inference

The suitability of a study for evaluating whether an observed association between an independent and a dependent variable can be interpreted as cause-and-effect is referred to as the "internal validity" of that study. It concerns the extent to which a study satisfies each of the criteria for assessing causality: empirical association, time order, and non-spuriousness.

"True experiments" (also known as "randomized controlled trials") have higher internal validity than observational (*non*-experimental) studies because they include design features to address those criteria (Chambliss and Schutt 2019). Many experimental studies are longitudinal, making them better than cross-sectional ones at establishing time order because they measure whether changes in the independent

variable preceded changes in the dependent variable, thus determining whether reverse causation can be ruled out. However, time order *alone* cannot establish causality—a reasoning error known as the post-hoc fallacy (Nordquist 2020).

Experimental studies are better than observational ones for establishing non-spuriousness because random assignment of cases into treatment and control group is intended to equalize potential confounders between those groups (Chambliss and Schutt 2019). In studies based on observational data, measuring and analyzing potential confounders can strengthen internal validity (Miller 2021).

Sample size does *not* influence ability to determine time order or non-spuriousness—two of the criteria for assessing internal validity.

Representativeness and Generalizability

Whether and to which other populations the results of a study can be generalized is known as its "external validity." A key criterion for assessing external validity is whether the distributions of the variables of interest are similar in the sample and target populations. In other words, the extent to which the sample is "representative" of the population to whom the findings are to be applied (Chambliss and Schutt 2019). The representativeness of a study sample is affected by several aspects of study design.

The *sample* generalizability of a study's results depends on how the set of cases was obtained, which affects the degree to which the sample is representative of the desired target population. Samples selected using probability (random sampling) methods are much better for obtaining representative samples than studies that use non-probability sampling methods. Random sampling methods are more commonly used for observational than for experimental studies, meaning that results of observational studies such as sample surveys often have better external validity than do experimental studies (Miller 2021).

Low response rates often reduce the external validity of a study because those who respond are typically very different from those who do not, resulting in an analytic sample that can have substantially different characteristics than the *intended* sample. It is also more difficult to determine external validity of samples drawn using non-probability methods because the amount and pattern of non-response cannot be determined.

The time structure of a study can affect external validity because samples from longitudinal studies may become less representative over time if cases with certain characteristics are more likely than others to drop out of the study. Such attrition does not affect cross-sectional studies (Chambliss and Schutt 2019).

Differences in context—when, where, and which cases were in the sample and in the target population—can substantially influence the degree to which findings can be generalized across populations (*cross-population* generalizability).

Contrary to what many people believe, sample size doesn't affect generalizability because it doesn't determine the representativeness of a sample (unless that sample is very small).

Statistical Significance

Inferential statistics assume that the sample is representative of the population in terms of the variables of interest. Samples drawn using probability sampling methods (random sampling) are more likely to produce representative samples than those using non-probability methods such as convenience sampling and quota sampling. Therefore, assessing statistical significance of patterns from data collected using *non*-probability sampling methods is problematic, and care should be taken in generalizing those conclusions (Utts and Heckard 2014).

Sample size is inversely related to the standard error of an estimate and thus also to the *p*-value and the width of the confidence interval associated with that estimate. As a consequence, results that are statistically significant based on a large sample might not be statistically significant if fewer cases had been included, and vice versa.

Neither the time structure nor whether a study is observational or experimental affect statistical significance of estimates.

Relationships among Facets of "Importance" of a Research Result

The fact that statistical significance, substantive significance, causality, and generalizability of a research finding are each affected by different facets of study design means that just because one of those facets of the "importance" of that result is satisfied does not guarantee that the other criteria for importance will also be met. Conversely, just because a study fares poorly on one of those facets does *not* necessarily mean that it will also fare poorly on the other elements of importance.

Statistical Significance and Other Facets of "Importance"

Consider the inferences that cannot be drawn based solely on statistical significance, regardless of whether a threshold or non-threshold approach was used to assess statistical significance:

- Statistical significance does *not* necessarily translate into substantive importance: the association between time spent playing video games and reading in Cummings and Vandewater (2007) was statistically significant but the effect was too small to represent a meaningful increase in reading time.
- Conversely, substantive importance does *not* ensure statistical significance:

a large effect might not be statistically significant, due to wide variation in the sample or a too small sample size. If the PSID sample had included more girls, the relationship between gaming and time with friends might have reached conventional level of statistical significance.

- In observational (*non*-experimental) studies, a statistically significant association does *not* necessarily imply causation: gaming time and reading time were correlated at the 99% confidence level in the PSID survey data, but that does *not* make gaming time a cause of reading time. In true experiments, where cases are randomized into treatment and control groups, however, the possibility of confounding is reduced, so statistically significant findings are typically interpreted as causal.
- Conversely, existence of a causal relationship does *not* guarantee statistical significance: random error could overwhelm a true causal effect if based on a very small sample.
- Inferential statistics assume a representative sample, but a representative sample does *not* ensure statistical significance, as with the lack of statistical significance of the association between time gaming and time with friends among girls based on the nationally-representative PSID data.

Consider, also, the relationships among the other dimensions of "importance."

Substantive Significance and Causality

- Evidence that an association is causal does *not* automatically mean that it is substantively significant: even if we have evidence of an inverse causal association between gaming time and reading time, that effect was so tiny as to be unworthy of effort to get teens to cut back on gaming time.
- Substantive importance (a "big effect") does *not* necessarily mean that an association is cause-and-effect, as with the relationship between gaming and time with friends.

Generalizability and Substantive Significance

- Substantive significance does not automatically translate into external validity: an estimate based on a convenience (non-random) sample cannot be generalized to a specifiable population.
- External validity does not guarantee substantive importance because an estimate based on a representative sample could be very small, as in the association between video game playing and reading time.

Facets of Applicability

- Internal validity does *not* guarantee external validity, and vice versa. Internal validity is determined by the time structure of the study that

collected the data, and whether it was experimental or observational; external validity by whether the sample is representative of the population (Chambliss and Schutt 2019). The PSID data used in the gaming study had high sample generalizability, but provided weak evidence for causality because neither time order nor non-spuriousness could be demonstrated.

    - **Caution**: presence of the word "random" in both terms means that students often conflate "random *sampling"* with "random *assignment,"* although those two aspects of study design are completely different and have distinct effects on how results can be interpreted. The former pertains to how cases were chosen for the study and affects its *external validity*; the latter determines how study participants are placed into groups to be compared, which affects *internal validity* (Utts and Heckard 2014; Miller 2021).

- Neither internal nor external validity determine the malleability of a risk factor.
    - Just because a result is based on a representative sample does *not* mean that it is possible to alter the risk factor, as illustrated by the challenges of reducing video game playing time among teenagers.
    - Simply because a risk factor has been shown to be a likely cause of the hypothesized outcome does *not* mean that that independent variable can be changed. Consider how difficult many people find it to lose weight and keep it off, even knowing that weight loss can substantially reduce their chances of diabetes, heart disease, or other serious illnesses.

Summary and Conclusions

In summary, the "importance" of a numeric answer to a real-world question is about much more than just *statistical* significance. It also relates to substantive importance, causality, generalizability, and whether an independent variable can be altered or serve as the basis for identifying a group to be targeted with an intervention, each of which is affected by different facets of study design and context. Unfortunately, if researchers don't specifically mention a *lack* of statistical significance, many readers will use the result as the basis for developing solutions to problems, even if the result doesn't meet one or more of the other criteria for "importance."

To provide a holistic understanding of the importance of a numeric result, researchers should present information about each of its aspects: The results section of a research paper should convey information about effect size, direction, and statistical significance. The discussion section should reiterate direction and size of the results in ways that convey the substantive importance for the topic and context,

and then consider factors affecting its practical importance, including internal validity and external validity of the study. For results that are intended to inform the design of an intervention, the discussion section should also address the extent to which the independent variable can be modified or used to target a subgroup of interest.

Finally, a great deal of misunderstanding about the "importance" of a numeric research finding can be averted if researchers present their results in ways that clearly distinguish between *substantive* significance and *statistical* significance. If you must use the term "significant," always precede it with a modifier—"statistical" or "substantive," or "practical." Alternatively, there are many synonyms for "significant" that more precisely identify which aspect of "importance" is being described (Miller 2015).

Practice Exercises

These exercises can be used as in-class group activities (working from an article provided to students) or assigned as homework for either individuals or groups.

1. In a local newspaper or magazine, find an article proposing a solution to a social problem based on results of a quantitative study. Find the original research paper being summarized by the newspaper or magazine article.
    a. Evaluate whether and how the article addresses each of these aspects of "importance":
        i. the substantive meaning (size and direction) of the results,
        ii. the statistical significance of the findings,
        iii. the internal validity of the study,
        iv. the external validity of the study,
        v. the extent to which the independent variable can be modified or used to focus on an at-risk group, and
        vi. whether the findings might differ by topic or context.
    b. Given your answers to those questions, write a short description of the appropriateness of the proposed solution, based solely on the results of that study.
2. In a journal article in your field, find an example of an association that is statistically significant based on the $p < 0.05$ convention.
    a. Evaluate whether the authors make it clear when they are discussing *statistical* significance.
    b. Consider whether the authors also discuss the practical meaning of the association and, if so, list which criteria they use to assess it.
    c. List any of the criteria for "importance" covered in this paper that the authors did *not* explicitly discuss.
    d. Investigate whether the article provides information that could be used to shed light on those aspects.
    e. Given your answers to those questions, discuss whether you agree with the authors' presentation of the overall "importance" of their findings.
3. Repeat the previous exercise, but for an example of an association that does *not* meet conventional criteria for statistical significance.

Acknowledgment

I would like to thank members of the National Numeracy Network, Stat Lit.org, and the Statistical Literacy section of the American Statistical Association for encouragement about my early work related to quantitative reasoning. I am also grateful to reviewers of *Making Sense of Numbers: Quantitative Reasoning for Social Research* and of earlier drafts of this paper for their feedback. Finally, I would like to acknowledge the contributions of the many students at Rutgers University who field-tested these ideas and instructional approaches.

References

Amrhein, Valentin, Sander Greenland, and Blakely McShane. 2019. "Retire Statistical Significance." *Nature*, 567, 305–307. https://doi.org/10.1038/d41586-019-00857-9

Chambliss, Daniel F., and Russell K. Schutt. 2019. *Making Sense of the Social World: Methods of Investigation (6th Ed.*). SAGE.

Cummings, Hope M., and Elizabeth A. Vandewater. 2007. "Relation of Adolescent Video Game Play to Time Spent in Other Activities." *Archives of Pediatrics and Adolescent Medicine,* 161(7), 684–689. https://doi.org/10.1001/archpedi.161.7.684

Daza, Eric J. 2020. "Confusing P-values with Clinical Impact: The Significance Fallacy." *Towards Data Science.* https://towardsdatascience.com/the-significance-fallacy-confusion-about-p-values-d7b5e530d0c

____. 2021. "Ditch 'Statistical Significance'—But Keep Statistical Evidence." *Towards Data Science*. https://towardsdatascience.com/ditch-statistical-significance-8b6532c175cb

McShane, Blakeley B., David Gal, Andrew Gelman, Christian Robert, and Jennifer L. Tackett. 2019. "Abandon Statistical Significance." *The American Statistician*, 73:sup1, 235–245. https://doi.org/10.1080/00031305.2018.1527253

Michael, Max, W. Thomas Boyce, and Allen J. Wilcox. 1984. *Biomedical Bestiary: An Epidemiologic Guide to Flaws and Fallacies in the Medical Literature*. Boston: Little, Brown & Company.

Miller, Jane E. 2015. *The Chicago Guide to Writing about Numbers (2nd Ed.)*. University of Chicago Press.

____. 2021. *Making Sense of Numbers: Quantitative Reasoning for Social Research.* New York: Sage Publications.

Nordquist, Richard. 2020. "What Is a Post Hoc Logical Fallacy?" ThoughtCo. https://www.thoughtco.com/post-hoc-fallacy-1691650 November 2020.

Salkind, Neil. J. 2016. *Statistics for People Who (Think They) Hate Statistics (6th Ed.).* SAGE.

Schneider, Dona, and David E. Lilienfeld. 2015. *Lilienfeld's Foundations of Epidemiology, Fourth Edition.* New York: Oxford University Press.

Thompson, Bruce. 2004. "The 'Significance' Crisis in Psychology and Education." *Journal of Socio-Economics*, 33, 607–613. https://doi.org/10.1016/j.socec.2004.09.034

Utts, Jessica, and Robert Heckard. 2014. *Mind on Statistics. 5th ed.* Independence, KY: Cengage, Brooks Cole.

Wasserstein, Ronald L., Allen L. Schirm, and Nicole A. Lazar. 2019. "Moving to a World Beyond '$p < 0.05$'." *The American Statistician*, 73:sup1, 1–19. https://doi.org/10.1080/00031305.2019.1583913

Ziliak, Stephen T., and Deidre N. McCloskey. 2008. *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives.* University of Michigan Press.