Journal of Research Practice
Volume 8, Issue 1, Article P2, 2012

ISSN: 1712-851X

Provocative Idea:

Statistical Reporting with Philip's Sextuple and Extended Sextuple: A Simple Method for Easy Communication of Findings

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

Abstract

The advance of science and human knowledge is impeded by misunderstandings of various statistics, insufficient reporting of findings, and the use of numerous standardized and non-standardized presentations of essentially identical information. Communication with journalists and the public is hindered by the failure to present statistics that are easy for non-scientists to interpret as well as by use of the word *significant*, which in scientific English does not carry the meaning of "important" or "large." This article promotes a new standard method for reporting two-group and two-variable statistics that can enhance the presentation of relevant information, increase understanding of findings, and replace the current presentations of two-group ANOVA, *t*-tests, correlations, chi-squares, and *z*-tests of proportions. A brief call to highly restrict the publication of risk ratios, odds ratios, and relative increase in risk percentages is also made, since these statistics appear to provide no useful scientific information regarding the magnitude of findings.

Index Terms: research education; research reporting & publishing; statistical reporting; research method; experimental design; quasi-experimental design; statistical analysis; statistical significance; effect size; risk ratio; odds ratio

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

Even now, in the twenty-first century, different authors report essentially identical—and oftentimes quite limited—information in numerous ways. This haphazard approach

hinders communication when all parties to the information do not fully understand one or more of the presentations, and can easily lead to misunderstandings when people with insufficient training attempt to interpret, but in fact misinterpret, the information. Consider, for example, the following three statistical results where we will assume each presentation is from a different two-group study on the same topic where there was a single statistical test in each study (alpha = .05). Study #1 found ANOVA f (1, 98) = 13.02, $p < .05$; Study #2 reported a Pearson correlation r (100) = .34, $p < .01$; Study #3 presented the result of a Student t-test, t (98) = 3.61, $p < .001$.

Can the vast majority of PhDs and scientific journalists look over these results and correctly determine which study found the largest association, which found the smallest, and which studies, if any, found an identical association? In this particular set of examples all of the findings on the magnitude of the association are identical. All of these analyses were derived from the same dataset and present equivalent information.[1]

The above examples provide relatively typical information as presented in contemporary research reports: they contain sufficient information for a knowledgeable scientist to determine the total N for the study as well as to calculate a measure of the magnitude of the association between groups (i.e., an *effect size*). However, it should not be the responsibility of the reader to know how to, and to be required to, calculate these values—such information should be clearly presented in research reports. Furthermore, many research reports fail to provide information many readers will want in order to usefully interpret the results of the reported study (e.g., the sizes of the separate groups rather than their combined total; effect size measures that depend on, or are independent of, these n values). Thus there is a need for improved reporting of research results in order to encourage or ensure complete reporting, and in order to report information in a way that is easy to compare and contrast across different studies. This should also make it easier for people to conduct meta-analyses—and make those analyses more complete and accurate—since today meta-analysts are often faced with either omitting some studies from their analysis due to missing information or to using effect size estimation procedures, rather than more precise calculations, due to limited information.

This article focuses on the reporting of research results with regard to statistical values that are or should be commonly reported across a wide range of social science research. The sextuple recommended below is not offered as a replacement of all statistical values that would typically be reported, rather, it is offered as a minimum set of information to report, in a standardized way that is easy for people, including non-scientists, to understand.

1. Information Needed for Two-Group Studies

Before recommending how to standardize the presentation of two-group analyses, it is important to determine what information is usually needed for proper scientific interpretation of the findings from two-group studies.

First, the most basic scientific question is: "Is the difference found between the samples larger than what might be expected due to purely chance variations?" That is, are the variables *statistically significantly* related? Note that this question is not asking if the finding is large, rather, it is asking if the finding (even if small) is larger than would be expected by chance alone. It has become a common practice in social science research to address this question by presenting *p*-values; *p*-values report the percent chance, expressed in decimal form (e.g., 4% is expressed as .04), of obtaining the presented finding, or a larger one, purely by chance. It appears, $p < .05$, $p < .01$, and $p < .001$ have come to be accepted as useful ways of reporting in response to this basic scientific question.

However, as types of analyses have increased, the importance of meta-analysis has been established, and larger samples that tend to produce small *p*-values have become more common, these imprecise, open-ended ranges should be replaced with exact *p*-values where practical. This may not be a critical improvement in scientific reporting, but it should pose no impediment and since the exact *p*-values provide more information than the ranges, they have the potential to improve scientific understanding. Indeed, researchers conducting meta-analyses may benefit when faced with reports that lack a clear effect size measure since the exact *p*-value may be usable to calculate an appropriate effect size statistic. The precise values may allow readers to perform Bonferroni correction to results when authors did not use this procedure (see below and Endnote 2) and this recommendation is also consistent with that made by the American Psychological Association (APA) Publications and Communications Board Working Group on Journal Article Reporting Standards (APA-JARS, 2008, p. 843, Table 1) as well as the newest APA publication manual (APA, 2009).

Second, a potential problem in science, particularly in the assessment of literature reviews—both articles that are reviews of the literature as well as the introduction that appears at the start of most primary research articles—is that isolated findings that appear to support a person's hypothesis may get cited as important findings when they are more likely to be the result of chance variation. When a finding is statistically significant but in fact was the result of chance variation, it is known as a *Type I error*. Since any finding may be the result of Type I error, the chance of there being Type I errors increases as more statistical tests are conducted. For example, if the chance of Type I error for a single test is 5% and there is only one test, then the overall chance of Type I error is 5%. However, if two statistical tests are conducted where each has a 5% chance of Type I error the overall chance of there being a Type I error rises to 9.75%; if there are 20 tests the overall chance of there being one or more Type I errors exceeds 64%.

To minimize the Type I error problem primary studies sometimes employ some form of correction procedure, for example, Bonferroni correction, but this information is rarely included in the brief presentations that appear in literature reviews. Thus, for example, a finding of $p = .0009$ may be presented in a literature review as "highly significant," but without going back to the original source one may not realize that the Bonferroni correction indicated that $p < .001$ indicated a significant difference between groups at the

$p < .05$ level; thus leading to a misinterpretation of the level of significance that was found.[2]

A better approach, in this author's view, is to always report the number of statistical tests that were or could have been performed, together with the number that were found to be statistically significant (using a threshold of 5%, i.e., alpha = .05). For example, a finding of $p = .001$ is far less important if it was the only 1 out of 20 tests that was statistically significant (since we expect 1 out of 20 tests to be statistically significant due to chance variation—Type I error—when alpha = .05), whereas this same value ($p = .001$) is of greater scientific import if 18 out of 20 tests were statistically significant (since the chance that all 18 statistically significant results were due to Type I error is small).

Thus, the second piece of information that a reader of a scientific report needs (after the *p*-value) is the ratio of the number of statistically significant findings from the study being presented and the total number of tests that could have been performed.[3]

The above two pieces of information (i.e., the number of statistically significant findings and the number of possible tests which could have been conducted) only tell the reader if a finding appears to be different from chance and provide a guide to assessing possible Type I error. The second major question scientists ask of their data is: "For a finding that is statistically significant, how large is the finding?"[4] That is, what is the magnitude of the association between group membership and the dependent variable? The answer to this question is complicated by the fact that many measures of association incorporate base-rate information (e.g., if one were to assess the difference in average IQ in a group of 30 compared to a group of 70 people, the result would not necessarily exactly equal the association for the same average difference when comparing two groups of 50 people). The question regarding the magnitude of the association, therefore, generally has two answers, or more precisely, there are two versions of the question, leading to the third and fourth pieces of information that are needed for a fuller understanding of the findings from two-group studies. The first version is: "In the real world, where group sizes (i.e., base-rates) may not be equal, to what degree are the variables related?" Worded in another way, this question becomes: "What is the *estimated association* (EA) for the population?" The second version is: "In the laboratory or other artificial setting, where there are an equal number of people in each group, how large is the difference between groups?" Worded in another way, it becomes: "What is the *group difference* (GD)?"

The real-world, unequal base-rate question (i.e., the EA) is best answered with a correlational statistic. Indeed, such statistics have been the primary method of reporting such findings (e.g., the use of the Pearson *r*, Student's *t*, ANOVA *f*, etc.). Importantly, correlational statistics generally incorporate base-rate information. Hence for the purpose of standardizing the presentation of statistical analyses we need to select a single preferred correlation-family statistic to represent the EA. The choice this author recommends is r^2 expressed as a percentage.

The r^2 value is the logical choice for many reasons. First and foremost, it has the property of equal intervals. Centimeters, for example, have this property. The magnitude change

going from 3 cm to 4 cm is the same as the magnitude change going from 15 cm to 16 cm. A 1 cm change indicates an identical magnitude change regardless of where it occurs. The magnitude change from $r = .1$ to $r = .2$, however, is not the same as the magnitude change from $r = .2$ to $r = .3$; thus the Pearson correlation is not easy for people to understand, and can even lead scientists astray in their data interpretation.[5] Since r^2 has equal intervals, it is a more natural choice for the EA. Additionally, when expressed as a percentage, the r^2 value has a range of 0 to 100, which is easy for people to remember and leads to easy interpretation (see Section 5 on describing the size of the association). Furthermore, r^2 represents the percentage of variance in one variable that is explained by the other variable, which is relatively easy to understand and conceptually what we want—a measure of the relatedness of the variables in the real world where 0% indicates no relationship and 100% indicates that if you know the value of one variable you can precisely calculate the other variable. Thus, this author recommends adopting $r^2$for the EA, expressed as a percentage.

Please note, because many analyses are conducted in order to gain knowledge that will be generalized to the human population thought of dichotomously (e.g., males compared with females; people who exercise regularly compared to those who do not), when practical the EA should reflect the true population split (rather than the split in the sample used for data collection). In other analyses this will not be possible, for example, when comparing two of three mutually exclusive groups (e.g., people with a body mass index below 25—not overweight, those in the 25 to 30 range—overweight, and those over 30—obese, in an analysis comparing only two of these groups). Thus if an r^2 is calculated using equal group sizes, as is common in an experiment where the true population base-rates are not equal (e.g., measuring the association between smoking and health problems where 50 smokers are compared to 50 non-smokers, drawn from a population where only 1 in 10 people are smokers), ther^2 will need to be adjusted to better estimate the association in the population with the EA.

It should also be noted that the correlational statistics just discussed operate under an assumption that the relationship being measured is a linear relationship. Just as it would be inappropriate to use a Pearson *r* as a measure of a curvilinear relationship, use of r^2 as suggested above similarly should not be misapplied to curvilinear relationships.

The equally-sized groups measure of association (i.e., the GD) needs to be a measure of the difference between groups, expressed in a standardized way; that is, a standardized mean difference such as Cohen's *d*, Hedges *G*, or Glass's delta. This author recommends that the measure be derived from the mean of one group minus the mean of the other group, divided by the standard deviation of the combined data.[6] In most social science research *d* is a logical choice, except that it calls for some level of statistical proficiency in order to be interpreted. A practical and easy to interpret method to communicate group differences is suggested below.

It is suggested that group differences be expressed using a measure that has a mean of 100 and a standard deviation of 15, much like intelligence quotient (IQ) scores such as the Wechsler Adult Intelligence Scale. Consequently, if we multiply *d* by 15 to create the

GD measure, we have a measure of the difference between groups that is expressed in the same units as IQ when measured by the well-known Wechsler IQ test, but one that can be used for almost any variable of interest (e.g., self-esteem, happiness, annual salary). This author refers to this scale as the IQEP scale (IQ Equivalent Points). Using IQEP difference scores also leads to easy-to-remember values for what constitutes small, medium, and large findings (see Section 5 on describing the size of the association), thus this author recommends adopting IQEP difference values for the GD measure.

Fifth and sixth, with the questions of statistical significance (i.e., Is the found difference unlikely to be due to chance? Is that difference likely to be Type I error?) and the questions of practical significance (i.e., How big is the EA, the estimated association between the variables in the real world? What is the GD, the average difference between groups?) addressed, the next important question is: "Was the study large enough to trust that the statistics are stable and reasonable estimates for the population from which the sample was drawn?" In the past, the answer to this question has often been misleading. For decades scientists have been focusing on large studies and reporting the total N for an analysis, but does knowing the total N really give us the information we need?

Consider two hypothetical studies that report identical IQEP difference values, where N = 2000 in both studies. If in one study there were 1000 people in the exposure group but in the other study there were only 20 people in the exposure group, we would properly put more trust in the stability and accuracy of the measures in the first study; we might put very little trust in the estimates of a study where there were only 20 people in one of the groups even though the total N for the two groups combined was 2000. Thus, although total N is important for many statistical issues and calculations, when assessing findings, this author suggests that it is critically important to know the size of the smallest group in the analysis. In two-group data there are only two sizes, hence this author recommends providing them both so the size of the smallest group is clear and the total N can be quickly calculated, if needed. For standardizing the presentation of control-group designs, the size of the control group should be listed first since that group's data provide the base for comparison.

2. Six Pieces of Information: Philip's Sextuple

This author recommends that the aforementioned six pieces of information be arranged into a sextuple as follows: (1) the exact *p*-value up to four decimal places if it is available, or one of the traditional ranges (e.g., "<.05") if an exact value is not available or if the first four decimal places are all zeros, (2) the count of statistically significant findings and the count of the number of possible statistical tests at that level (see Endnote 3), separated by a slash ("/"), (3) the estimated association (EA), which is the r^2 value expressed as a percentage (e.g., "9.0" or "9.0%" when $r = .30$), using only one decimal place; if this value reflects the base-rates in the population as a whole (rather than the base-rates in the study sample) it should be underlined to indicate that it is a true EA, (4) the group difference (GD), which is the IQ Equivalent Points (IQEP) difference between groups, expressed with one decimal place, (5) the size of the control group, and (6) the size of the exposure group.

When any of the six pieces of information is not available, a question mark ("?") should be used to mark the missing information. This notation can draw attention to potentially important information that is lacking, minimizing the possibility of misinterpretation of the statistics that are presented. If these suggestions become accepted, this method of denoting missing information may also encourage better study design and reporting, since researchers are likely to prefer to conduct studies and analyses that will provide all of the recommended information.

When any of the six pieces of information is being estimated (e.g., due to incomplete or imprecise reporting in the primary study or limitations of the study design), a lowercase "c" (i.e., *circa*) should be appended to the front of the estimate (e.g., "c10.2").

For example, the sextuple for the various presentations in the first paragraph of this article would be written as: (0.0005, 1/1, 11.7%, 10.2, 50, 50). Alternatively, to increase clarity it could be written as: (p = 0.0005, 1 out of 1, EA = 11.7%, GD = 10.2, $n_{control}$ = 50, $n_{exposure}$ = 50).

3. Calculating the Statistics

Calculating the statistics for the sextuple requires very little effort beyond what researchers are already calculating. They will normally already have at least half of the information: the exact *p*-value and the size of the two groups. If they used any of the common tests for differences between groups (i.e., f, r, t, χ^2, or z), the calculated statistic can usually be converted to the EA using the formulas in Appendix A which have been adapted from those for calculating r presented by Rosenthal (1991) and Rosenthal and Rosnow (1991). The GD (i.e., IQEP difference value) can be calculated directly or from the group sizes, means, and standard deviations as shown in Appendix B.

4. Presenting Additional Information Using Philip's Extended Sextuple

The sextuple recommended here provides all of the most critical information typically provided in a primary or secondary report, and more. Nevertheless, there are cases where additional information should be presented to further elucidate the results of a study. There are three cases that will occur with sufficient frequency that their data presentation should be standardized along with the above.

Case 1. Reporting the group means on the dependent measure (e.g., the means for each group on the Beck Depression Inventory) and perhaps the standard deviations.

Case 2. Reporting the absolute "risk" or percent affected for each group when the dependent measure is dichotomous (e.g., the percentage of each group that has been diagnosed with clinical depression).

Both of the above cases involve presenting two additional statistics, and these two cases will not occur for the same sextuple. Thus, when this type of information is being presented an *extended sextuple* can be used where the seventh position reports the mean

or risk for the control group and the eighth position reports the mean or risk for the exposure group. To minimize confusion, percent risk values should always include the percent symbol ("%").

If an author wants to present both means and standard deviations, these paired values can be written together separated by a full colon (e.g., "95:15" would indicate a mean of 95 and standard deviation of 15).

Case 3. Reporting confidence intervals for the EA (i.e., r^2) and/or the GD (i.e., IQEP difference).

Over the past decade there has been increased interest in encouraging authors to report confidence intervals so that readers can better assess how precise a reported statistical value is likely to be (e.g., APA, 2009; APA-JARS, 2008; Des Jarlais, Lyles, Crepaz, & the TREND group, 2004). This is particularly important since most research is based on samples drawn from a larger population, hence the true population value is merely being estimated by the sample data—providing confidence intervals can help readers better assess likely upper and lower bounds on the true population values. To report 95% confidence intervals for the EA and the GD, that is, for "effect size"[7] statistics, the statistics can be enclosed in angle brackets with the confidence limits appearing before and after the brackets (e.g., if the GD is reported as "10<15<20" this indicates that the GD is 15 with a 95% confidence interval running from 10 to 20).

Since 95% confidence intervals are so common, this author recommends they be recognized as the de facto standard. If authors wish to report 99% confidence intervals these should encapsulate the 95% confidence intervals (e.g., "g<h<i<j<k" would indicate that *i* is the EA or GD with a 95% confidence interval running from *h* to *j*, and a 99% confidence interval running from *g* to *k*).

Thus, *Philip's extended sextuple* is merely the six pieces of information previously presented, plus the average score for each group, or alternatively, the percentage of each group that was affected or had the trait being measured. The seventh item corresponds to the fifth item (i.e., the number of people/items in the control group) and the eighth item presents the average or percent affected for the other group. If an average is followed by a colon (":") and another number, that value is the standard deviation.

5. Describing the Size of the Association

It is useful to have standard conventions for interpreting and describing the size of a finding. That is, it is useful to have commonly accepted qualitative descriptors (e.g., words such as *small*, *medium*, and *large*) for describing the quantitative findings from studies. Cohen (1988) successfully and suitably set the conventions for describing correlations: $r = .1$ indicates a small association, $r = .3$ indicates a medium association, and $r = .5$ indicates a large association. These recommendations have stood the test of time. This author only recommends a slight adjustment to one of these base values, and adds a clarification and extension as we go from reporting r to r^2.

Although not explicitly stated, this author believes Cohen intended his recommendations to be understood as thresholds that, if reached, lead us to use a particular term (as opposed to being the midpoint of the ranges implied by the descriptors). Regardless of his intent, that is what is recommended here. Thus, the values recommended below are intended to be interpreted as thresholds.

If we simply translate Cohen's suggestions directly, we see that the thresholds for the descriptors *small*, *medium*, and *large* for EA values would be 1%, 9%, and 25%, respectively. However, given the important goals of easy communication and of scientists and non-scientists being able to remember the thresholds, adjusting the 9% to 10% makes sense (this is equivalent to an $r = .3162$ rather than $r = .3000$). This author recommends further extending this approach to include two additional ranges: *very small* defined as less than 1%, and *very large* defined as 50% or greater (equivalent to an $r = .7071$ or larger). These ranges and descriptors are presented in Table 1.

Table 1. *Recommended Descriptors for Variously Sized Associations*

Descriptor	EA (i.e., r^2)	GD (i.e., IQEP difference)
very small	less than 1%	less than 3
small	1% up to 10%	3 up to 10
medium	10% up to 25%	10 up to 20
large	25% up to 50%	20 up to 30
very large	50% or more	30 or more

Although there are cases (i.e., extreme base-rate splits, e.g., when studying something that has a very low prevalence rate) where data properly indicate a medium or large GD together with a small or very small EA, when the base rates are equal the descriptors for EA and GD should match, with only minor discrepancies. If we translate the EA thresholds of 1%, 10%, 25%, and 50% directly to IQEP differences we would have the thresholds: 3.0, 10.0, 17.3, and 30.0. The 17.3 is the only problematic figure, which this author suggests be rounded up to 20.0 (equivalent to an $r = .5547$ rather than $r = .5000$) to create an easy to remember progression: 3, 10, 20, 30. These ranges are also displayed in Table 1.

It should be noted that Cohen's (1988) recommendations for interpreting standardized mean difference values are generally of lower magnitude than those recommended here. Cohen's argument appeared to largely rest on an assumption that the dichotomous variable (e.g., group membership) is an artificial dichotomy, thus the use of the point-biserial *r* (which is equivalent, for example, to using the *t* from a *t*-test) would lead to underestimation; he therefore illustrated a conversion from point-biserial *r* to biserial *r*, showing near equivalence in the later value. Although in the past, before the advent of modern computers and widespread education in statistics, many variables may have been artificially dichotomized, today this should be a rare problem—when it occurs we should not adjust our descriptors, rather, we should recognize and note that artificial dichotomization can produce underestimates of association and if possible, the analyses should be re-conducted using interval or continuous measures. Thus, since the values

presented above are appropriate for natural dichotomies (e.g., male/female; placebo/treatment) as well as constructs that are dichotomously defined or conceptualized (e.g., has clinical depression/does not have clinical depression), this author recommends they be adopted (cf. Hopkins, 2002).

Although this author recommends the descriptors in Table 1 be adopted for widespread use, it is important to remember that there is a difference between the small-to-large continuum presented in Table 1 and the unimportant-to-important continuum. A large increase in the risk of a person experiencing one extra night of insomnia in the coming year is less important than a small increase in the risk that this person dies in the coming year. We must not forget the context and definition of the things we are measuring. Nevertheless, when using standardized measures such as EA and GD, conventional interpretations can greatly smooth communication and provide a common base from which to interpret conclusions.

6. Two-Variable Analyses

The foregoing presentation focused on two-group analyses. Two-group analyses, however, are just a special case of two-variable analyses (i.e., where one variable is the dichotomous variable, "group membership"). The sextuple and extended sextuple presented above are equally applicable to the more general two-variable case, with one thing to note.

The issue to note is that the interpretation of the estimated association (EA) may be slightly different. For example, if an analysis compares pre- and post-tests on some variable for one group, there is no EA in the sense of dividing the world population into two groups and examining the real-world association. As mentioned above, this difference in interpretation also occurs for two-group analyses where the whole population cannot be divided into two mutually exclusive groups.

Nevertheless, the sextuple and extended sextuple present the needed information from these types of studies as well; we must merely remain aware that the meaning of the EA in these contexts is limited to the r^2 value coming out of the two-variable comparison, or alternatively, not report the EA (i.e., reporting the GD as the only effect size statistic), however, this author does not favor this latter option as it limits the available information.

In the case where pre-post analyses are being presented for a single group, the 6th position in the sextuple (normally the size of the second group) should be indicated as not applicable ("n/a").

7. Risk Ratios, Odds Ratios, and (Relative) Increased Risk Percentages

The recent popularity of reporting risk ratios (RRs), odds ratios (ORs), and logistic regression ORs in the social and behavioral sciences appears to stem from the ease with which any RR, OR, or logistic regression OR greater than 1 can be arbitrarily and invalidly declared to indicate a large, substantial, important, or significant problem or

association. Readers are reminded that relative measures (e.g., ratios) can never by themselves indicate if an association or risk is small or large (cf. Stadel, Colman, & Sahlroot, 2005). Readers are also reminded that in addition to other limitations, one cannot validly assume that an OR can be used as an estimate for an underlying RR (cf. Holcomb, Chaiworapongsa, Luke, & Burgdorf, 2001; Sackett, Deeks, & Altman, 1996).

Importantly, different studies producing identical RRs—or identical ORs—cannot be assumed to reflect similar findings. Consider the examples presented in Tables 2 through 6. In absolute terms we see that in the example in Table 2 there is a 2.6% increase in risk; Table 3 shows a 5% increase in risk; Table 4 shows a 32% increase in risk; Table 5 shows a 50% increase in risk; and Table 6 shows a 50% increase in risk in a case of an extreme base-rate split.

Table 2. *2.6% Increase in Risk Example With Equal Base-Rates*

	Impaired	Healthy
Control group	1	499
Exposure group	14	486

Table 3. *5% Increase in Risk Example With Equal Base-Rates*

	Impaired	Healthy
Control group	5	495
Exposure group	30	470

Table 4. *32% Increase in Risk Example With Equal Base-Rates*

	Impaired	Healthy
Control group	20	480
Exposure group	180	320

Table 5: *50% Increase in Risk Example With Equal Base-Rates*

	Impaired	Healthy
Control group	50	450
Exposure group	300	200

Table 6. *50% Increase in Risk Example With Vastly Different Base-Rates*

	Impaired	Healthy
Control group	100	900
Exposure group	6	4

Clearly, Tables 2, 3, 4, and 5 show very different cases, with the effect size increasing from example to example. Table 7 presents the r^2 and IQEP difference values, along with the RRs and ORs, for each of these examples.

Table 7: *Analyses of Examples Presented in Tables 2, 3, 4, 5, & 6*

Row #	Absolute increase in risk	Equal base-rates?	r^2	IQEP	RR	OR
1.	2.6% (Table 2)	yes	1.1%	3.2	14.0	14.4
2.	5.0% (Table 3)	yes	1.9%	4.1	6.0	6.3
3.	32% (Table 4)	yes	16.0%	12.0	9.0	13.5
4.	50% (Table 5)	yes	27.5%	15.7	6.0	13.5
5.	50% (Table 6)	no	2.6%	14.8	6.0	13.5

Notes. r^2 = the square of the correlation coefficient; the coefficient of determination. IQEP = IQ equivalent point difference value; *d* multiplied by 15. RR = risk ratio. OR = odds ratio.

As can be seen in the first four rows of Table 7, both r^2 and IQEP difference values properly reflect the increasing size of the association between the variables. A comparison of Table 7 rows 4 and 5 shows that both r^2 and IQEP difference values properly reflect and distinguish between the two "size" questions; row 5 showing that the problem is small in the population as a whole (r^2 = 2.6%) yet medium at the individual level (IQEP difference = 14.8).

Please note, the examples in rows 2 and 4 produce identical RRs and the examples in rows 3 and 4 produce identical ORs. These simple examples show that neither RRs nor ORs can be used to infer the size of the association between variables. Importantly, in this set of examples the smallest increased risk (Table 7, row 1) produces the largest RR and OR. Clearly, RRs and ORs should never be referred to as effect sizes since such terminology promotes misunderstanding (since "large" ratios do not imply large findings). Examples 4 and 5 additionally demonstrate that ORs cannot be assumed to approximate RRs (instead, they sometimes create highly exaggerated impressions). Such measures can be useful in advocacy where one can fallaciously argue, for example, "There is more than a 14-fold increase in risk!" but such exhortations are inappropriate in science and objective journalism when based on relative measures such as RRs and ORs (including logistic regression ORs). The publication of RRs, ORs, and related measures such as relative increased risk percentages[8] as effect size measures should be banned from scientific writing and left to the literary field, as Mark Twain noted[9] there are three types of lies: lies, damned lies, and statistics. RRs and ORs are examples of the third type of lie. The r^2 (i.e., EA) and IQEP difference values (i.e., GD), optionally supplemented with the actual, absolute risk percentages for each group, provide the needed information.

8. Strengths of the Sextuple and Extended Sextuple

The sextuple this author recommends here effectively presents all of the primary information that is traditionally reported in two-group studies, but does so in a standardized, unified way (i.e., independent of the type of statistical analysis performed). The sextuple can be used to completely replace most of the various presentations of two-group statistics such as those used for *t*, *f*, and *r* as well as those for two-group χ^2 and the *z*-test of proportions when the dependent variable is dichotomous. Hence the

aforementioned five disparate presentations can be replaced with a single, more complete presentation. This makes data interpretation and communication of findings easier and less prone to error.

The primary benefits of the sextuple are that it uses easy to understand statistics and encourages fuller reporting of findings which can promote a more complete and correct understanding of the findings than many traditional presentations provide.

The use of this approach has the potential to (1) clearly present information on statistical significance, practical significance, and study size in an easy to read format; (2) encourage the clear reporting of effect sizes and confidence intervals (cf. APA, 2009; APA-JARS, 2008; Des Jarlais et al., 2004) thereby improving understanding of a given study's findings as well as providing valuable information for use in future meta-analyses or other research syntheses; (3) draw attention, in context, to the size of the association of the relationship examined (cf. Cohen, 1994, 1995; Thompson, 2002; Tukey, 1969); (4) encourage researchers to report the means or percentages that were found so that people familiar with the measurement scales will be able to gain a sense of the absolute difference found;[10] (5) minimize false claims of practically significant results from studies not exhibiting statistical significance while alerting scientists to the potential need for further research in the area (e.g., large *p*-value coupled with large effect sizes); (6) minimize mistaking Type I errors for important findings (cf. Tukey, 1969, 1991); (7) minimize false claims of practically significant results from studies evidencing trivially sized findings (e.g., small *p*-value coupled with very small effect sizes and large *n* values) while visually demonstrating that large studies produce small *p*-values regardless of the size of the association (cf. Cohen, 1994, 1995; Tukey, 1991); (8) help people distinguish between social-level and individual-level problems (e.g., small EA coupled with large GD); (9) minimize the chance of people generalizing from studies where one or more of the statistics is based on only a few participants (e.g., small *n* in one or both groups); and (10) lessen communication problems between scientists and journalists (e.g., since easy to interpret statistics are presented, there should be fewer miscommunications caused by use of the word *significant* when it does not mean important or large).

Importantly, the sextuple is easy to understand for both scientists and non-scientists, especially as regards the reporting of effect size via the estimated association (EA: r^2 expressed as a percentage) and the group difference (GD: IQEP difference value). Appendix C presents an introduction to reading and understanding the sextuple for scientific journalists and non-scientists interested in the results of scientific research. At this time, this author knows of no drawbacks or limitations to this method of reporting. Therefore, this author recommends that this, or a similar form of reporting, become the standard for reporting on two-group and two-variable analyses.

For more than a decade there have been calls for effect size reporting to be required in the reporting of research results when such measures are practical and feasible. The 5th edition of the APA publication manual did, indirectly, require it (see APA, 2001, pp. 22-23) and the 6th edition makes this more clear—". . . complete reporting of . . . estimates of appropriate effect sizes and confidence intervals are the minimum expectations for all

APA journals" (APA, 2009, p. 33). Regardless of whether or not the sextuple approach presented above comes to be widely adopted, this author reminds readers that when a peer-reviewed journal article makes an assertion such as "a substantial association has been found" or "these large findings show" or similarly "these findings demonstrate the importance of" the authors are asserting conclusions, ostensibly based on the findings, which are inherently based on an assessment of the effect size found. If the authors did not calculate an effect size measure, then they are guilty of fabricating their conclusions; if they did base their conclusions on one or more effect size measures, then it seems there can be no good reason not to convert such measures into something easily understood by others. The EA and the GD can easily accomplish this. Editors, peer reviewers, and journalists should require authors to provide easy to understand statistical findings when such is practical and feasible. The EA and/or the GD are practical and feasible for the vast majority of social science research when comparing two groups or examining two variables. This author strongly urges publishers to require easy to understand reporting of findings, and additionally urges journalists to ask for such measures whenever they are faced with a conclusion that is inherently based on an assessment of the size of an association, but where a clear statistic providing that size has not been provided or where only potentially misleading statistics, such as *p*-values, risk ratios, or odds ratios have been presented.

Endnotes

1. When the dependent variable is dichotomous, chi-square (χ^2) results as well as the results from a *z*-test of proportions are also equivalent to the results of *f*, *t*, and *r*. Note also that all of the reported *p*-values are correct, since the exact *p*-value is $p = 0.00049$ for these analyses.

2. When applying Bonferroni correction a difference below the corrected *p*-value threshold only indicates that the groups differ at the overall alpha level (Bland & Altman, 2006). For example, if alpha is .05 and there are five tests, the corrected threshold is .01 (i.e., .05 / 5). If one test has an exact *p*-value of .009, although it is correct to report that $p < .01$, this finding only indicates a difference between groups at the .05 level.

3. If a study reports statistics from different levels of analysis the total number of tests is the total for the appropriate level. For example, imagine a study that was designed to investigate the contribution of salary (high or low) to some other variables, performing the tests separately for males and females, where the variables of interested were depression, self-esteem, and eating disorders. It started with an omnibus test for each sex, then did three follow-up tests for each sex. The total for omnibus tests would be 2, whereas the total for follow-up tests would be 6 (3 for men and 3 for women). Hence if a literature review cited only the self-esteem finding for women, the correct total would be 6.

4. In the case of meta-analysis, this question applies even when the results are not statistically significant.

5. For example, a study that determined that an $r = .2$ reduced to $r = .1$ after controlling for some confounding variable might state that the association was reduced by 50%. This, however, would be incorrect since *r* is not an equal-interval measure. The correct figure would be that the association was reduced by 75% (4% reduced to 1%).

6. Note that Cohen (1988) defined *d* with the denominator being the standard deviation of either group (since the null hypothesis being tested assumes both groups are sampled from the same population and thus will, on average, have identical standard deviations). Using the standard deviation of the combined data is more appropriate since (1) it provides a larger sample upon which to calculate the standard deviation thus reducing random error in this statistic, (2) when the two groups really are from different populations, in real life social science research any difference in standard deviations is unlikely to be due exclusively to the effects of the independent variable (arising instead, for example, from confounding variables), thus, using a control group's standard deviation will frequently underestimate the "true" standard deviation, which leads to exaggerated effect sizes, and (3) this approach directly follows from the logic of null hypothesis testing. Consequently, the combined data should be used to calculate the standard deviation used to calculate *d*. It should be noted, however, that when it has already been clearly determined that the null hypothesis is incorrect, an alternative *d*-statistic might be preferred (in the case of true experiments which rarely occur in social science research Glass's delta might be appropriate, and in many social science experiments Hedge's *G* might be a logical choice).

7. This author put the term *effect size* in scare quotes to remind the reader that although referred to with the term *effect*, in most quasi-experimental and survey research these measures are measures of association. They often do not represent the size of a cause-and-effect relationship between group membership and the dependent variable—even when a cause-and-effect relationship has been hypothesized and that hypothesis led to conducting the study. In general, the effect size is an upper limit on the magnitude of the true effect, with the lower limit being zero.

8. Relative increased risk percentages are just RRs expressed another way. For example, a RR of 1.2 represents a 20% relative increase in risk. Relative increased risk measures further promote miscommunication since authors frequently do not specify if the provided figure is relative or absolute. Since in the English language the unmarked (i.e., default) meaning is absolute, when a scientist finds an increase in risk from 1.0% to 1.2% and describes it as a "20% increase in risk!", the reader may, and in the English language should, think the *absolute* increase was 20%, rather than one fifth of 1%.

9. For a discussion of the origins of the saying, see Rees (n.d.).

10. While advocating for reporting statistics in their original units, rather than in dimensionless units (such as *r*), Tukey (1969, pp. 89) wrote: "Being so disinterested in our variables that we do not care about their units can hardly be desirable."

References

APA Publications and Communications Board Working Group on Journal Article Reporting Standards. (2008). Reporting standards for research in psychology: Why do we need them? What might they be? *American Psychologist*, *63*(9), 839-851. DOI: 10.1037/0003-066X.63.9.839

American Psychological Association. (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: Author.

American Psychological Association. (2009). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author.

Bland, J. M., & Altman, D. G. (2006). Multiple significance tests: The Bonferroni method. *BMJ*, *310*, 170.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

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

Cohen, J. (1995). The earth is round ($p < .05$): Rejoinder. *American Psychologist*, *50*, 1103.

Des Jarlais, D. C., Lyles, C., Crepaz, N., & the TREND Group. (2004). Improving the reporting quality of nonrandomized evaluations of behavioral and public health interventions: The TREND statement. *American Journal of Public Health*, *94*(3), 361-366.

Holcomb, W. L., Chaiworapongsa, T., Luke, D. A., & Burgdorf, K. D. (2001). An odd measure of risk: Use and misuse of the odds ratio. *Obstetrics & Gynecology*, *98*, 685-688.

Hopkins, W. (2002). *A new view of statistics: A scale of magnitudes for effect statistics*. Retrieved December 1, 2012, from http://www.sportsci.org/resource/stats/effectmag.html

Rees, N. (n.d.). *The most quoted remarks*. Retrieved December 1, 2012, from http://www1c.btwebworld.com/quote-unquote/p0000149.htm

Rosenthal, R. (1991). *Meta-analytic procedures for social research* (Revised ed.). London: Sage.

Rosenthal, R., & Rosnow, R. L. (1991). *Essentials of behavioral research: Methods and data analysis* (2nd ed.). New York: McGraw-Hill.

Sackett, D. L., Deeks, J. F., & Altman, D. G. (1996). Down with odds ratios! *Evidence-Based Medicine, 1*(6), 164-166.

Stadel, B. V., Colman, E., & Sahlroot, T. (2005). Misleading use of risk ratios [Correspondence]. *Lancet, 365*, 1306-1307.

Thompson, Bruce (2002). "Statistical," "practical," and "clinical": How many kinds of significance do counselors need to consider? *Journal of Counseling and Development, 80*, 64-71.

Tukey, J. W. (1969). Analyzing data: Sanctification or detective work? *American Psychologist, 24*, 83-91.

Tukey, J. W. (1991). The philosophy of multiple comparisons. *Statistical Science, 6*, 100-116.

Received 20 July 2012
Accepted 25 November 2012

Appendix A:

Formulas for Calculating the Estimated Association (i.e., EA, r^2)

The EA can be precisely calculated from numerous two-group statistics when the sample sizes are in proportion to their prevalence in the population. If the two-group statistics came from an artificial setting (e.g., a survey that included equal numbers of PhDs compared with non-PhDs when in the real world this is not a 50/50 split) the EA can be estimated or listed as "n/a". Note that these formulas should not be used for statistics that were based on three or more groups. N in the following formulas indicates the number of participants for independent group tests, or the number of correlated/matched pairs for correlated group tests.

EA from correlation, Pearson correlation, point-biserial correlation, *r*:

$$EA = 100 \times (r)^2$$

EA from Student *t*-test, *t*-test, *t*:

$$EA = 100 \times \left(\frac{t^2}{[t^2 + (N-2)]} \right)$$

EA from analysis of variance, ANOVA, *f*:

$$EA = 100 \times \left(\frac{f}{[f + (N-2)]} \right)$$

EA from chi-square (χ^2) for 2x2 tables:

$$EA = 100 \times \left(\frac{\chi^2}{N} \right)$$

EA from proportions analyzed with the *z*-test of proportions:

$$EA = 100 \times \left(\frac{z^2}{N} \right)$$

Appendix B:

Formulas for Calculating the Group Difference (i.e., GD; IQEP Difference Value)

The GD can be precisely calculated using the following formula:

$$GD = 15 \times \left(\frac{(\overline{X}_1 - \overline{X}_2)}{S_{grand}} \right)$$

where:

n_1 = number of people (or things) in group 1
n_2 = number of people (or things) in group 2
N = $n_1 + n_2$

$\overline{X}_1$ = mean of group 1
$\overline{X}_2$ = mean of group 2
$\overline{X}_{grand} = \frac{n_1\overline{X}_1 + n_2\overline{X}_2}{n_1 + n_2}$

S_1 = standard deviation of group 1 (n_1 minus 1 in the denominator)
S_2 = standard deviation of group 2 (n_2 minus 1 in the denominator)
S_{grand} = standard deviation of the combined groups (N minus 1 in the denominator)

If you do not have the raw data available, S_{grand} can be precisely calculated from the group sizes, means, and standard deviations:

$$S_{grand} = \sqrt{\frac{[(n_1 - 1) \times S_1^{\,2}] + [(n_2 - 1) \times S_2^{\,2}] + [n_1 \times (\overline{X}_{grand} - \overline{X}_1)^2] + [n_2 \times (\overline{X}_{grand} - \overline{X}_2)^2]}{(n_1 + n_2 - 1)}}$$

Appendix C:

Interpreting Philip's Sextuple and Philip's Extended Sextuple

C1. Philip's Sextuple

Philip's sextuple brings together 6 important pieces of information when appraising or evaluating the results of a study comparing two groups (e.g., smokers & non-smokers, men & women, placebo group & real medication group). The presentation is standardized to allow easy interpretation and to allow easy comparisons between different studies.

The first piece of information is formally known as a *p*-value. In most social science research, if this value is below .05 then we say the finding is *statistically significant*. This only means that the difference between the groups was larger than expected by chance alone. It does not mean the difference between groups was large or important.

The second piece of information is a ratio. It tells you how many statistical tests in the study were found to be statistically significant, compared to the number of tests that were or could have been performed. This is important because "statistically significant" results do sometimes occur by chance alone (in general, in social science research this occurs at least 5% of the time). Hence values near "1/20" suggest that the findings are the product of chance, even if they are "statistically significant."

The third piece of information is known as the *estimated association* or EA. It is a number between 0 and 100; it may or may not be reported with a percent sign. Zero means there is no relationship between group membership and the other variable and 100 means there is a perfect relationship. Perfect relationships never really happen in social science research, but were one to occur, it would mean that there is a mathematical equation that could perfectly predict a person's score based on knowing which group they are in. This third piece of information is often an assessment at the societal level (and may be underlined to indicate this), and thus is important to look at when thinking about public policy, national health, and other issues that involve a large population and trying to understand the relationship between group membership and a potential outcome variable in the community at large.

Put another way, the estimated association tells you how well knowing one piece of information (i.e., group membership) predicts the other variable (e.g., 0% predictive ability, 100% predictive ability). In general, an EA value below 1 is considered *very small*, values from 1 to about 10 are *small*, from 10 to around 25 are *medium*, and from 25 to 50 are *large*. It is rare to find associations larger than 50 but they could properly be described as *very large*. If these values appear small to you, remember that most things we measure (e.g., self-esteem) are influenced by many factors, hence a finding of 25% really is a potentially important finding, although it is not the complete picture (i.e., it does not provide 100% predictive ability).

Note that in real life we must consider costs and benefits for most things, hence you might hear that a very small association should be viewed as important (e.g., if a drug given to thousands of people saved only a few lives, but had no side effects and cost almost nothing, although the association between taking the drug and living longer might be very small at the population level, it may nevertheless be worth taking the drug since you might benefit and there is little or no risk or cost associated with taking the drug), or that a medium association is too small to worry about (e.g., since there might be side-effects, for example, from taking a drug). The words *very small*, *small*, *medium*, *large*, and *very large* are thus starting points for assessing things.

The fourth piece of information is known as the *group difference* or GD. It is a number that expresses how different, on average, the two groups are from each other; you can think of this as looking at average differences at the individual level. This is somewhat different from the third piece of information in that this fourth piece of information assumes group sizes are equal. Note that if something had a large impact on a very small percentage of people, this fourth piece of information would show the large impact (an individual level comparison), whereas the estimated association would properly show a small finding (since, at the societal level, the association is small, since so few people are affected).

The group difference is designed to be interpreted in the same way as IQ scores, but can be used to measure anything. Hence if the thing being measured really was IQ, a 7 point GD would really mean a 7 point difference in IQ. If you have a sense of how big a difference of a given value is in IQ points, then when you see this fourth piece of information you can have a good idea of how different the two groups scored on whatever was being measured (e.g., self-esteem, happiness, support for a particular political candidate, etc.). In interpreting the group difference, values less than 3 are generally considered to be *very small*, from 3 up to 10 are *small*, from 10 up to 20 are *medium*, from 20 up to 30 are *large*, and values greater than 30 are *very large* (and rarely occur since most things we measure are affected by many different factors).

The fifth piece of information merely reports the number of people in the control group, or if there was no control group (e.g., when comparing males and females) the number of people in one of the groups. The sixth piece of information reports the number of people in the other group. This information is important since you should not trust the measures of the size of the association between groups if there was only a small number of people in one of the groups.

If the EA or GD values are placed between two others like this: "10<15<20", the measurement is the number in the middle (i.e., 15) and the other two numbers provide what is known as a 95% confidence interval. In this example, the scientist would be saying that based on the data collected, the true measure probably lies somewhere between 10 and 20, with 15 being the best guess. You may also sometimes see even more numbers like this: "8.4<10<15<20<21.6". This is just like the prior case, except that it additionally reports a 99% confidence interval (running from 8.4 to 21.6, with the best estimate being 15).

C2. Philip's Extended Sextuple

Philip's extended sextuple is merely the six pieces of information discussed above, plus the average score for each group, or alternatively, the percentage of each group that was affected or had the trait being measured. The seventh item corresponds to the fifth item (i.e., the number of people in the control group) and the eighth item presents the average or percent affected for the other group. If an average is followed by a colon (":") and another number, that extra number is the standard deviation, a measurement scientists and mathematicians use to measure how much variation there is in a group of scores.

Notes. If some of the information is not available, a question mark should be used to note that the information was not available. If a "c" (standing for "circa") precedes a value, it means the exact value could not be calculated but an estimate was possible and the presented value is the estimate. If "n/a" is listed instead of a value, this indicates that due to the study design that particular value is not applicable (e.g., the size of the 2nd group, when reporting an analysis that looked a one group of people at two points in time).