The illusion of certainty and the certainty of Illusion:
A case study of misunderstandings in scientific articles

Tom Lang
Tom Lang Communications and Training International, Kirkland, WA, USA

Correspondence to:
Tom Lang
10003 NE 115th Lane
Kirkland, WA 98933
USA
tomlangcom@aol.com
+1 425 242 1370

This article was reprinted from: Lang T. The illusion of certainty and the certainty of illusion: a case study of misunderstandings in scientific articles. *AMWA J*. 2013;28(3): 105-9. Copyright 2013 by Tom Lang Communications and Training International.

Abstract

Critical thinking is necessary to edit a scientific article. However, in addition to questions about the language, we can also question the assumptions, documentation, and implications of the research, in a process I call "analytical editing." A text with unverified assumptions, missing documentation, and unconsidered implications can lead readers into believing that they understand an article when they do not, creating the "illusion of certainty." Here, I present an example of the analyses needed to understand a single sentence; a case study, if you will, of analytical editing. A close look at the sentence raises several important questions about meaning, measurement, statistical analyses, how data are presented, and how results are interpreted. Analytical editing, in conjunction with traditional substantive editing, allows editors to increase their professionalism and value-added to clients.

The single biggest problem in communication is the illusion that it has taken place.
George Bernard Shaw[1]

Science is based on writing. Only writing allows science to be recorded, evaluated, and reproduced and enables it to be systematic, cumulative, and public; the characteristics that distinguish it from authority, intuition, and tradition as a way of establishing "truth."

Publication–the final stage of research–depends on writing, as does evidence-based medicine, which is *literature-based* medicine.[2]

Given the importance of writing in understanding and advancing science, one would think that physicians and researchers would be provided full support in preparing publications. However, at least in clinical medicine, such support is often inadequate. Researchers are not expected to do their own literature searches and so are given access to librarians. They are not expected to do their own data analysis and so are given access to statisticians. They are not expected to render their own graphs and drawings and so are given access to medical illustrators. But for some reason, we expect them to do their own writing–to communicate technical information accurately, completely, clearly, and economically, in words and images–without specific training, and often without the support of professional medical writers and editors. Thus, we shouldn't be surprised that a large portion of the scientific literature is not immediately, accurately, and completely understandable.

One of the most important lessons I have learned in almost 40 years of editing is that the certainty we believe we have about understanding even a simple, straightforward sentence is often illusory. The sense of certainty is so strong that we don't even think to question the meaning. Only on closer examination does the illusion become apparent. Further, such sentences are found in most scientific articles, which is to say, these illusions are also a certainty in the scientific literature.

I encountered a good example of a sentence in which the actual meaning differs remarkably from its apparent one. In this article, I pose some questions that need to be answered if this sentence is to be understood correctly. These questions are part of what I call "analytical editing," or editing to assure that research designs and activities are documented appropriately and explained adequately.[2] Analytical editing seeks to meet the needs of evidence-based medicine by making sure the evidence itself is completely and clearly reported. Analytical editing does not require us know medicine. It does require that we know how medical research is conducted–or at least what questions to ask about the research–as well as the standards to which this research should be documented. A task often left to peer reviewers, analytical editing can be done by trained editors and, in conjunction with traditional substantive editing, allows editors to increase their professionalism and value-added to clients.

The Example

This sentence was in the results section of a poorly written abstract: *One group of patients was significantly less depressed than the other*. The sentence seemed straightforward, but the more I analysed it, the more questions I had.

The questions

Question 1: What is the context of the sentence?

The sentence was the second in the results section of the abstract. Taken by itself, the sentence could have been a description of the patients at baseline, an incidental finding that might confound the results, or a result itself. Given the context of the article–a study of a new antidepressant–it was probably the result of the study.

Meaning is a product of message and context, in the same way that the meaning of a picture is a product of image and background (Figure 1). Change the context, and the same message has a different meaning. *The wall was built to scale* means something different to an architect than it does to a climber. For this reason, the context of every scientific article needs to be clear to rule out other interpretations made possible by different contexts. One function of a good introduction is to put the research in the proper context.

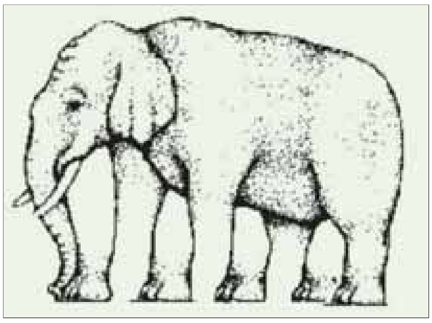

Figure 1. The "figure-ground" effect that becomes apparent from trying to make sense of this image is similar to what happens when we interpret a written message in different contexts. The context determines the meaning to some extent.

Question 2: Who was studied?

The article stated that the participants were women outpatients with moderate-to-severe depression being treated at a university hospital. The two groups mentioned in the sentence were the treatment and control groups of the study, something the sentence could have said, "Patients in the *treatment* group were significantly less depressed than were patients in the *control* group." We also need to know the patients' age, diagnosis, how they were selected for the study (the sampling method and eligibility criteria), other health conditions, and so on.

How the sample size was determined also needs to be explained. Especially in randomised trials, sample size should be determined with a power calculation. Basically, a power calculation tells investigators how many patients they need to enroll in a study to have, say, an 80% chance of detecting a difference of a given size *if such a difference actually exists in the population from which the sample was taken*. Investigators rarely get a chance to study an entire population. Instead, they have to study of a sample of that population. However, there is a chance that the sample won't include patients that express the difference of interest, a problem called "sampling error." The power calculation estimates the size of the sample likely to be large enough to include patients that express the difference at a degree of uncertainty acceptable to investigators.

In "underpowered" studies–studies that did not enroll enough patients to detect the desired difference–the lack of a statistically significant difference doesn't mean the groups are similar, it means the study was inconclusive: "absence of proof is not proof

…the lack of a statistically significant difference doesn't mean the groups are similar, it means the study was inconclusive: "absence of proof is not proof of absence."

of absence." The difference of interest is usually the smallest considered to be clinically important, so we have to determine this difference and whether the study enrolled enough patients to have a reasonable chance (often 80% or 90%) of detecting it.

Question 3: What was studied?
Depression can be treated in several ways, so the treatment needs to be described in detail. If the treatment is a drug (as it was in this example), we need to know the generic name, manufacturer, dosage, route of administration, and perhaps the indications, possible side effects, and the degree to which each group took the medication as planned. The rate of protocol adherence is usually higher in in-patient studies than in outpatient studies, for example.

Question 4: How was depression measured?
All study variables must be defined in objective, measurable terms. In this case, we need to know how depression was measured. Was the diagnosis based on a physician's judgment, a self-report questionnaire, or some other way? The text said that "All patients completed the Beck Depression Inventory before and after treatment." The Beck Depression Inventory is a common, validated instrument for measuring depression. This information was encouraging. Many authors do not say how they measured their variables, often because "my readers will know." Right.

Question 5: What type of comparison is being made?
In a study with two groups in which both pre- and post-treatment values are measured, two comparisons are possible. The within-group comparison looks at the changes between pre- and post-test values for each group, whereas the between-group comparison looks at the differences between groups at the beginning or end of the study. In a study like this one, both comparisons are likely. However, the sentence in question says that one group was less depressed than the other, so we have to ask whether the statement refers to a *between-group comparison*–at the end of the study, mean depression scores in one group were lower than the mean of those of the other (and presumably the baseline scores were similar)– or a *within-group comparison* – the change in depression scores during the study was greater in one group than in the other (and the baseline scores were not necessarily similar).

Question 6: How large was the difference between groups?
The authors reported that "The mean depression score of the treatment group was 38% lower than that of the control group." Fine, but results expressed only as percentages are *always* suspect. Numerators and denominators should always be available for all percentages.[3]

There is an old laboratory joke about how 33% of the rats lived, 33% died, and the last one got away. It is also usually true that a 50% reduction from 2 to 1 is not the same as a reduction from 2,000 to 1,000. Hence, the need to provide numerators and denominators when reporting and interpreting percentages

Mean values can also be a problem. If Bill Gates walks into a room, the average income of people in the room skyrockets, but nobody makes any more money. Here, it is possible that the lower mean depression scores represent not an overall decrease in the severity of depression but rather an effect caused by a few patients who responded unusually well to treatment (Figure 2).

Question 7: What does the author mean by "significantly"?
In medical writing, *significant* should be reserved for its statistical meaning, but the term is still often used to mean *markedly* or *substantially*.[2-4] An accompanying *P* value or a 95% confidence interval usually indicates that the term is used for its statistical meaning, but not always. In the present example, *significant* was used in its statistical sense.

The most common reporting error in medical articles is confusing statistical significance with clinical importance.[2,3] Relying on *P* values to interpret results is often easier than considering whether a result is clinically important. However, even when used appropriately, *P* values themselves must be reported correctly. We need to know the actual *P* value (*P*=0.03, not *P*<0.05); the alpha level (usually 0.05) that defines the threshold of statistical significance; the statistical test used to calculate the *P* value; whether the assumptions of the test have been met by the data (eg, whether the data are independent or paired); whether the test was 1- or 2-tailed; and the statistical software program used in the

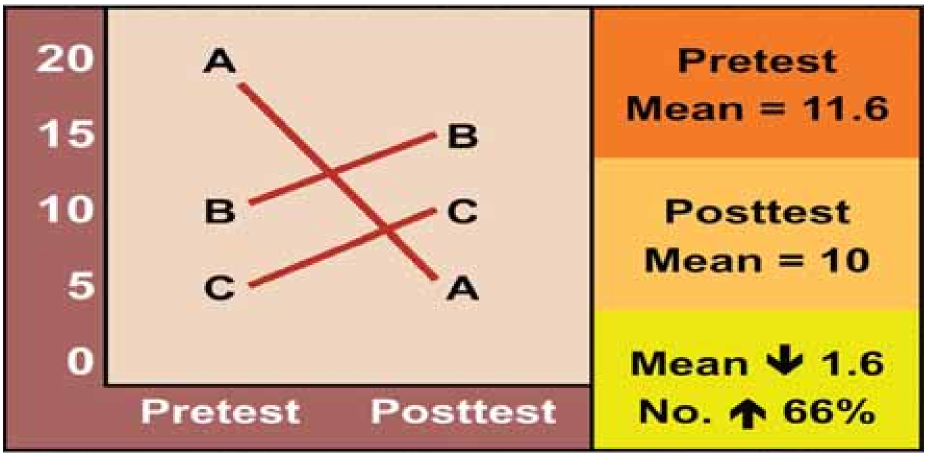

Figure 2. The problem of reporting a change in group means or the number of patients in whom change occurred. Here, the large change in patient A has had a disproportionate affect on the mean of all three patients. Thus, the data can be reported either as the fact that the mean decreased by 1.6 points, from 11.6 to 10 (14%), or that 67% of the patients had increased values. (Of course, the 67% is only 2 of 3, but it's still 67% . . .)

analysis (to establish its validity).[3]

Returning to the manuscript at hand, had the authors said something like, "One group was less depressed than the other (*P*=0.02)," we would have known that "significant" was used in its statistical meaning.

Question 8: How precise is this estimate of the difference?

The results of most biomedical studies are, in fact, estimates, and estimates require a measure of precision.[3] In medicine, this measure is usually the 95% confidence interval. I think of the interval as being the range in which the mean difference is expected to occur in 95 of 100 similar studies and in which the difference would be outside the range in the remaining 5 of the 100.[3]

Confidence intervals are useful because they keep the interpretation focused on the effect size and therefore on the medicine, not the *P* value.[3] Confidence intervals that contain both clinically important and clinically unimportant values ("heterogeneous" intervals) suggest that, even if the difference in means is statistically significant for the current trial, the estimate is not probably not precise enough to conclude that the treatment will likely be effective in 95 of 100 similar trials. In other words, the result is clinically inconclusive.

Typically, a larger sample size gives a more precise estimate (a narrower confidence interval). What is important is not the width of the confidence interval but its "homogeneity." When the confidence interval contains only clinically important values, or only clinically unimportant values, then we have a more definitive answer to the research question.

Ideally, the authors would have written something like: "The difference between means was 3 points (95% confidence interval, 1.5 to 4.5 points)." But they didn't.

Question 9: What is the measurement scale for depression?

The Beck Depression Inventory is a scale that runs from 0 to 63 (Figure 3). Scores of 0 to 9 indicate no or minimal depression; 10 to 18, mild depression; 19 to 29, moderate depression, and 30 to 63, severe depression.[5]

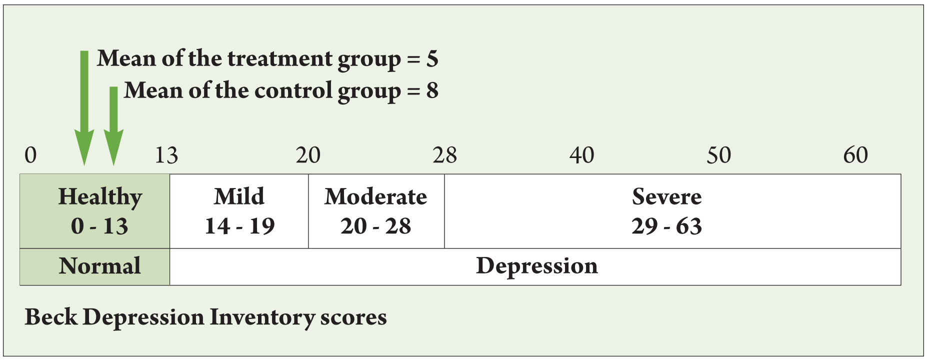
Mean of the treatment group = 5
Mean of the control group = 8
0
13
20
28
40
50
60
Healthy
0 - 13
Mild
14 - 19
Moderate
20 - 28
Severe
29 - 63
Normal
Depression
Beck Depression Inventory scores

Figure 3. The Beck Depression Inventory is a common, validated instrument for measuring depression. To understand the measurement, however, we must answer several questions: 1. Is the scale linear? That is, does a 3-point difference at one end of the scale mean the same thing as a 3-point difference at the other end? 2. What is the smallest difference in scores that is clinically meaningful? 3. Are their any threshold scores

So, the 3-point difference between means, and its 95% confidence interval, has to be interpreted accordingly.

When we know the scale, we can also infer something about the baseline values. Remember, the text said that "All patients completed the Beck Depression Inventory before and after treatment." It is reasonable to conclude, then, that all women had Beck scores of at least 20 at baseline, and we hope the text will confirm this fact. The results are reported as the means of the post-treatment Beck scores, but it would be nice to know the mean baseline values of both groups. In some studies, if mean baseline values are close to normal, even the best treatment may show little effect because the range over which the means can drop is limited.

Question 10: What is the smallest clinically meaningful difference?

When reporting and interpreting results, the effect size (say, the differences between means) is usually more important than the *P* value. The effect size can be interpreted clinically, whereas a *P* value cannot.[3]

The authors revealed that after the intervention, the difference between the means of the treatment and control groups was 3 points. However, a difference, to be a difference, must make a difference. The "critical effect size" (the minimum clinically important difference) for the Beck Inventory was not given. (It turns out to be 5 points.[6] More on this later.)

What are we to make of this 3-point difference? Does it matter whether the difference crosses one of the threshold scores that define a different degree of depression? Does it matter whether the difference occurs at the low end or high end of the scale? Pain measured on a 10-point scale may be nonlinear; that is, a reduction from 9 to 8 may be greater than a reduction from 4 to 3.[7] We don't have to know whether the scale reflects a linear relationship among scores, however, we just have to ask authors if it is (Figure 3). Don't be surprised if they don't know.

Question 11: What were the actual mean values of both groups?

Now the illusion became apparent. A table showed that the mean score was 5 in the treatment group and 8 in the control group. These values are consistent with the 3-point difference between means and with the 38% lower score of the treatment group ([8 – 5]/8 = 0.38 Î 100 = 38%). However, both means are in the normal range (scores less than 9; Figure 4), so describing the result as "one group is less depressed than the other" is incorrect and misleading. The scores also differ by less than 5 points, so the difference is not clinically important.[6] The authors seemed to have based their interpretation solely on a significant *P* value, without considering the clinical implications of the results. "Why" could be a most interesting question here.

Question 12: What was the proportion of patients in each group who were still depressed after treatment?

The example compared the means of two groups. However, a common error in clinical research is to report changes or differences in means rather than indicating how many

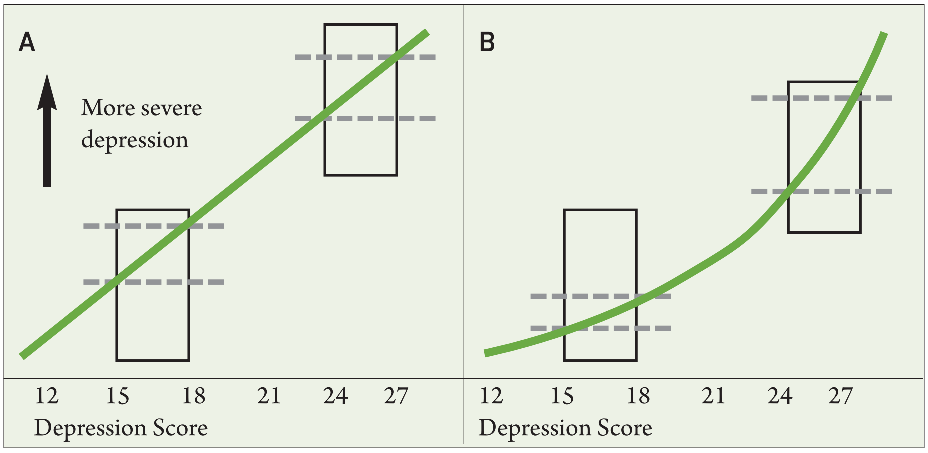

Figure 4. Measurement scales may or may not be linear. A. If the scale is linear, a 3-point change at the high end means the same thing as a 3-point change on the low end: the distance between the dotted lines is the same in both rectangles. B. If the scale is not linear, where the distance between the dotted lines in the two rectangles is different, the importance of a 3-point change depends on where that change occurs on the scale.

patients got better or worse (Figure 2). It would have been helpful to know how many patients were no longer depressed by the end of the study.

The issue here is the "unit of observation." I once edited a manuscript describing a study of 25 eyes, but it never said how many patients were involved. The unit of observation was eyes, not patient. The primary outcome of interest – the unit of observation – is in the protocol, but, as in the example, how patients responded is often and surprisingly not given.

Question 13: Is the drug likely to be generally effective?

Determining the effectiveness of the drug was the purpose of the study. The authors' claim that "*one group of patients was significantly less depressed than the other*" was supposed to mean that the drug was effective. They should have written something like: "After treatment, 72% (38/53) of the treated patients and 49% (27/55) of the control patients scored 9 or below on the Beck Inventory (95% CI for the 23% difference, 2% to 41%)," but they didn't. Instead, given the small effect size (3 points on the Beck scale in which 5 points is the smallest important difference), the fact that both means were in the healthy range, the lack of a confidence interval, and not knowing how many patients were no longer depressed at the end of the study, it does not seem reasonable to agree with the authors that the drug was effective.

However, we also can't conclude that the drug was ineffective. The difference was statistically significant, if clinically irrelevant. The drug did reduce the mean of the treatment group from well above 19 to 3, which supports the claim of efficacy, but the mean in the control group may have been reduced to a similar degree. All we can say is that they study was not well conducted, not well reported, or both.

Conclusions

Not all sentences are this involved, but many are and require analysis as detailed as the example presented here. Analytical editing can take time – and skill, training, and experience. What makes good writing and editing valuable is that they reduce readers' time, effort, and uncertainty about the meaning of a text, and they don't create the illusion of clarity. The problem is that many scientific articles are poorly written and poorly edited. Worldwide, authors are generally not skilled in communicating technical information in writing and do not receive adequate editorial support, and most journals provide only superficial copyediting. This situation pretty much assures that readers of the scientific literature will regularly encounter the "illusion of certainty" and therefore must be prepared to accept the "certainty of illusion."

Acknowledgement

My thanks to Bart Harvey, MD, PhD, MEd, FRCPC, FACPM, for his review of the manuscript and help with the calculations.

References

1. Cited in: Caroselli M. Leadership skills for managers. New York: McGraw Hill Professional; 2000.
2. Beck Depression Inventory. 2013 [cited June 6, 2013]. Available from: http://en.wikipedia.org/wiki/Beck_Depression_Inventory.
3. Lang T. How to write, publish, and present in the health sciences: a guide for clinicians and laboratory researchers. Philadelphia: American College of Physicians; 2010.
4. Lang T, Secic M. How to report statistics in medicine: annotated guidelines for authors, editors, and reviewers, 2nd ed. Philadelphia: American College of Physicians; 2006.
5. Style Manual Committee, Council of Science Editors. The CSE manual for authors, editors, and publishers. 7th Ed. Reston, VA: Council of Science Editors in cooperation with the Rockefeller University Press; 2006.
6. Hiroe T, Kojima M, Yamamoto I, Nojima S, Kinoshita Y, Hashimoto N, *et al*. Gradations of clinical severity and sensitivity to change assessed with the Beck Depression Inventory-II in Japanese patients with depression. Psychiatry Res. 2005;135(3):229-35.
7. Aubrun F, Langeron O, Quesnel C, Coriat P, Riou B. Relationships between measurement of pain using visual analog score and morphine requirements during postoperative intravenous morphine titration. Anesthesiology. 2003;98(6):1415-21.

Author information

Tom is an international consultant and educator in medical writing/editing, critical appraisal of clinical research, and scientific publications. He is past President of the Council of Science Editors and current Treasurer of the World Association of Medical Editors. He also served on the CONSORT and PRISMA groups and wrote *How to Write, Publish, and Present in the Health Sciences.*