
**Review**

**LAS COSAS NO SIEMPRE SON LO QUE PARECEN: USO INAPROPiado DE LAS ESTADÍSTICAS EN LOS DEPORTES Y LAS INVESTIGACIONES DE SALUD**

**THINGS AREN’T ALWAYS AS THEY SEEM: INAPPROPRIATE USE OF STATISTICS IN SPORTS AND HEALTH RESEARCH**

Beck, Travis W.¹; Young, Kaelin C.²; Stock, Matt S.³.

¹*University of Oklahoma*

²*Wichita State University*

³*Texas Tech University*

Correspondence to:

Travis W. Beck  
University of Oklahoma  
Department of Health and Exercise Science  
Tel: (405) 325-1378  
Email: tbeck@ou.edu

Edited by: D.A.A. Scientific Section  
Martos (Spain)

Received: 4/1/2015  
Accepted: 1/6/2016

J Sport Health Res  
ISSN: 1989-6239
Las estadísticas son a veces objeto de abusos en la investigación en deporte y salud. Al igual que hay muchos medios excelentes y muy útiles para transformar los datos, también hay varias técnicas que llevan a la confusión. Independientemente de que estos errores sean intencionales o no, el resultado final es el mismo en el sentido de que influyen negativamente sobre práctica clínica y el avance de la ciencia. Por desgracia, la revisión por los pares no siempre es capaz de sacar a la luz estos errores. Por lo tanto, la publicación de un manuscrito no garantiza que los datos que contiene hayan sido analizados correctamente. Por otra parte, estas prácticas incorrectas pueden a veces ser tomadas como modelo por los estudiantes e investigadores noveles que citan trabajos anteriores y asumen que la técnica es aceptable "porque estudios anteriores en este área las han utilizado". El propósito de este documento es describir algunas de estas prácticas. Con este documento pretendemos aumentar la conciencia acerca de estas técnicas, así como alentar a los investigadores a utilizar sus propias habilidades de pensamiento crítico.

**Palabras clave:** Estadística, medición, investigación.

**Keywords:** Statistics, Measurement, Research.
INTRODUCTION
An assumption often made by the general public is that peer review guarantees quality of the methods used to collect and analyze the data from published studies. Unfortunately, the peer review system is far from perfect and has received much criticism for inhibiting scientific advancements and allowing use of inappropriate statistical techniques (Horrobin 1982, 1990, 1996, 2001; Rothwell and Martyn 2000). The latter is the focus of this paper. The abuse of statistics is certainly not a new topic. Huff (Huff 1954) was one of the first to describe some of these methods in his extremely popular book entitled: How To Lie With Statistics. Since that time, a number of different review articles have been written about how statistics can sometimes be deceiving (Glantz 1980; Tharyan 2011; Prescott and Civil 2013; Gore et al. 1977; Weech 1974). Interestingly, the origin for many of these techniques can be traced to advertising, where scientific integrity is often compromised in favor of the profit margin. It is not at all uncommon to see advertisements use things like deceiving graphs, lack of an appropriate control group, and inappropriate sampling in an effort to persuade the naive consumer into buying what they think is a “scientifically proven” product. Other misconceptions come from statistics textbooks (Brewer 1985). Even the best-selling books have been found to contain half-truths, definitional errors, and inference-based mistakes (Brewer 1985). Regardless of the source, the migration of some of these techniques into peer-reviewed research is dangerous because it not only obscures the truth, but also sets a precedent for the types of methods that are considered appropriate science. Is a treatment effect that appears to be real simply an artifact of the way that the data were analyzed? Is it appropriate to use a given technique simply because it was used in a previous study? How common are these errors in research, and how can someone that is not trained in statistics identify them? Unfortunately, there are no universal answers to these questions. Those looking for a “cookbook” approach to proper data analysis are likely to be disappointed. Every data set should be analyzed carefully and with a full understanding of the implications involved with using a given method. There are, however, several statistics-related mistakes and deceiving methods that are somewhat common in sports and health research. The purpose of this paper is to describe some of these methods, why they are used, and the misconceptions that they can create. We will not, however, vindicate previous studies that have used them. As such, we will not be citing the investigations that have used these methods. Rest assured that these studies are out there. The purpose of this paper is not to go on a “witch hunt”, but rather to expose these methods and encourage other researchers to be aware of them when they are reading the literature and analyzing their own data. It is our hope that this paper will bring increased awareness of these methods and improve the quality of sports and health research.

IGNOING STATISTICAL ERROR
Most introductory statistics courses provide at least a cursory description of statistical error. This concept is generally presented in a theoretical context with a table similar to that shown in Figure 1. The two possible errors (Type I and Type II) are committed by making an incorrect decision regarding the truth of the null hypothesis when performing a null hypothesis significance test (NHST). The null hypothesis was originally developed by the English statistician Ronald Fisher (Fisher 1935) and is always stated in terms of no difference (when comparing means) or no relationship (when performing correlation or regression). A Type I error is therefore committed when the researcher(s) incorrectly rejects the null hypothesis and, therefore, falsely concludes that the data provides sufficient evidence of a mean difference(s)/relationship(s). Type I errors are often referred to as false positives (Figure 1) because they erroneously suggest a positive treatment effect.

However, a Type II error occurs when the researcher(s) incorrectly retains the null hypothesis, thereby falsely concluding that the data provide insufficient evidence of a mean difference(s)/relationship(s). Type II errors are sometimes referred to as false negatives (Figure 1) because they incorrectly suggest that there is not enough evidence to support a treatment effect. The fact that these errors are often presented in a theoretical context (such as in Figure 1) leads many researchers to underestimate their importance, since one can never know with absolute certainty if they have committed a Type I or a Type II error. For example, studies with very large sample sizes run the risk of committing a Type I error, since miniscule and even meaningless treatment effects can be
statistically significant with enough power. In contrast, investigations that have very small sample sizes are prone to Type II errors, where large and clinically meaningful treatment effects may not be statistically significant due to insufficient power.

<table>
<thead>
<tr>
<th>Reject Null Hypothesis</th>
<th>Null Hypothesis Is True</th>
<th>Null Hypothesis Is False</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Type I Error</td>
<td>Correct Outcome</td>
</tr>
<tr>
<td></td>
<td>False Positive</td>
<td>True Negative</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Fail To Reject Null Hypothesis</th>
<th>Correct Outcome</th>
<th>Type II Error</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>True Positive</td>
<td>False Negative</td>
</tr>
</tbody>
</table>

Figure 1. Traditional presentation of statistical error in null hypothesis significance testing (NHST)

Thus, it is important for researchers to always be cognizant of these error rates and what affects them each and every time they perform an experiment. For the NHST to be considered a valid test for evaluating the effectiveness of a given treatment, the factors that determine statistical significance (e.g., sample size, experimental control, in addition to effect size) must be acknowledged and closely regulated.

WHAT DOES THE NULL HYPOTHESIS REALLY MEAN?

One of the most common mistakes made by researchers stems from a misinterpretation of what the null hypothesis actually means. As discussed by Fisher (1935, p. 16) in his original definition: “In relation to any experiment we may speak of this hypothesis as the “null hypothesis,” and it should be noted that the null hypothesis is never proved or established, but is possibly disproved, in the course of experimentation. Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis.” Unfortunately, retaining the null hypothesis often leads researchers to conclude that “the mean values being compared are equal”, or “there is no relationship between variables X and Y.” Such interpretations are incorrect. Retaining the null hypothesis simply means that, based on the data provided, there is not enough evidence to suggest that the mean values are different, or that there is a relationship between variables X and Y. This is akin to concluding that there might be mean differences or a relationship between X and Y, but the evidence provided by the data is not substantial enough to draw this conclusion for a given confidence level. Severe abuse of the null hypothesis often occurs in experiments where the researchers’ goal is to show equality. For example, researchers may be interested in showing that the benefits provided by two different resistance training programs are equal. So, they collect the data and compare the mean values from the two different programs [e.g., with a t-test or an analysis of variance (ANOVA)]. The result of the statistical test(s) leads the researchers to retain the null hypothesis and conclude that the two resistance training programs are the same in terms of their outcomes. This conclusion is incorrect! It may very well be that the two programs provide the same benefits, but the NHST is not the test to evaluate this! A different test should be used because the NHST is not capable of proving equality, only inequality (with a given confidence level, of course). As discussed thoroughly by Altman and Bland (1995), absence of evidence is not evidence of absence.

It is important to point out that misinterpretation of the null hypothesis is not a new concept. Altman and Bland (1995), Cohen (1990, 1994), Thompson (1992), and many others have discussed this issue. The famous statistician John Tukey (1991, p. 100) wrote: “It is foolish to ask ‘Are the effects of A and B different?’ They are always different – for some decimal place.” The fact that ANY mean difference can be made “significant” with a large enough sample size illustrates a fatal flaw of NHST, and one that many critics cling to. Thompson (1992, p. 436) commented on this issue by stating: “Statistical significance testing can involve a tautological logic in which tired researchers, having collected data on hundreds of subjects, then conduct a statistical test to evaluate whether or not there were a lot of subjects, which the researchers already know, because they collected the data and know they are tired. This tautology has created considerable damage as regards the cumulation of knowledge.”

An unfortunate reality is that misinterpretation of the null hypothesis and the limitations of NHST in general has influenced science heavily. It is not at all uncommon for preference to be given to manuscripts that find “significant” results. Or, even worse, manuscripts are rejected because they find “non-significant” results. This has indirectly led to
considerable statistical mischief amongst researchers in an effort to get their work published. As a side note, it is hard to blame them when tenure and promotion, salary increases, funding opportunities, and professional notoriety in general are judged, to a certain extent, on number of publications. And if preference is going to be given to those studies with “significant” results, then researchers must either resort to methods for improving their chances for “significance” (some of which are completely unethical), or be content with having their work rejected on a somewhat regular basis. At this point, it is important to acknowledge the fact that the conflict between using statistics as a tool to advance science and doing what is necessary “just to keep your job” is not new, and has even spurred the development of journals dedicated specifically to studies that fail to find statistical significance. These journals include the Journal of Negative Results, Nonsignificance, the Journal of Non-Significant Results, and the Journal of Articles in Support of the Null Hypothesis, among others. Nevertheless, continued ignorance by reviewers and journal editors regarding the limitations of NHST has initiated the development of a number of different data tampering methods. As intelligent consumers of the literature, it is important to be able to identify these tricks to reduce the spread of misinformation and promote the advancement of legitimate science.

DECEIVING GRAPHS

Two very common tricks used by researchers to make graphs deceiving are: (1) inappropriate scaling of the y-axis, and (2) use of error bars based on the standard error of the mean (SEM), rather than the standard deviation (SD) (Huff 1954). For example, Figures 2a and 2b show data from a pre vs. post control group intervention study.

With just a cursory inspection, it would appear that the treatment effect shown in Figure 2b is much larger than that in Figure 2a. However, the data contained in both graphs are exactly the same. The only difference is that the y-axis in Figure 2b has been scaled inappropriately to make the differences between the treatment and placebo groups appear to be larger than what they actually are. If the reader was attentive to the fact that these differences are actually minimal, then he/she would likely interpret the data differently. However, the author(s) should also not be attempting to deceive the reader. It is important to point out that the use of deceiving graphs is prominent not just in medical research, but also in other fields, such as engineering. In a field where critical decisions are based on calculations and interpretations, researchers must always strive to present the data in such a way that reflects its true technical quality (Drummond and Vowler 2011), as opposed to what “looks the best”.

A similar strategy has been used by researchers to make the variability in data appear to be small. Using error bars based on the SEM, rather than the SD, makes the variability appear to be less, even though there is little justification for performing this step. When asked why they used the SEM, rather than the SD, it is not uncommon for researchers to state something like “it looks better” or something similar. This is hardly a reasonable justification and many times is indicative of the confusion that exists regarding the information contained in the SD and the SEM (Altman and Bland 2005). By definition, the SD is the average deviation of the scores away from the mean value. The SEM, however, is an estimate of the amount of error that can occur when the mean value from a random sample is used as a predictor of the mean from the population from which the sample was drawn. Thus, even though the SEM is derived from the SD (i.e., SEM = SD/√N), the reader should not be distracted by the fact that the two measures contain different information. The SD reflects the actual variability observed in the sample, whereas the SEM is a probability-based estimate of the proximity to the true population mean. If this is the information that the reader would like to convey with the SEM, then it should certainly be used. However, using it to
decrease the sizes of error bars in an effort to appease reviewers is not justified.

PERCENT CHANGE
Transformation of data into units that can be universally understood is a useful practice in situations where the units are confusing and/or there is excessive between-subjects variability. In many cases, the data are transformed into units that represent percent change away from baseline. Sometimes, this transformation is justified. But, there are many other times where it is not. The transformation of data into change score or percentage change units has limitations that have been acknowledged by several authors (Cronbach and Furby 1970; Vickers 2001; Kaiser 1989; Törnqvist et al. 1985). A very important, and sometimes misunderstood limitation of transforming data to percent change is that it places restrictions on the data. In particular, transformation to percent change assumes that if the scores for any given variable are small, then the amount of change that occurs away from the baseline will also be small. Similarly, if the scores are large, then the change away from baseline will also be large. Furthermore, the baseline value plays just as big of a role in determining percentage change as does the actual change away from the baseline. Two individuals that demonstrate the same absolute change away from baseline, but have different baseline values, will consequently have different percentage changes. Is one of these individuals “better” simply because he/she started at a lower baseline value? If not, then why is percentage change being used, because it completely transforms the data and gives the impression that the adaptations from the two individuals were different? For example, Figure 3 shows the results from two subjects that participated in a basic strength training research study.

Subject 1 begins the study at a strength level of 100 lbs. and improves to 110 lbs. after the training, resulting in a 10% improvement. Subject 2 starts the study at a strength level of 200 lbs. and improves to 210 lbs. after the same training, resulting in only a 5% improvement. Does it make any sense to conclude that the training was more effective for subject 1? We would argue that it does not, because both subjects improved their strength level by the same absolute amount, but subject 2 is being penalized because they started the study at a higher strength level. A similar example can be seen with body fat percentage. Subject 1 begins with a body fat percentage of 30% and is able to reduce it to 20% with diet and exercise, resulting in a 33% change. Subject 2 begins with a body fat percentage of 20% and is able to reduce it to 10% with diet and exercise, resulting in a 50% change. Should we conclude that the diet and exercise used by subject 2 was more effective at reducing body fat percentage? Certainly not, since both subjects demonstrated the same absolute change in body fat percentage, and the only reason that subject 2 appeared to improve more based on percent change is because they started out at a lower % fat value. This illustrates a dangerous, and often under-appreciated consequence of transforming data to percent change units. The magnitude of the values that get reported is dictated not only by the size of the change, but ALSO by the size of the baseline value. All participants must start at the same baseline value in order for percent change to be useful.

INAPPROPRIATE SAMPLING
The selection of research subjects that are not representative of the population of interest is an unfortunate strategy that can be used to deceive readers that do not understand proper sampling (Huff 1954). For those involved in research, it seems quite obvious that the sample used by the investigators should be representative of the population of interest.
After all, we as researchers cannot rightfully state that the results from a given group of individuals apply to those that are different from them in some critical characteristic. As elementary as this sounds, this general rule often gets broken (sometimes intentionally) by researchers. For example, a strength and conditioning researcher might be interested in assessing the effectiveness of a new training methodology for improving the strength of Division I American Football players. It seems logical, then, that the researcher should select a random sample of Division I American Football players. Unfortunately, this is easier said than done (for a variety of reasons), and, as a consequence, the researcher ends up using a sample that is convenient, but not very representative of Division I American Football players. Another example occurs in the context of advertisements for nutritional products/strategies. Specifically, these products/strategies can, in some cases, be tested on underweight/malnourished individuals, for which any nutrition will be beneficial, regardless of whether or not it is some “special” type of nutrition.

APPROACHING SIGNIFICANCE AND HIGHLY SIGNIFICANT

The phrases “approaching significance” and “highly significant” are used much too frequently in research, and illustrate a general misunderstanding of what statistical significance testing is, and what it is not (Carver 1978; Schafer 1993; Shaver 1993). All too often, researchers will set their Type I error rate (e.g., \( p = 0.05 \)), and then interpret a significance level of, for example, 0.06 as “approaching significance”, and a significance level of, for example, 0.001 as “highly significant”. Their argument for justifying their interpretation of the former situation is that if their treatment effect would have been just a little bit larger, then it would have met or exceeded the 0.05 standard and been considered “statistically significant”. Similarly, their justification for their interpretation of the latter was that the very low significance level must have been due to a large treatment effect. While both of these arguments can be true, they also ignore the fact that the magnitude of the treatment effect is not the only factor that determines the significance level. Namely, the sample size and variability in the sample are just as important. The fact that ANY treatment effect, no matter how miniscule or meaningless, can be made statistically significant, or even “highly significant” with a large enough sample size (see Figure 4) or low enough variability is illustrative of the problem associated with using these phrases.

What’s even worse is when the researcher(s) interpret the results that “approached significance” in the same context as if they had actually achieved significance. Such a strategy abuses the NHST to the point where you have to question why it is even being done, and illustrates a general misunderstanding of what significance actually means (Lew 2012). After all, if \( p = 0.06 \) is approaching significance, and we can interpret that the same way we would if \( p = 0.04 \), then maybe \( p = 0.07 \) is approaching significance too, and maybe \( p = 0.08 \) as well? The point is that there has to be a cutoff set, and that cutoff should ALWAYS be set before the study. The argument that “these results would have been statistically significant if the treatment effect had been slightly larger” is akin to arguing “this miniscule treatment effect would have been statistically significant if the sample size had been larger.” Why should a statistically significant effect from 30 subjects be any more meaningful than an equivalent, but not statistically significant effect from 20 subjects?

The fact that the significance level is affected by both sample size and the magnitude of the treatment effect also illustrates the importance of doing a priori
sample size estimations. When these estimations are performed correctly, then there is absolutely no reason to hedge the significance level. Establishing an effect size that provides a meaningful difference allows the researchers to estimate a sample size that will provide statistical significance, given, of course, that a “real” treatment effect is observed. Sample sizes that are less than this estimation will generally be underpowered, while larger sample sizes will be overpowered. This is the appropriate application of sample size estimation and the NHST. A priori estimation of sample size removes any need to hedge the interpretation of the results from the NHST.

UNDERPOWERED SIGNIFICANCE TESTS

The information presented in the previous section leads quite well into this section because, as stated previously, statistical power, sample size, and the significance level are all inter-related. Thus, achieving an appropriate level of power is dictated mostly by sample size, since the significance level is usually fixed (e.g., at 0.05). It is an unfortunate reality that a priori sample size estimations are often not performed, and, as a result, many research studies are underpowered. This has been well documented in the field of educational psychology (Cohen 1962, Rossi 1990), but is also quite prevalent in medical research (Tressoldi et al. 2013; Brody et al. 2013; Bouwmeester et al. 2012; Bacchetti 2010; Button et al. In Press). With underpowered studies, the researcher(s) must try to decipher if the effect observed in their study was truly not significant, or if they had committed a Type II error due to a small sample size. When one considers the amount of time and effort that is placed into research, it is a bit disheartening to consider the fact that these underpowered studies might have a 50% chance (or even less) of the treatment effect being statistically significant. A strategy that is sometimes used to combat this is to measure a lot of variables with the hope that “something will turn up significant”. This strategy is obviously not acceptable, as it completely ignores the compounded Type I error rate when many significance tests are performed. For the often used 5% Type I error rate, one can expect that one out of every twenty tests will be significant based on chance alone. Additionally, this strategy calls into serious question why the research is even being done in the first place. Are the investigators trying to advance science with novel empirical based evidence? Or, are they just trying to get published at any cost?

STATISTICAL VS. CLINICAL SIGNIFICANCE

When used properly, clinical significance can be a very useful supplement to statistical significance. Briefly, clinical significance refers to a treatment effect that has clinical importance, even if it is not statistically significant (Fethney 2010). Of course, an effect can be both clinically and statistically significant. And, in a perfect situation, the sample size would be large enough such that if the treatment effect was clinically significant, it would also be statistically significant. However, these perfect sample sizes are sometimes hard to achieve, particularly with some clinical populations. Thus, “clinical significance” can be a useful tool for researchers to use in situations where statistical significance alone will not suffice. Unfortunately, use of clinical significance also gets abused by researchers that want to justify trivial treatment effects with small sample sizes as still being important. For example, it is common for researchers to state something to the effect of: “although these findings were not statistically significant, the change in the mean values could be considered clinically significant”. In some regards, “clinical significance” is similar to the previously mentioned “approaching significance” issue, in the sense that both provide an alternative to the researcher when they feel that they have a real treatment effect that is not statistically significant. The debate of statistical vs. clinical significance is certainly not new, and it is not the purpose of this section to review them comprehensively. However, we will try to dispel some myths regarding clinical significance, as well as describe some situations where its use is acceptable.

First of all, if the researcher(s) is going to use the term “clinical significance” as a part of their interpretation, then they must provide a description for the minimum difference that they would consider clinically significant. Furthermore, justification for considering this difference to be clinically significant with evidence from the researcher’s discipline must be provided. It is certainly not appropriate for a researcher(s) to deem results clinically significant simply because they find them interesting. Second, the researcher(s) must provide some sort of legitimate justification for why they had a limited sample size.
The primary criticism of the term “clinical significance” is that if the treatment effect was, in fact, clinically meaningful, then the researchers should have used a large enough sample size to provide the power necessary to prove that this effect is greater than what can be expected by chance alone. Any resistance to this argument illustrates an under-appreciation for the fact that NHST is a numbers game, with the final decision being dependent on having the right combination of numbers (i.e., large enough effect size and sample size, and low enough variability). It could even be argued that “clinical significance” perpetuates underpowered studies because it provides researchers with an excuse for why their experiment failed to find a treatment effect. Furthermore, if the researcher(s) is going to base their interpretation on “clinical significance”, then why even perform any statistical significance testing at all? Why not just base the conclusions on a subjective interpretation of what is “clinically significant”?

To be fair, it is important to point out some situations where the use of clinical significance is justified. For example, studies that have power issues (either overpowered or underpowered) can benefit from clinical significance because it provides the researcher(s) with a frame of reference that can be used when the information provided by the NHST is suspect. For example, overpowered studies might have trivial (but statistically significant) changes that should be examined more carefully to determine if they have clinical relevance. Similarly, underpowered investigations that show large (but not statistically significant) changes should be evaluated more closely to see if their magnitude is clinically important. In addition, the instances where clinical significance is used should have clearly defined standards for what is deemed a “clinical” change, and, if possible, it should be used in conjunction with NHST.

**INAPPROPRIATE INTERPRETATION OF INTERACTIONS AND MAIN EFFECTS**

The development of widely available personal computers and statistical software packages has allowed researchers to use increasingly complex research designs that reveal new and interesting information. An unintended consequence of these developments is that researchers can sometimes lose intimate contact with their data and become confused with the analysis. For example, 2-way factorial ANOVAs (both within-within and between-within) are very common in biomedical research. An understanding of the possible outcomes of these analyses, and how to properly follow them up, is a bit less common. Typical issues include interpretation of significant main effects when the interaction is also significant, interpretation of a significant interaction when the simple effects are not significant, and understanding the factor(s) that can cause the various types of interactions. We will now discuss each of these problems and how to handle them.

It is fairly common for factorial ANOVAs to result in a significant interaction and at least one significant main effect. The researcher must then decide which provides the best evidence of a treatment effect, if in fact, they are confident that the treatment effect actually exists. In a lot of cases, researchers want rules that can be applied to help them handle complex statistical problems. And, generally speaking, the main effects are less meaningful when there is a significant interaction (Keppel 1991). However, the nature of the interaction and what causes it are also important factors that should not be overlooked. For example, an interaction that is entirely, or even just partially, a function of baseline differences should be interpreted differently than one where there are no baseline differences, and the interaction is due to a treatment effect. Similarly, a main effect that occurs in the presence of an interaction that is not quite statistically significant should be interpreted with an equal amount of caution.

Another relatively common occurrence involves finding a statistically significant interaction with no significant pairwise differences (e.g., simple effects). This is another situation where the researcher(s) must be careful with their interpretation. Apparent evidence of a treatment effect (due to the significant interaction) with no pairwise differences (due to non-significant post hoc tests) can make it very tempting for researchers to use a variety of different post hoc tests until they find one that provides statistical significance. This is obviously not an appropriate strategy, even though the alternative places the researcher(s) in a bit of a conundrum. However, such is the drawback of NHST. These tests are excellent with large sample sizes and big effect sizes. But,
when one (or both) is compromised, then statistical significance, and hence, the researcher’s interpretation, becomes a numbers game. Although it may be difficult for the researcher(s) to “hedge their bet”, such is the nature of research. Inconclusive evidence will always be inconclusive, no matter what post hoc test is used. Such should be the researchers’ interpretation.

**INAPPROPRIATE USE OF WITHIN-GROUP COMPARISONS**

The 2 × 2 between-within design is commonly used in research and is most appropriately analyzed with a 2-way ANOVA. However, researchers sometimes inappropriately analyze data from this design by performing only within-subjects comparisons. This is, in effect, breaking down a multiple-factor problem into several separate, single-factor problems. In cases where the comparison for the treatment group is significant and that for the control group is not, then the researchers might conclude that there was a significant treatment effect. Is this the correct conclusion? Despite its prevalence, this practice is absolutely not acceptable, since it increases the Type I error rate to as much as 50% for two groups, and 75% for three groups (Bland and Altman 2011). Why then, is this practice so prevalent? As discussed by Bland and Altman (2011), the most logical answer is that researchers blindly copy the methods from other studies without a real understanding of what they are doing or the limitations of the tests that they are performing. This blind acceptance of methods can be dangerous, and should be discouraged to prevent the drawing of conclusions that are not necessarily true. As discussed previously, the correct approach for this design is the 2-way ANOVA.

"SIGNIFICANT" CHANGE VERSUS "MEANINGFUL" CHANGE

When a biological variable is measured multiple times, changes in the scores for any given subject can be due to three sources: (1) measurement error, (2) biological variability, and (3) variability due to the effect of a treatment imposed by the investigator. In a perfect world, measurement error and biological variability are minimal, and the treatment effects that we generate with our studies are large. As such, “statistically significant” changes are also “meaningful” changes, and the statistics make our jobs as researchers very easy. The “yes” versus “no” decision provided by the NHST matches up with the clinically meaningful changes, and all is good with the world. Unfortunately, as we all know, the research world is far from perfect, and this imperfection can create conflict when we try to interpret data. The primary problem comes when the biological variability or the measurement error (or both) are large in relation to the change that is expected to come from a “real” treatment effect. Underestimation or complete disregard for the magnitude of these errors can create misperceptions on the efficacy of a given treatment, as well as increase the rate of inappropriate diagnoses in clinical practice. Consider, as an example, the data shown in Table 1.

![Table 1. Data for a blood chemistry variable (IU/L) measured from three separate groups of subjects before (PRE) and after (POST) an intervention. All values are means ± SD](image)

<table>
<thead>
<tr>
<th>Group</th>
<th>PRE</th>
<th>POST</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>28.04 ± 7.50</td>
<td>28.36 ± 6.85</td>
</tr>
<tr>
<td>2</td>
<td>29.61 ± 9.42</td>
<td>27.61 ± 8.02</td>
</tr>
<tr>
<td>3</td>
<td>34.69 ± 10.83</td>
<td>31.11 ± 9.97</td>
</tr>
</tbody>
</table>

These data represent the mean ± SD values for a blood chemistry variable (IU/L) measured from three separate groups of subjects before (PRE) and after (POST) an intervention. The appropriate analysis for this data set is a two-way (group × time) mixed factorial ANOVA, with group membership as the between-subjects factor, and time as the within-subjects factor. The result of this ANOVA is a statistically significant (p ≤ 0.05) group × time interaction. Appropriate follow-up analyses include one-way ANOVAs across groups at the PRE and POST time points, as well as paired-samples t-tests to compare the PRE versus POST scores for each group. The result of the follow-up analyses indicated that there were no mean differences (p > 0.05) among the groups at either the PRE or POST time points, and the change from PRE to POST for group 1 was not statistically significant (p > 0.05). Groups 2 and 3, however, both demonstrated significant (p ≤ 0.05) decreases from PRE to POST. A tempting conclusion, then, based on the results from these statistical analyses, would be that the intervention imposed on groups 2 and 3 significantly decreased this particular blood chemistry variable, with changes in the mean values of 2.00 and 3.58 IU/L, respectively.
Now consider that the measurement error for this blood chemistry variable is particularly high, with a least significant change of 6.65 IU/L. The mean changes of 2.00 and 3.58 IU/L are clearly within the measurement error for this technique, despite the fact that they were considered statistically significant. This simple example demonstrates the extreme importance of calculating reliability statistics for each dependent variable that is measured in a study. Many journals are beginning to require that authors provide reliability information for their dependent variables, and Weir (2005) provided an excellent guide that can be used for assessing reliability. However, reliability assessment must be built into the design of the study, such that multiple baseline measurements are made before the intervention is imposed. For some variables, it is not necessary to assess reliability every single time that a study is performed. For example, a laboratory may assess reliability for a given strength measurement, and that reliability information can be used for multiple studies, given, of course, that the testing procedures were the same for all studies. For other variables, however, reliability must be assessed for every study. A good example of this situation is for blood chemistry variables (like the one discussed above) that are examined with assay kits. Even when the instructions provided by the manufacturer are followed very closely, the measurement error from using these kits is relatively high. Without reliability information, and determining the least significant change, researchers run a high risk of committing a Type I error. Their assumption is that any given value is close to the true value for that individual, when, in fact, it could easily be far from it due to measurement error. The same logic also applies to variables that have high biological variability. For example, the maximal oxygen consumption rate (VO₂max), which is one of the most commonly measured variables in Exercise Physiology, has a trial-to-trial variability of as much as 5.6%, with biological variability accounting for at least 90% of this error, and measurement error accounting for less than 10% (Katch et al. 1982). Regardless of the source (biological variability versus measurement error), however, the end result is the same in the sense that it reduces reliability and increases the Type I error rate.

CONCLUSIONS
This paper has described 11 common misconceptions and techniques that mislead (sometimes unknowingly) readers of scientific journal articles. Whether or not these mistakes are intentional or unintentional, the end result is the same in the sense that it can lead to false conclusions in research, errors in clinical practice, and, potentially, misdiagnoses. This paper was by no means intended to be a comprehensive listing of all misconceptions. Instead, we have simply tried to describe some of the more common mistakes that we have seen in the sports and health literature. It is important to point out that there is no uniform solution to these problems. Thus, researchers seeking a “cookbook” approach to statistical analysis will likely be disappointed. However, many of these problems could be avoided, or at least minimized, by performing a priori sample size estimations, building reliability assessments into the study design, and using effect sizes and confidence intervals in conjunction with the results from the NHST. A clear understanding of the limitations of the NHST is also very important. In the end, researchers should always remember that statistics is a numbers game, and the numbers can be pushed in any given direction to distort the truth. However, a hallmark of good research is when the numbers reflect the truth.

REFERENCES
4. Bland, J.M., & Altman, D.G. (2011). Comparisons against baseline within randomized groups are often used and can be highly misleading. Trials 12:264-270.


29. Rossi, J.S. (1990). Statistical power of psychological research: what have we gained in


