CREST CENTRE FOR RESEARCH AND EVIDENCE ON SECURITY THREATS

20-MINUTE READ | GENERAL

SUBSTANCE OR SNAKE OIL?

How to evaluate a written claim of efficacy regarding a product or service.

> **How do we decide which studies are valid and what they mean?**

INTRODUCTION

Should your running shoes be chosen such that they match the shape of your feet to prevent running-related injuries? Evidence from randomised controlled trials and observational cohort studies suggest so. But other investigations have shown that ‘motion control shoes’ (shoes designed to limit foot rotation inward and downward) are more protective for experienced runners. How should you know which advice to follow? How can the information be sorted – if there are ‘studies’, how do we decide which studies are valid and what they mean?

STEP 1: IDENTIFY THE CLAIM

The first step is to identify what the authors are claiming or what question(s) they are asking; there may be a primary question as well as ancillary questions. These should be well-defined and understandable. For example, “...In this study, we: (1) investigated whether running shoes equipped with motion control features modified injury risk in regular leisure-time runners and (2) if this influence depended on foot morphology.”[1]

STEP 2: IDENTIFY THE METHOD

Examine how the evidence was *obtained*: this could be via observation, in which case the phenomenon is measured and otherwise not interfered with; it could be via *correlation*, in which case there is a measure of at least two variables but no manipulation of either: or via *experimentation*, in which case at least one variable is deliberately manipulated and at least one other variable is measured for how it is impacted by that change.

A variable is any factor, trait or condition that can exist in differing amounts of types.

ADDITIONAL DATA

A natural next step might be to go online and look for additional information. Certainly, there is much information to find; it has been estimated that by 2025, there will be 463 exabytes of data created globally each day (an exabyte is 10004 bytes, or 1,000,000,000,000,000,000 bytes).

Liebowitz, J. (Ed.). (2020). Data Analytics and AI. CRC Press.

OBSERVATIONAL DESIGN

The phenomena of interest are only measured and otherwise not interfered with. For example, the injuries and health problems[2] suffered by ultramarathon runners during a 219-km, five-day stage race were recorded. The measures were the total numbers and percentages of each subsequent clinical encounter with a health professional and their respective health problems. An *observational cohort* study is an observational study where the study participants are linked in some manner. For example, the ultramarathon runners could have been only those between the ages of 18 and 25.

A *case study* is an in-depth observation of a single individual, family, event, or other entity where multiple types of data may be included (psychological, physiological, biographical, etc.). A case study most likely will include a variety of quantitative and qualitative measures. Although case studies allow for an intensive analysis of an issue, they are limited in terms of generalisability. For example, a five-year case study of a women marathon runner's strategies and tactics measured body composition, maximal oxygen uptake, and running economy.[3] The authors found an 8% increase in race performance in the first three years but no improvement after that. The limitation of such studies is primarily that the results may not generalise to others, e.g. male marathon runners.

CORRELATIONAL DESIGN

In a correlational design, the variables of interest are not directly manipulated by the investigator but are selected because there is reason to think that they are related. At least one variable is measured directly. For example, the number of injured runners and the number of injuries per runner were compared for 107 runners that ran barefoot versus 97 who ran in shoes. Fewer injuries were observed in the barefoot runners.[4]

It is tempting here to infer a causal relationship – that is, that the shoes were the cause of more injuries. But with these data only, such an inference is not justified. There may have been another reason for the difference – and in fact, the study noted that barefoot runners ran fewer miles.

STEP 2: IDENTIFY THE METHOD

EXPERIMENTAL DESIGN

If we want to know whether shoes cause running injuries, we need to run an experiment.

An experiment is uniquely informative because it permits the study investigators to infer a causal relationship between at least two variables, one that is deliberately manipulated (the *independent variable*) and one that changes in value due to that manipulation (the *dependent variable*). Causal inferences are not possible for correlational or observational studies.

The issue of how study participants are selected is critical to experimental design. There are two broad categories of selection: matching participants on selected characteristics or sampling randomly.

MATCHING PARTICIPANTS

For example, an experiment was conducted to evaluate how three stability categories of running footwear (neutral, Nike Pegasus), stability, Nike Structure Triax, or motion control, Nike Nucleus) were associated with the occurrence of running-related pain in a population of women training for a long-distance running event.[5] The type of running shoe was the independent variable. The running-related pain was the dependent variable.

Women in each group were selected so that they were equivalent for height, weight, body mass index, passive Hallux dorsiflexion range of motion (the extent of backward bending of the big toe) and Q-angle (the angle formed between the muscles of the front of the thigh and the tendon that extends down from the quadriceps muscle in the thigh). By *matching* subjects on these variables, differences among the groups as a function of shoe type could not subsequently be attributed to differences among those possibly *confounding* variables.

SNAKE OIL

The term 'snake oil' most likely comes from a 19th-century remedy for joint pain brought into the U.S. with the arrival of Chinese laborers who built a U.S. transcontinental railway in the mid-1800s. Snake oil had long been a folk remedy in Chinese medicine. Later analyses showed that Chinese water-snake oil contains 20% eicosapentaenoic acid (EPA), one of the two types of omega-3 fatty acids that reduce inflammation, blood pressure, and cholesterol.

Graber, C. (2007). Snake Oil Salesmen Were on to Something. Scientific American, 1.

RANDOM SAMPLING

An alternative method of participant selection, usually considered more robust than matching (which can be difficult if the variables are complex), is to sample randomly from a larger group (i.e. from the population to which the study authors want to generalise their findings).

For example, 372 recreational runners were randomly given either a motion control or a standard version of a regular running shoe model and were followed up for six months regarding running activity and injury.[6] This is an instance of a *randomised controlled trial*.

A good random sampling procedure is one where a large number of samples of the same size can be randomly selected from the larger population. Random sample means just that: finding an authoritative index of random numbers (generated by a flip of a fair coin or by tables of random numbers) and using those to determine which runner would wear which shoe.

STEP 2: IDENTIFY THE METHOD

CONSIDER SAMPLE SIZE

Understanding how many subjects are needed depends on the expected impact of the independent variable (e.g. type of shoe) on the dependent variable (e.g. running activity and injury). Generally, the study authors will have prior knowledge that allows them to mathematically estimate this impact. In the motion-control shoes study, the authors noted that “Given an expected injury rate of 22% and 35% in the two groups … respectively, and a desired power of 0.8 and an α-level of 0.05, a total of 364 runners were required to test our main hypothesis.” The injury rate was estimated based on a previous study, and a reference to that study was provided. A “power of .8” means that the authors estimated that they would get a statistically significant difference between the shoe conditions 80% of the time. An “α-level of 0.05” indicates the authors accepted a 5% risk of concluding that the shoes make a difference exists when in fact there was no actual difference.

It should be noted that while α-levels are common statistics, numerous other statistics can be used and there is some controversy among statisticians as to which are most appropriate.

CONSIDER HOW THE STUDY PARTICIPANTS WERE FOUND

In the motion-control shoes study, participants were recruited via advertisements in local newspapers and on specialised Internet sites. How participants were selected is very important because this determines the people to whom the study results can be generalised. Given that the authors of the study were from the University of British Columbia in Vancouver, Canada and the Nike Sports Research Center in Beaverton, Oregon, U.S.A., the results of the study could most confidently be generalised to runners living in the northwest part of the U.S. However, it is common for assertions to be made more broadly (i.e. runners in general).

HUMAN SUBJECTS PROTECTIONS?

If the study involves humans, consider how they were treated. People who serve as participants in a research study have certain rights: they must be aware of any risks, be knowledgeable about the possible benefits from participating, be able to withdraw from a study at any point, and be treated in a dignified manner. The report of research with human subjects as participants should reference the authors’ adherence to all human subject protection regulations. In the UK, a study protocol must be reviewed and approved of ahead of implementation by a National Ethics Committee. In the motion-control shoes study, the authors noted that “All participants received a full description of the study protocol and provided written informed consent for participation. All procedures were approved by the National Ethics Committee for Research (ref 201211/04).”

STEP 3: WHAT EVIDENCE IS OFFERED?

Identify the evidence that is provided in support of the claim. Evidence must be *empirical* – that is, observation-based and seen, smelt, tasted, or heard directly or by using instruments to enhance these senses. (An alternative to empirical is inferential, which is not sufficient.)

STATISTICAL SIGNIFICANCE?

There are many measures of what is referred to as statistical significance. Statistical tests tend to vary by study design, including the nature of the measurements. In general, an outcome that is 'statistically significant' indicates that the reported differences cannot be attributed to chance alone. In an experimental design, an inference is made that the differences are most likely attributable to variations in the independent variable. For example, that the shoes were responsible for the differences in injuries.

A common statistic is a p-value; a value of $p<0.05$ means there is a less than 5% probability that the same effect would be found if everyone in the larger population were fitted with the same shoes and then assessed for running-related injuries – and that the findings are not the result of sampling error. Generally, $p>0.05$ is not acceptable. In a correlational design, an r-value is often reported, where r varies between -1 and +1 and reflects the strength of the relationship between two variables. A negative r-value means that as Variable 1 increases, Variable 2 decreases; a positive r-value means that as Variable 1 increases, so does Variable 2. An r-value of 0 means the two variables are unrelated; as the r-value approaches $|1|$, the variables are more strongly related.

WERE DIFFERENCES FOUND?

The outcome of a study may provide support for an initial claim. However, some outcomes may not support the claim or may support an alternative claim. It is important to note that finding no statistically significant difference (e.g. no difference in running-related injuries due to shoe type) does not allow the inference that there is no comparable difference in the larger population, e.g. that such shoes make *no* difference. Finding no statistically significant effects also could be due to inaccurate measurements, sloppy methods, or *noise*, which is unexplained variance within the samples. Too much noise means that if the study were repeated, it is not likely to produce the same outcome.

IDENTIFY POSSIBLE CONFOUNDING VARIABLES

A confounding variable is a variable that could account for a difference between groups or conditions other than the independent variable(s).

In the matched-participants study of shoe stability types (neutral, stability, or motion control), the shoes were de-identified so the study participants did not know which brand they were given. This is important because participants may have had different attitudes towards or experiences with the various shoes – these potentially confounding variables had to be eliminated as possible causes for the outcomes obtained. Ensuring that participants in an experiment do not know which group they are in is a prerequisite to a robust experimental study. In addition, the experimenters were *blind*, i.e. they were not aware of what shoe each participant was given. Each shoe pair was coded by a coworker not involved in the study before distribution. The code was broken only after the completion of data collection. When both the study participants and the experimenters are unaware of which condition each participant is in, the design is referred to as *double-blind*.

STEP 4: EVALUATE THE CLAIM

The authors will summarise their findings and in so doing, are likely to make inferences about the original claim. For example, the shoe study authors may conclude that the motion control shoe used resulted in both a greater number of injured runners and missed training days than the other two shoe categories. However, additional questions should be considered, and a good research report will include some discussion of these issues.

VALIDITY

How valid is the claim the authors make? There are various sorts of validity.

Internal validity

Internal validity refers to whether changes in the dependent variable are due to manipulation of the independent variable, e.g. were injuries due to the shoes or could the injuries be accounted for by some other variable? Could the authors rule out other factors? In fact, the authors of the Nike shoe study noted that there were differences among the groups in body weight, which may have accounted for their results.

External validity

External validity refers to the extent to which the outcomes of the study can be generalised to other settings (*ecological validity*), other participants or subjects (*population validity*), and over time (*historical validity*).

There are additional types of validity, mostly relevant to assessing tests that are devised to measure human characteristics or traits (such as IQ, anxiety, psychopathy, school readiness, etc.). For example, a *prospective study* design was used to examine whether personality factors predispose runners to injury.[7] Forty runners completed a personality test and were followed for one year during which they reported their training mileage. Runners with high scores on a Type A personality inventory experienced more injuries, especially multiple injuries. What can we make of these results?

CONSIDER CONTENT VALIDITY

Content validity is of two sorts: *face content validity*, which refers to whether the test assesses what it claims to assess (e.g. here, Type A traits, which include being competitive, impatient, easily upset and associating self-worth with achievement), and *construct validity*, which is whether the test is faithful to underlying theoretical concepts (e.g. here, that the test measures traits consistent with the underlying assumption that such traits are related to coronary heart disease).

CONSIDER CRITERION-RELATED VALIDITY

Criterion-related validity is relevant to the relationship of the test to other measures of the same construct. There are two sorts: *concurrent validity*, which is the extent to which the outcome of the test correlates with outcomes on similar tests (e.g. that a person taking the Type A Self-Rating Inventory – the test used in the study – would score similarly if they a took a different personality test). *Predictive validity* is the extent to which the test predicts later performance on a related construct (e.g. a prediction is made based on their Type A score that they will more likely participate in races, and this is found to be so).

RELIABILITY

How reliable is the outcome? Reliability refers to the trustworthiness or consistency of the outcome. Is the measurement free enough of *random error*? If the study were repeated with the same or a similar sample, would the outcome be essentially the same? If the study were repeated with the same sample later in time, would the outcome be essentially the same?

Random error is an error that is due to chance alone. Random errors are nonsystematic and occur arbitrarily when there are unknown or uncontrolled-for factors that affect the variable being measured or the process of measurement. Random sampling ensures that random error is equally distributed across sample.

LOOK FOR POSSIBLE BIAS

The source of the evidence also is relevant. Source bias means that the researcher or research agency has a prejudice for or against the claim outcomes. To avoid bias the claimant should not gain or lose from the claim outcomes, conduct or pay for the research providing the supporting evidence, and/or selectively choose the evidence presented. In the Nike shoe study, might we wonder if the author that is employed by Nike might be biased towards one or other of the Nike shoes?

Recall that the experimenters didn't know which participants had which shoes – they were blind to the assignment of shoes – so this might make such bias improbable.

INDICATORS OF A GOOD STUDY

The study is published in a peer-reviewed journal. This means that the study was critically assessed by qualified experts who were blind to the identity of the authors (i.e. the reviewers were not given the authors' names).

The authors acknowledge and discuss the limitations of their research. This might mean that they point to the peculiar nature of the sample (e.g. all the runners were in the UK, so that generalisation to EU runners may not apply) or to the method (e.g. the authors used self-report of running mileage, running a risk that people might cheat on such reports).

The authors provide content information. This may be a review of previous, related studies, references to related reviews, or documents such as reports that are relevant to the topic.

The authors relate their findings to previous findings, noting similarities and/or differences. If there are differences, some possible reasons for the differences are offered.

Assertions as to previous findings or current knowledge are appropriately referenced, and there is no attempt to hide findings as 'proprietary.' Anything other than common knowledge (e.g. the sun rises in the east, oceans vary in depth, etc.) should include a reference to a reputable source.

> **Anything other than common knowledge (e.g. the sun rises in the east, oceans vary in depth, etc.) should include a reference to a reputable source.**

INDICATORS OF A POOR STUDY

No context is provided – that is, the authors do not put their study in a larger scientific context. Science is cumulative so that there will always be relevant previous research. The authors fail to exhibit knowledge of this previous research.

STEP 4: EVALUATE THE CLAIM

Poor sampling methods – that is, samples are not appropriately matched or are not selected randomly from a larger, relevant population.

The methods are not sufficiently described – the reader should be able to essentially repeat the study based solely on the information offered in the report.

The method is not scientific. As noted, many methods are scientific, however, providing only anecdotal data is not scientific.

The authors fail to provide a rationale for their choice of sample size.

If the study uses experimental methods, and the sample sizes are either very small or very large, the study may lack validity. If the sample size is very small, it is unlikely that a difference among samples could be found, in which case the study authors might have missed a difference that exists in the larger population. If the sample size is very large, there is the danger of finding a difference that is statistically significant but very small, and therefore not very meaningful in terms of any application of the findings to real-life situations.

If the study looks for correlations between two or more samples in a large number of measures, then the sample size(s) should be much larger than the number of measures.

The authors hide some aspects of their data or their methods under the rubric of 'proprietary' or 'sensitive/classified.' This does not mean the study is problematic per se, but it does mean that the reader cannot properly evaluate the study to know how to interpret or make use of the results.

In addition to information overload, people are subject to numerous logical fallacies when confronted with arguments for or against a product, relationship advice, healthcare precautions, and so on. Some of these are:

- Argument from ignorance: since something has not been proven false, it is therefore true. For example, it has not been shown that female runners do not prefer Nikes, so we infer that they do.
- Appeal to popularity: A proposition is argued to be true because it is widely held to be true. A good example of this is choosing a running shoe based on online ratings.
- Appeal to authority: An appeal to authority is inappropriate if the person is not qualified or if experts in the field disagree.
- False analogy: When two objects or events A and B are shown to be similar, the assumption that if A has property P, so does B. For example, if male runners prefer Nikes, so do female runners.
- Affirming the consequent: The argument of the form, If A then B. B, therefore A. For example, 'Male runner prefer Nikes. So if a runner prefers Nikes, he is male.'

Evaluating research reports is a matter of judgment. The more expert one is in the research area, the better one might be in making such judgments. There is no easy formula.

The table (**next page**) is a brief overview of some of the important questions that should be asked when evaluating a research report.

QUESTIONS TO ASK WHEN EVALUATING RESEARCH

Questions to ask	Judgments to make	
	👍	👎
Does the design allow for causal inferences?	Experimental: Yes	Observational, correlational: No
Are study participants properly selected?	Random samples: Yes	Matched samples: Depends on matching method
Are sample sizes appropriate?	Too small: Might miss effects that are actually present	Too large: Risk meaningless outcome(s)
Are there ‘human subjects protections?’	Yes, referenced	No, no report
Are the findings empirical?	Yes: Scientific	No: Descriptive, not scientific
Were there statistically significant differences among groups or conditions?	Yes: Can assume difference exists in population	No: No conclusion can be reached
Are the findings valid?	Yes: Lack confounding variables	No: Significant confounding variables possible
Are the findings reliable?	Yes: Could be repeated with same outcome	No: Unrepeatable with same or different participants
Any hidden data?	No such claims; transparency good	Claims of proprietary or ‘sensitive/ classified’: Cannot judge
Bias possible among authors?	No bias evident: No reason to suspect findings	Author(s) may have equity in outcome: Bias makes findings suspect No bias evident: No reason to suspect findings

Table 1. Basic steps to evaluate a written claim of efficacy regarding a product or service.

READ MORE

1. Malisoux, L., Chambon, N., Delattre, N., Gueguen, N., Urhausen, A., & Theisen, D. (2016). Injury risk in runners using standard or motion control shoes: a randomised controlled trial with participant and assessor blinding. British Journal of Sports Medicine, 50(8), 481–487.

2. Scheer, B. V., & Murray, A. (2011). Al Andalus Ultra Trail: an observation of medical interventions during a 219-km, five-day ultramarathon stage race. Clinical Journal of Sport Medicine, 21(5), 444–446.

3. Jones, A. M. (1998). A five-year physiological case study of an Olympic runner. British Journal of Sports Medicine, 32(1), 39–43.

4. Altman, A. R., & Davis, I. S. (2016). Prospective comparison of running injuries between shod and barefoot runners. British Journal of Sports Medicine, 50(8), 476–480.

5. Ryan, M. B., Valiant, G. A., McDonald, K., & Taunton, J. E. (2011). The effect of three different levels of footwear stability on pain outcomes in women runners: a randomised control trial. British Journal of Sports Medicine, 45(9), 715–721.

6. Malisoux, L., Chambon, N., Delattre, N., Gueguen, N., Urhausen, A., & Theisen, D. (2016). Injury risk in runners using standard or motion control shoes: a randomised controlled trial with participant and assessor blinding. British Journal of Sports Medicine, 50(8), 481–487.

7. Fields, K. B., Delaney, M., & Hinkle, J. S. (1990). A prospective study of type A behavior and running injuries. J Fam Pract, 30(4), 425–429.

COPYRIGHT

IMAGE CREDITS

Copyright ©2021 R. Stevens / CREST (CC BY-SA 4.0)