Journal of Research Practice
Volume 11, Issue 1, Article P1, 2015

ISSN: 1712-851X

Provocative Idea:

The Lay Public's Misinterpretation of the Meaning of 'Significant': A Call for Simple yet Significant Changes in Scientific Reporting

Philip Tromovitch
Science and Engineering Research Institute
Doshisha University
1-3 Tatara Miyakodani, Kyotanabe-shi
Kyoto-fu, 610-0394, JAPAN
ptromovi@mail.doshisha.ac.jp

Abstract

A large national sample of U.S. adults ($n > 1{,}000$) was queried regarding their interpretation of the term *significant* when used in regard to scientific findings. The vast majority provided incorrect interpretations of the meaning with only 5.8% providing a reasonably correct interpretation. Most respondents who reported they hold doctoral degrees provided incorrect interpretations. Given the widespread misinterpretation of this term, scientific journals should require—not merely recommend—that all usages of the term *significant* be prefaced with an adjective (e.g., statistically, practically, clinically) and that the meaning of statistical significance be reviewed prior to its first usage. Additionally, all claims regarding the size of a finding should be required to be supported with appropriate effect size statistics to ensure that statistical significance is not misrepresented as indicating practical significance.

Index Terms: scientific writing; research reporting; communication; statistical significance

Suggested Citation: Tromovitch, P. (2015). The lay public's misinterpretation of the meaning of 'significant': A call for simple yet significant changes in scientific reporting. *Journal of Research Practice*, *11*(1), Article P1. Retrieved from http://jrp.icaap.org/index.php/jrp/article/view/477/411

1. The Issue

The present author embarked on a series of multinational data collections to assemble data for a variety of scientific investigations primarily focused on psychological constructs and sexual variables, but also including secondary data on other potentially important areas of interest to the author. The project is titled *The Multinational Life Experience and Personality Project* (MLEPP). The analyses presented here focus on one of the secondary goals of the project: to investigate the lay public's interpretation or misinterpretation of the meaning of the term*significant* in describing scientific and statistical findings. This report is limited to the U.S. sample collected from April 2014 through August 2014 because the relevant data were only collected from this sample.

In most areas of science the goal is to understand something through measurement, and the method is to use measurements based on samples. For example, if one wanted to measure the difference in coughing frequency between smokers and non-smokers, one would not collect data from all humans on earth (i.e., the *population* of smokers and the *population* of non-smokers), instead, one would collect data from a *sample* of smokers and a *sample* of non-smokers and compare the data from the samples. Consequently, it is important to be able to assess the likelihood that the differences one finds between samples, in fact are likely to exist between the larger populations from which the samples were drawn. For a finding of any given size, it is possible to calculate the probability of obtaining that finding (or a larger one) purely due to chance.

A statistically significant finding is a measurement that a researcher is declaring to be, within some probability of error (usually 5 percent in the social sciences on a per analysis basis), different than would be expected by chance variation alone (i.e., there is only a 5 percent chance of obtaining a finding of that size or larger purely by chance if there is no difference in the populations from which the data were sampled). In contrast to this, the basic English definition of the term *significant* is focused on a meaning of importance, often with the implication that the thing being described is large. For example, the *New Oxford American Dictionary*'s first definition of the term is: "sufficiently great or important to be worthy of attention; noteworthy" (Oxford University Press, 2012). Thus, a significant finding may in fact be insignificant. Consider the following examples.

Imagine that a researcher hypothesizes that when people enter an auditorium or stadium where there is an equal amount of seating on both sides of a central entrance, that people with higher IQs tend to sit on one side and people with lower IQs tend to sit on the other. This researcher goes to a suitable auditorium and administers the Welsher IQ test to all of the attendees and finds a 0.5 point average IQ difference: $n_{right\text{-}side} = 1{,}000$, $n_{left\text{-}side} = 1{,}000$, $t = 0.75$, $p > .45$ and concludes that the finding is not significant (i.e., finding a half point IQ difference between two $N = 1{,}000$ samples is not particularly unlikely if there is no IQ-seating effect). This same researcher then performs a replication at a large sports stadium and again finds a 0.5 point average IQ difference: $n_{right\text{-}side} = 10{,}000$, $n_{left\text{-}side} = 10{,}000$, $t = 2.357$, $p < .05$ and declares the finding to be "significant." Clearly, the importance of such a small difference in IQ should not be called significant in its general English connotation. The finding in both studies shows the same thing, a half-point IQ difference, it simply took a very large sample to demonstrate that this very small difference was probably not due to chance alone.

Thus, if scientists, researchers, professors, and journalists writing about scientific findings report something as *significant*, the public may misinterpret the meaning, wrongly believing that the finding is large or important based on their familiarity with the general English usage of the term; alternatively, they may suspect the term has a different meaning in science and mistakenly guess at that meaning (e.g., since science often deals with precise measurement, they might mistakenly guess that the term *significant* in science refers to accuracy or precision of measurement).

As part of the 2014 U.S. data collection for the MLEPP, participants who indicated that English was their native or first language were asked one of three questions in order to determine the public's interpretation or misinterpretation of the meaning of the term *significant* when used by scientists, or when used in reference to scientific findings.

2. Methods

2.1. Data Collection and Participant Recruitment

The questionnaire was developed using a web-based system that allows the creation of complex computer-based surveys where the researcher can control which questions are displayed to which respondents (e.g., only display a particular question if the answer to a prior question was a given value), response randomization (i.e., randomize the order of the response-choices for a multiple-choice type question to control for order effects within a question), as well as more complex situations (e.g., display only one of a set of three questions to any given respondent).

National quota sampling was employed to create a U.S. national sample of adults aged 18-59 with approximately equal numbers of males and females and good coverage of the 18-59 year-old age range. The quota sampling appears to have succeeded in forming a national sample with good representativeness and coverage across numerous demographic variables (Tromovitch, 2015). A data quality demerit-based rating system was created to rate probable data quality. The analyses presented here used the full dataset (i.e., all responses with 14 or fewer demerits). Detailed information on the survey design, deployment, and data gathering, as well as a presentation of the basic demographics of the 2014 U.S. sample has been published elsewhere. Please see Tromovitch (2015) for more detailed information.

2.2. The Questions

Respondents were asked one of three questions to elicit their interpretation of the meaning of the term *significant* in the context of the reporting of scientific findings. Three questions were used as a form of built-in replication, to ensure that variations in the way the answers were solicited did not meaningfully impact the overall conclusions of the research. Two of these questions were free-response (i.e., text-box) questions, one was a multiple-choice (i.e., radio button) question. Most respondents were asked the multiple-choice question because this type of question allows for objective evaluation of the accuracy of the answers, whereas free-response questions require subjective evaluation by raters to categorize the answers. The order of the possible responses for the multiple-choice question was randomized for each participant to prevent order effects. The multiple-choice question read:

> When scientists declare that the finding in their study is “significant,” which of the following do you suspect is closest to what they are saying:
>
> - the finding is large
> - the finding is important
> - the finding is different than would be expected by chance
> - the finding was unexpected
> - the finding is highly accurate
> - the finding is based on a large sample of data

If there is no miscommunication between science and the public, the multiple-choice question has only one correct answer (“the finding is different than would be expected by chance”); respondents providing this answer were coded as *correct*. If, however, the public mistakenly uses the basic English definition of the term *significant* when it is used in reference to scientific findings, such respondents should have chosen either of the *large* or *important* responses; respondents providing either of these distracters were coded as *incorrectly using general English*. The remaining three distracters were coded as *other incorrect answer*. Failure to provide an answer was coded as *blank* (n = 5).

In the first mini-wave of data collection the multiple choice question only included the first four options (n = 30), in all later waves of data collection respondents were provided with all six options displayed above (n = 581), thus if no respondent knew the correct definition and all respondents guessed at random from the provided options, the mathematical expectation is that 17.1 percent of the responses would be correct by chance.

Unfortunately, the question did not include an “I do not know” option, hence blank responses may include both respondents who did not know the answer and chose not to guess (who should count toward the total number of respondents who did not know the correct answer) as well as respondents who chose not to answer the question (who should be excluded from analysis since it is not clear if they do or do not know the correct answer). The conservative analytical approach was taken and blank responses were excluded from analysis, thus the percentage of correct answers, provided below, may be an overestimate.

Instead of the multiple choice question, some respondents were asked one of two free-response questions and provided with space for a textual answer. The free-response questions differed from each other in that one explicitly provided the adjective *statistically* before the word *significant* whereas the other did not, however the implicit version clearly indicated that the question was focused on a scientific context. The explicit (i.e., adjective-prefaced) version of the question was worded:

> Scientists sometimes conclude that the finding in their study is “statistically significant.” If you were updating a dictionary of modern American English, how would you define the term “statistically significant”?

The implicit version of the question was:

> Scientists sometimes conclude that the finding in their study is "significant." If you were updating a dictionary of modern American English, how would you define the term "significant" for this context?

The textual answers were initially categorized by the researcher into the closest corresponding option used in the multiple-choice question (6 possibilities), or categorized as *other* (i.e., the respondent provided an incorrect definition that could not be matched to one of the distracters), *uncategorizable* (e.g., one respondent wrote: "when a gradual count has been identified"), *do not know* (e.g., "no idea"), *significant other* (e.g., "One's partner or other half"), *textual nonresponse* (i.e., the respondent wrote something but appeared to be intentionally not answering the question, e.g., the single word "assume" or the character string "dfdsfdsfdfdf"), or *blank*. There was no obvious pattern or predominant theme to the contents of the *other* category. Thus, there were 12 initial categories. The responses were then recoded as *correct*, *incorrectly using general English*, or *other incorrect answer* (*other* and *uncategorizable* were coded as *other incorrect answer*). Data from respondents coded as *blank* (n = 38) as well as data coded as *textual nonresponse* (n = 12) were excluded from analysis since it cannot be determined if they did or did not know the correct answer; data coded as *significant other* (n = 13) were also excluded from analysis on the basis that these respondents most likely did not read the question closely enough to know what was being asked and thus it cannot be determined if they did or did not know the correct answer. Responses coded as *do not know* were used in calculating the total N (since these respondents did answer the question and did not know the correct answer).

As a check on the author's categorizations, a second rater with a PhD in linguistics was provided with the textual answers and a scoring key. Seventy-two percent of the textual answers were initially coded identically by both raters. With the exception of items rated by either coder as *correct*, all discrepancies in coding were quickly resolved by discussion (i.e., 100 percent inter-rater agreement was achieved). Most discrepancies stemmed from deciding how to categorize a response that could reasonably be placed in more than one category or from the threshold used to distinguish between two categories (e.g., one rater coded "a breaking new way of thinking or a new method" as *important* and the other rater coded it as *other*; after discussion this answer was coded as *other*). Since correct answers are the most critical item for the analyses presented in this article and the author did not want to underestimate the public's level of correct understanding of the term *significant*, the conservative approach of using broad standards for coding an answer as correct was employed. As part of this approach, all items rated as correct by either rater were treated as correct.

3. Results

The results from the multiple-choice question appear in Table 1. As can be seen in the last column of the table, of the n = 611 respondents only 8.5 percent selected the correct answer. Most respondents selected an answer consistent with the general English meaning of the term *significant*. In order to explore whether possessing a doctoral degree impacted the respondents' understanding of the term, data from respondents who indicated they had completed a doctoral degree were tabulated separately from the data from the other respondents and subjected to a two (groups) by three (answer categories) chi-square analysis. The doctoral degree holders responses statistically significantly differed from the other responses ($\chi^2 = 19.9$, $p < .001$). Of the fifteen doctoral degree holders, six selected the correct answer and nine selected an incorrect choice. It should be noted that the research was not designed specifically to assess doctoral degree holders' knowledge, hence background data on these doctoral degree holders such as their field of research and details about their scientific preparation were not collected, limiting the generalizability of these findings (see Conclusions and Recommendations).

Table 1. *Percentage of responses to the multiple-choice question asking for the definition of the term* statistically significant

	Respondent completed a doctoral degree?		
	No	**Yes**	**Combined**
	***n* = 596**	***n* = 15**	***n* = 611**
correct answer	7.7%	40.0%	8.5%
incorrectly using general English	50.8%	40.0%	50.6%
other incorrect answer	41.4%	20.0%	40.9%

Note. The first 30 respondents were only provided with 4 answer choices, the later 581 respondents were provided with 6 answer choices (see main text), hence if all respondents guessed at random the mathematical expectation is that 17.1% of the responses would be correct.

The results from the explicit free-response question appear in Table 2. As shown in the last column of the table, of the $n = 250$ respondents only 4.0 percent provided an answer that was categorized as correct. All ten answers categorized as correct appear in the appendix. Although the question wording did not suggest the possibility of responding by indicating one did not know the answer, somewhat more than 10 percent of the respondents explicitly indicated they did not know the meaning of the term.

Table 2. *Percentage of coded responses to the explicit (i.e., adjective-prefaced) free-response question asking for the definition of the scientific term* statistically significant

	Respondent completed a doctoral degree?		
	No	**Yes**	**Combined**
	***n* = 245**	***n* = 5**	***n* = 250**
correct answer	3.7%	20.0%	4.0%
incorrectly using general English	42.9%	20.0%	42.4%
other incorrect answer	41.6%	60.0%	42.0%
indicated: do not know the meaning	11.8%	0.0%	11.6%

The results from the implicit free-response question appear in Table 3. As shown in the last column of the table, of the $n = 242$ respondents only 0.8 percent provided an answer that was categorized as correct. Both answers categorized as correct appear in the appendix.

Nearly 10 percent of the respondents explicitly indicated they did not know the meaning of the term.

Table 3. *Percentage of coded responses to the implicit free-response question asking for the definition of the term* significant *when used in a scientific context*

	Respondent completed a doctoral degree?		
	No	**Yes**	**Combined**
	***n* = 240**	***n* = 2**	***n* = 242**
correct answer	0.8%	0.0%	0.8%
incorrectly using general English	62.5%	100.0%	62.8%
other incorrect answer	28.3%	0.0%	28.1%
indicated: do not know the meaning	8.3%	0.0%	8.3%

To determine if the explicit and implicit versions of the question led to different patterns of results, a two (questions) by four (answer categories) chi-square analysis was conducted. The result was statistically significant ($\chi^2 = 23.0$, $p < .001$). To determine if the explicit and implicit versions of the question led to different levels of correct responding rather than merely a different pattern of responding, a two (questions) by two (correct; not correct) chi-square analysis was also conducted. This chi-square was also statistically significant ($\chi^2 = 5.2$, $p < .03$). As seen in the tables, the explicit (i.e., adjective prefaced) version of the question produced a significantly higher rate of correct responses and a lower rate of mistakenly using the general English definition of the term.

4. Conclusions and Recommendations

The concept and language of statistical significance has been taught in introductory statistics and research methods courses for well over half a century, with the basic terminology being taught at least a century ago. It seems likely that this may be the most basic concept taught in such courses in recent history. For example, in 1921 in his book titled *A First Course in Statistics*, Jones explained the terminology by writing “unless the observed differences fall outside recognized limits it is said that they are *not significant* of any difference other than such as might quite well be accounted for by random sampling alone” (p. 133; italics in original). Even more than a century ago (probably before the use of *p*-values came into use), the meaning of the scientific term *significant* was clearly being taught. For example, in 1911 Yule wrote:

> [T]he question again arises whether this difference may be due to fluctuations of simple sampling alone, or whether it indicates a difference between the conditions subsisting in the universes from which the two samples were drawn : in the latter case the difference is often said to be **significant**. (Yule, 1911, p. 262; bold in original)

Addressing the issue with more modern wording, Campbell and Stanley’s 1963 chapter on experimental and quasi-experimental designs points out that “statistical tests of significance come in for the decision as to whether or not the obtained difference rises above the fluctuations to be expected in cases of no true difference for samples of that size" (p. 28).

Across all three questions testing knowledge of the meaning of *statistical significance* (*N* = 1103) the vast majority of the U.S. national sample of adults provided an incorrect definition of the scientific term *significant*. Indeed, the percentage of correct answers to the multiple-choice version of the question was less than half of what would be expected by random guessing. Most respondents provided a definition that was consistent with the general English meaning of the term but that is unrelated to the meaning of the term in science. These data clearly point to a critical level of misinterpretation of scientific findings, in part due to science's failure to use widely understood terminology. Editors, peer reviewers, and authors should reduce the usage of the term *significant* in scientific publications, replacing it with wording that is less likely to cause misinterpretation (e.g., rather than writing that a finding is "significant", instead writing that the finding is "different than expected by chance"). This recommendation should be especially applied to the writing of abstracts and discussion sections since these are the sections of a scientific article that are most likely to be read by non-scientists. Usage in results sections, however, should also be scrutinized and improved. Consider, for example, the last sentence of the results section of this article which includes the wording ". . . a significantly higher rate of correct responses . . ."; will all readers, or any readers, know if this author was using the scientific or the general English term in this instance?

Since the typical misinterpretation of *statistical significance* is that a finding is large or important even though it might really be trivial in size, authors should be required to provide easy to understand effect size statistics (e.g., the *population association* or *IQ equivalent points*; IQEP, see Tromovitch, 2012) if they make any explicit or implicit claims regarding the size or magnitude of their findings. Indeed, asserting a claim of having found a large finding without providing an effect size measure could be considered scientific misconduct since one would be asserting something as fact without any apparent evidence to support the claim.

The sample included very few doctoral degree holders (*n* = 22), most of whom received the multiple-choice version of the question. That most provided incorrect answers suggests there may be deficiencies in doctoral level education in the United States. Since the study was not designed for the purpose of specifically testing doctoral degree holders, no data was collected on the area of their specialties, hence it is possible that most of these doctorates were in areas where a basic understanding of statistics is not emphasized. It should be noted, however, that in addition to the limitation of a small sample size, it is not known if any of these individuals were practicing scientists, professors, or others who would be in a position to misinform the public regarding scientific findings. Future research should examine the knowledge level of doctoral degree holders who influence the public, such as active professors, researchers, expert witnesses, consultants, and scientists, to determine if a reform of doctoral education is required. Such a reform might be simple, as it might only be necessary to ensure that all doctoral programs require at least one semester of statistics training and properly test student knowledge at the end of all required, basic, introductory statistics and research methods courses—and not allow any student to pass such a course who does not have the requisite knowledge. Similarly, ensuring the ability to read and understand basic statistical presentations could be made a required part of existing qualifying and comprehensive examinations, at least in fields where scientific reports make use of inferential statistics.

Although the explicit (i.e., adjective prefaced) version of the question produced a significantly higher rate of correct responses than the implicit version, this may be an example of where a significant difference is not significant. The 3.2 percent increased rate of correct responses provided when the word *significant* was prefaced with the adjective *statistically* suggests that the use of this adjective helps some readers realize that the general English definition is not in use. Hence this author recommends that editors and peer-reviewers require all usages of the term *significant* to be prefaced with an appropriate descriptor (e.g., "statistically" when indicating a finding was not likely to be due to chance alone; "practically" when a finding is both statistically significant and large; and "clinically" when a finding is statistically significant and large enough to suggest a need for action in psychology or a medical field; see Thompson, 2002), however, given that 96 percent of the respondents could not correctly define the term even when the adjective was used suggests that this may not lead to a meaningful (i.e., significant) improvement in the communication of scientific findings. Thus it is important for all articles reporting "significant" findings to explain the meaning of the term.

When one considers the ease with which false-positive findings can be published (i.e., an apparently statistically significant finding, but one that in fact is not accurate for the populations under study; see Simmons, Nelson, & Simonsohn, 2011; for a more mathematical treatment see Ioannidis, 2005) and the numerous misinterpretations regarding the meaning of *p*-values (see Cohen, 1994), it is clear that science's goal—understanding the truth and making that knowledge available to others—is currently hindered by multiple factors. Misinterpretation of the meaning of the scientific term *statistically significant* should be brought to an end, either through improved education, enforcement of the use of less ambiguous terminology, or both. Junior high school and high school science courses (and mathematics courses, if possible when doing a section on probability) could perhaps introduce the terminology of science to students so that most citizens will have the basic knowledge needed to avoid misunderstanding basic scientific reports. Additionally, journals should consider policies to improve scientific communication which include but are not limited to (1) requiring all usages of the term *significant* to be preceded by an appropriate adjective, (2) requiring all articles that use the term *statistical significance* to define the term prior to its first usage, and (3) requiring all assertions of having found a large association to be supported by easy to understand effect size statistics such as the population association or IQEP.

Acknowledgement

This research was supported in part by a grant-in-aid from the Science and Engineering Research Institute of Doshisha University.

References

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Boston, MA: Houghton Mifflin.

Cohen, J. (1994). The earth is round (*p* < .05). *American Psychologist*, *49*(12), 997-1003.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*,*2*(8), e124. Retrieved from http://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.0020124

Jones, D. C. (1921). *A first course in statistics*. London, UK: G. Bell and Sons (reprinted by BiblioLife).

Oxford University Press. (2012). *New Oxford American dictionary* (3rd ed.). Computer-based version installed with Apple computers running OS X 10.9 Mavericks.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359-1366.

Thompson, B. (2002). “Statistical,” “Practical,” and “Clinical”: How many kinds of significance do counselors need to consider? *Journal of Counseling and Development*,*80*(1), 64-71.

Tromovitch, P. (2012). Statistical reporting with Philip’s sextuple and extended sextuple: A simple method for easy communication of findings. *Journal of Research Practice*, *8*(1), Article P2. Retrieved from http://jrp.icaap.org/index.php/jrp/article/view/323/270

Tromovitch, P. (2015). The multinational life experience and personality project (MLEPP): Data collection and demographics of the 2014 U.S. and Japanese national samples. *The Science and Engineering Review of Doshisha University*, *55*(4), 308-316.

Yule, G. U. (1911). *An introduction to the theory of statistics*. London, UK: Charles Griffin.

Appendix:

Coding of Textual Responses

Twelve textual answers were coded as correct definitions of the term *significant* in a scientific context (see main text for question wordings and further details). No grammar, spelling, or other corrections were made to the answers below.

The following ten answers were coded as correct for the explicit version of the question.

1. “findings are different than random chance”
2. “statistically more than happen stance”
3. “notable variance from a base line. Conclusive.”
4. “Data that deviates enough from average or baseline to indicate a clear trend, from which a conclusion is obvious.”
5. “Non-coincidental; likely correlation between factors”
6. “A result that is not likely to occur randomly, but rather is likely to be attributable to a specific cause.”
7. “That something is not left only to chance.”
8. “A finding that has a strong probability of rejecting the null hypothesis.”
9. “95 percentile”
10. “A big enough colleration of numbers were it cannot simply be circumstanial”

The following two answers were coded as correct for the implicit version of the question.

1. “The likelihood that the result/findings are not due to chance and probably true.”
2. “statistically higher than normal”

Received 19 January 2015 | Accepted 12 February 2015 | Published 16 February 2015